Semantic Role Labeling Tutorial: Part 3

Semi- , unsupervised and cross-lingual approaches

Ivan Titov

NAACL 2013

Universiteit van Amsterdam

Universität des Saarlandes

Shortcomings of Supervised Methods

- Supervised methods:
  - Rely on large expert-annotated datasets (FrameNet and PropBank > 100k predicates)
  - Even then they do not provide high coverage (esp. with FrameNet)
    - ~50% oracle performance on new data [Palmer and Sporleder, 2010]
  - Resulting methods are domain-specific [Pradhan et al., 2008]
  - Such resources are not available for many languages
- How can we reduce reliance of SRL methods on labeled data?
  - Transfer a model or annotation from a more resource-rich language (*crosslingual transfer / projection*)
  - Complement labeled data with unlabeled data (*semi-supervised learning*)
  - Induce SRL representations in an unsupervised fashion (*unsupervised learning*)

Much less mature area than supervised learning for SRL

Outline

- Crosslingual annotation and model transfer
- Semi-supervised learning
- Unsupervised learning

Exploiting crosslingual correspondences: classes of methods

- The set-up:
  - Annotated resources or a SRL model is available for the source language (often English)
  - No or little annotated data is available for the target language
- How can we build a semantic-role labeller for the target language?
  - If we have parallel data, we can project annotation from the source language to the target language (*annotation projection*)

    [Pado and Lapata, 2005; Johansson and Nugues, 2006; Pado and Pitel, 2007; Tonelli and Pianta, 2008; van der Plas et al., 2011]
  - If no parallel data, we can directly apply a source SRL model to the target language (*driect model transfer*)

    [Kozhevnikov and Titov, 2013]

Crosslingual annotation projection: basic idea

- Start with an aligned sentence pair

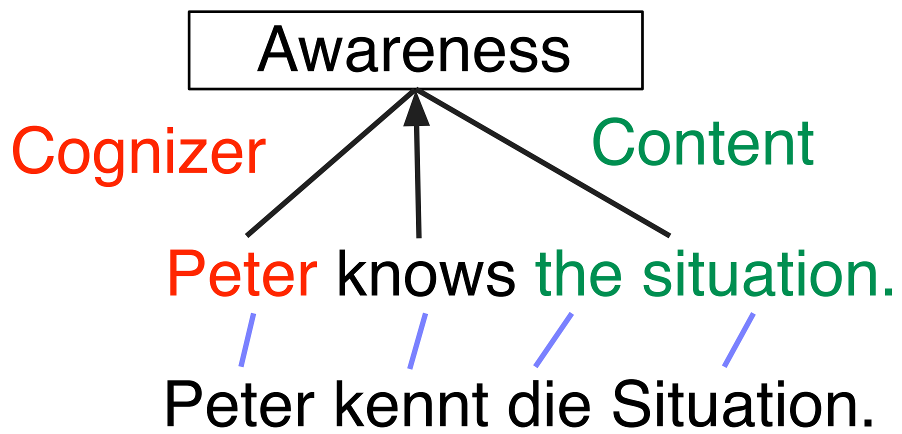
Peter knows the situation.

Peter kennt die Situation.

Example from
Sebastian Pado

Crosslingual annotation projection: basic idea

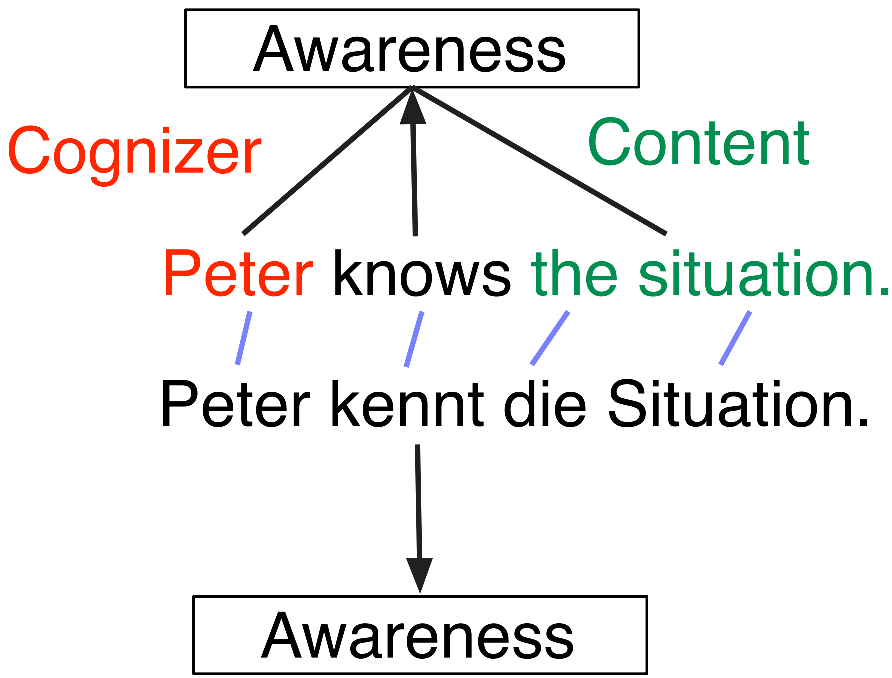
- Start with an aligned sentence pair
- Label the source sentence



Example from
Sebastian Pado

Crosslingual annotation projection: basic idea

- Start with an aligned sentence pair
- Label the source sentence
- Check if a target predicate can evoke the same frame

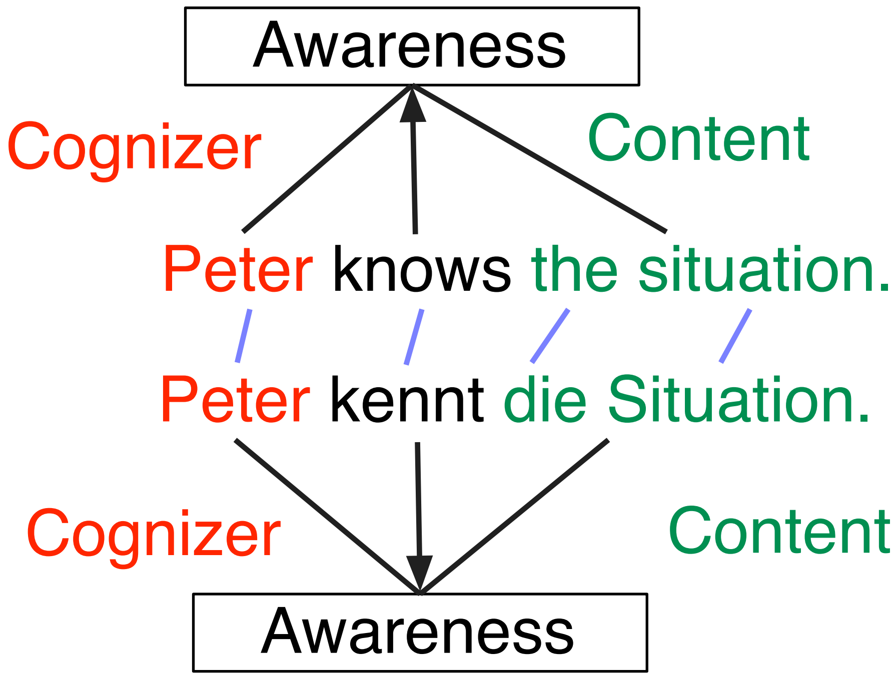


Example from
Sebastian Pado

Crosslingual annotation projection: basic idea

- Start with an aligned sentence pair
- Label the source sentence
- Check if a target predicate can evoke the same frame
- Project roles from source to target sentence

How do we project?

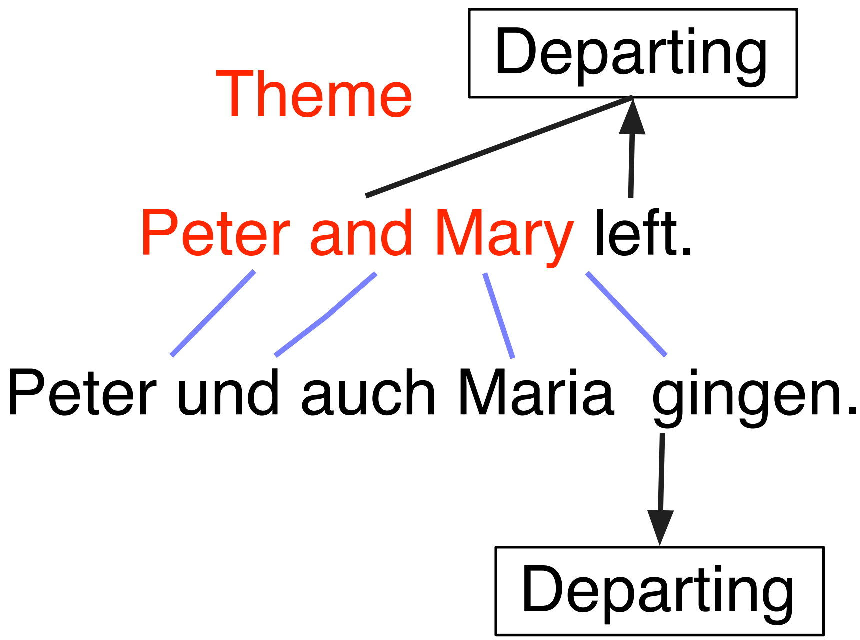


Assumption: role structure is preserved across languages

Example from Sebastian Pado

Word-based projection

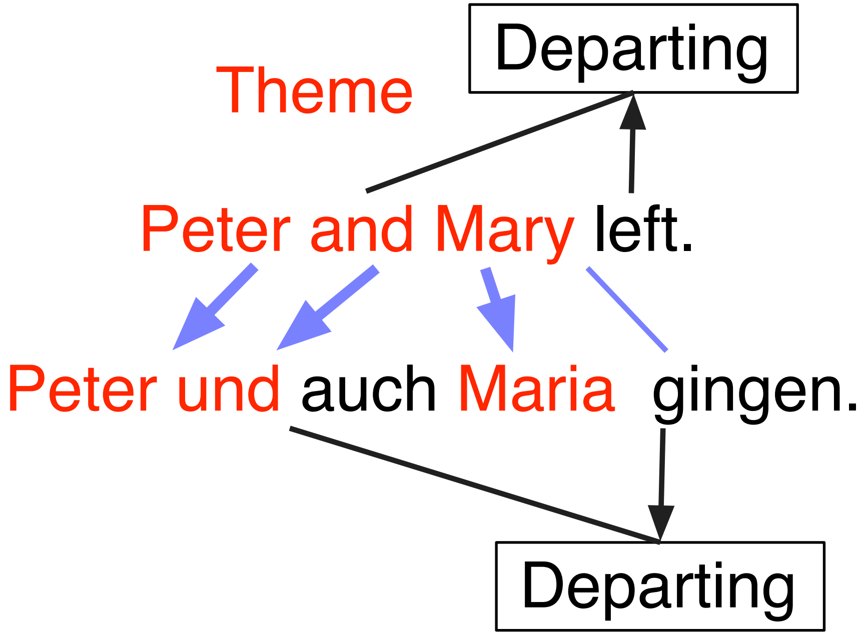
- For each source semantic role:



Example from
Sebastian Pado

Word-based projection

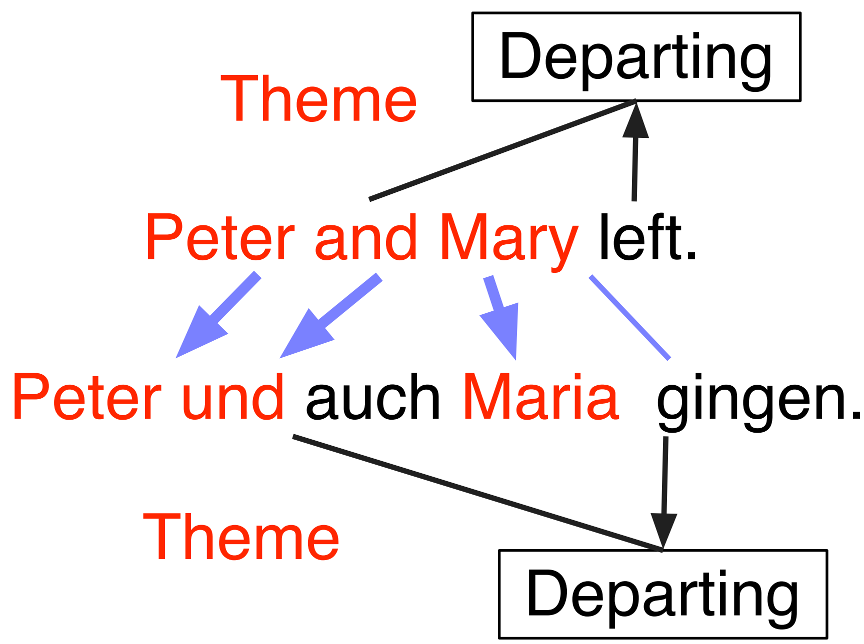
- For each source semantic role:
  - Follow alignment links



Example from
Sebastian Pado

Word-based projection

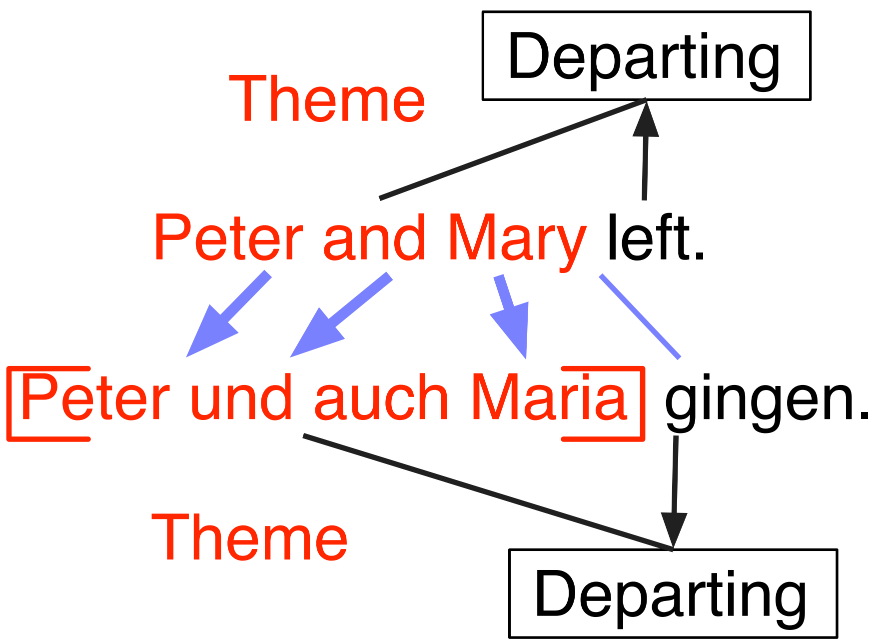
- For each source semantic role:
  - Follow alignment links
  - Target role spans all the projected words



Example from
Sebastian Pado

Word-based projection

- For each source semantic role:
  - Follow alignment links
  - Target role spans all the projected words
  - Ensure contiguity



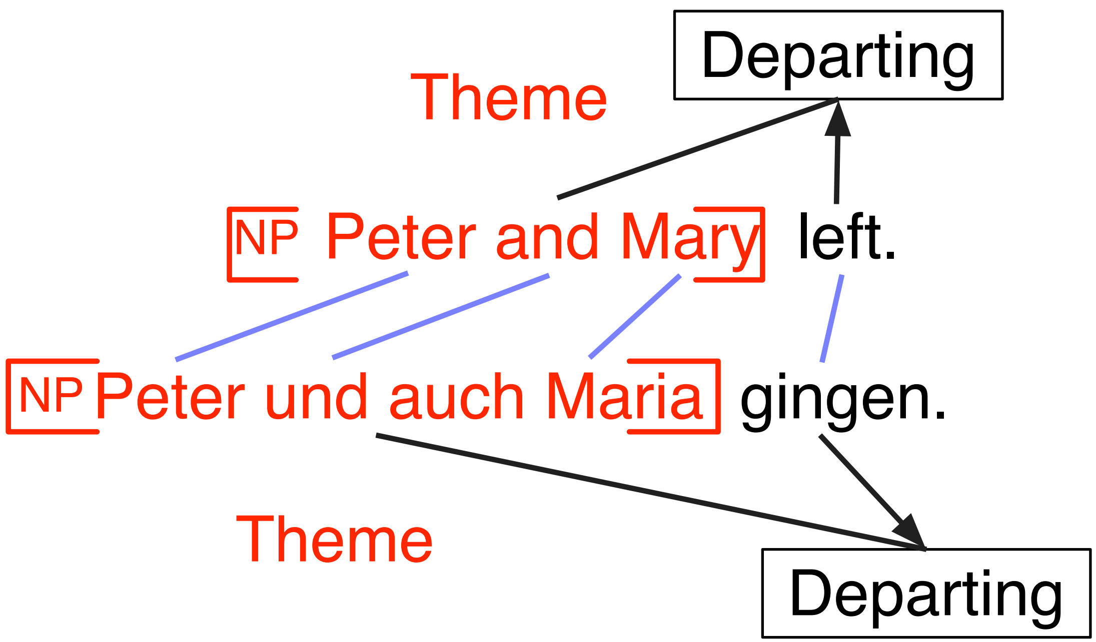
Noisy because of errors and omission in word alignments

Example from
Sebastian Pado

Syntax-based projection

- Find alignment between constituents
- For each source semantic role:
  - Identify a set of constituents in the source sentences
  - Label aligned constituents with the semantic role

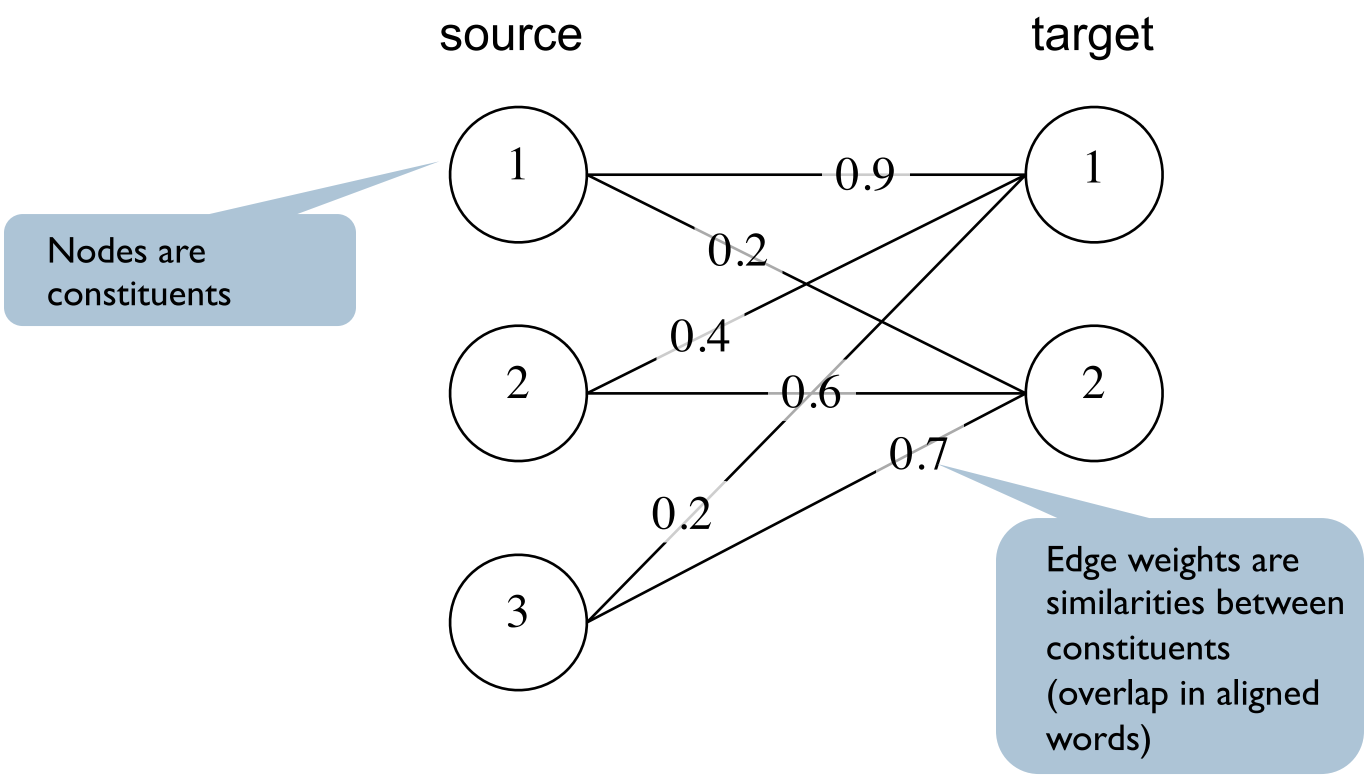
We have an alignment between words, how do we get one for constituents?



Example from
Sebastian Pado

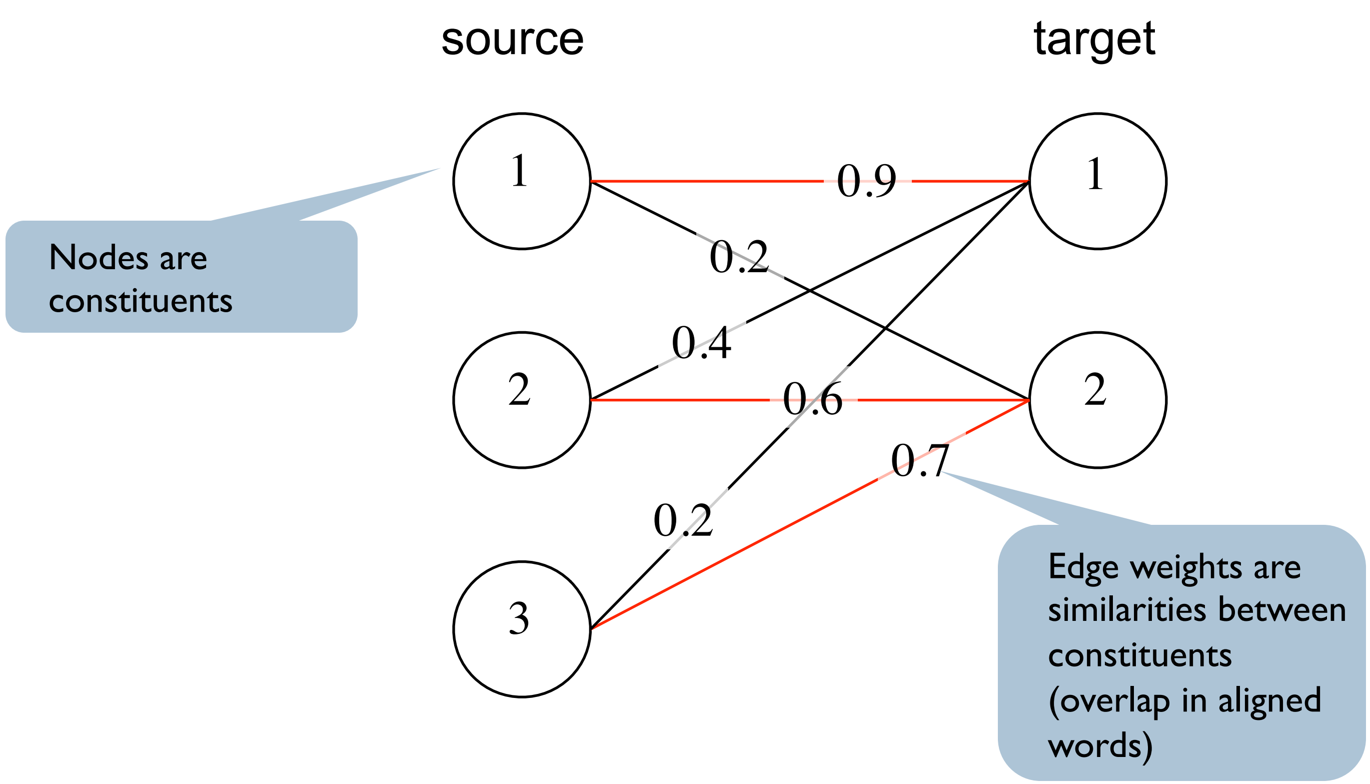
Syntax-based projection

- Define semantic alignment as an optimization task on a graph
- Graph for each sentence pair



Syntax-based projection

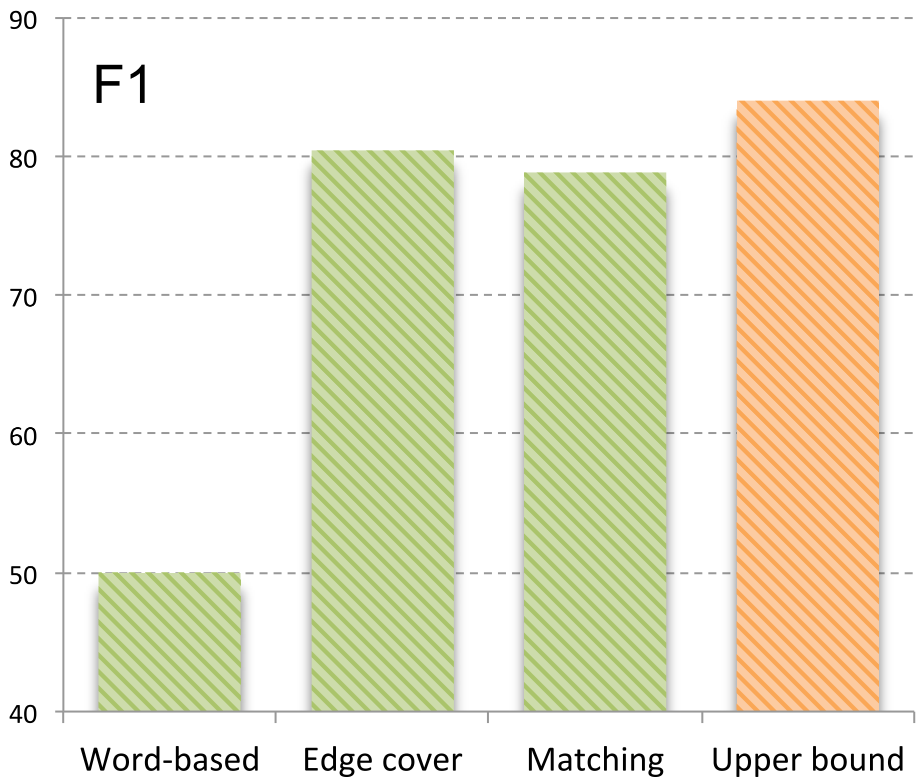
- Define semantic alignment as an optimization task on a graph
- Graph for each sentence pair



- Choose an optimal alignment graph, maybe with some constraints:
  - Covers all target constituents (edge cover)
  - Edges in the alignment do not have common endpoints (matching)

Evaluation

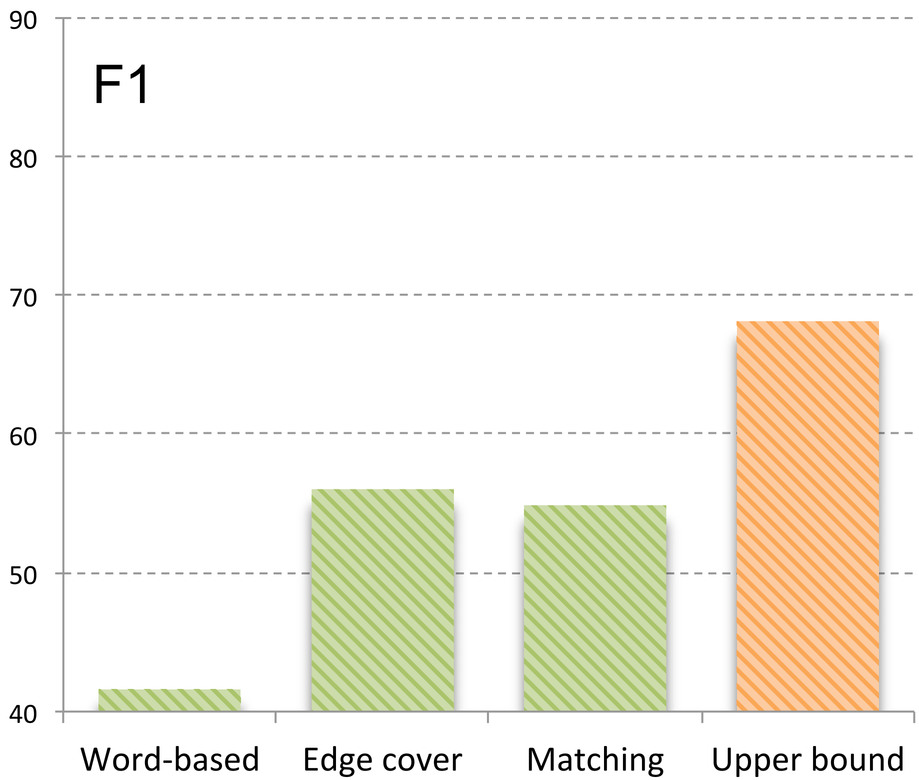
- English to German, FrameNet-style representations
- Manual syntax (for 2 languages), manual SRL for source, auto alignments



The evaluation is limited to sentences where a frame is preserved in translation

Evaluation

- English to German, FrameNet-style representations
- Auto syntax (for 2 languages), auto SRL for source, auto alignments



The evaluation is limited to sentences where a frame is preserved in translation

The projected annotation can be used train a semantic-role labeller for the target language

For semantic-role dependency representations word-based transfer (along with an heuristic for treating prepositions and conjunctions) is more competitive

Outline

- Crosslingual annotation and model transfer
  - Annotation projection
  - Direct transfer
- Semi-supervised learning
- Unsupervised learning

Direct transfer of models

- Is there a simpler (?) method which does not (directly) require parallel data?

- Direct transfer (DT) of models:
  - Train a model in one language
  - Apply to sentences in another language

- Is this realistic at all?
  - Requires (maximally) language-independent feature representation
  - Have been tried successfully for syntax [Zeman and Resnik, 2008; Tackstrom et al., 2012]
  - Performance depends on how different the languages are

[Kozhevnikov and Titov, 2013]

Language independent feature representations

- Instead of words use either
  - cross-lingual word clusters [Tackstrom et al., 2012] or
  - cross-lingual distributed word features [Klementiev et al., 2012]
- Instead of fine-grain part-of-speech (PoS) tags use coarse universal PoS tags

  [Petrov et al., 2012]
- Instead of rich (constituent or dependency) syntax either use either
  - unlabeled dependencies or
  - transfer syntactic annotation from the source language before transferring semantic annotation and use it

Language independent feature representations

- CoNLL-2009 data (dependency representation for semantics)
- Target syntax is obtained using direct transfer
- Only accuracy on labeling arguments (not identification)

For the identification task relative performance between the methods is similar

Language pair	Direct transfer	Annotation projection
English to Chinese	*70.1*	*69.2*
Chinese to English	*65.6*	*61.3*
English to Czech	*50.1*	*46.3*
Czech to English	*53.3*	*54.7*
English to French	*65.1*	*66.1*

DT achieves comparable performance to AP and does not (directly) require parallel data

A SRL model trained on projected sentences (word-based projection on top of dependencies)

Outline

- Crosslingual annotation and model transfer
- Semi-supervised learning
- Unsupervised learning

Semi-supervised learning: classes of methods

- There are three main groups of semi-supervised learning (SSL) methods considered for SRL:
  - ***methods creating surrogate supervision:*** automatically annotate unlabeled data and treat it as new labeled data (annotation projection / bootstrapping methods)
  - ***parameter sharing methods*:** use unlabeled data to induce less sparse representations of words (clusters or distributed representations)
  - ***semi-unsupervised learning*:** adding labeled data (and other forms of supervision) to guide unsupervised models

We will discuss these methods towards the end of the tutorial

Outline

- Crosslingual annotation and model transfer
- Semi-supervised learning
  - methods creating surrogate supervision
  - parameter sharing methods
- Unsupervised learning

Creating surrogate supervision

How do we choose examples?

1. Choose examples (sentences) to label from an unlabeled dataset
2. Automatically annotate the examples

How do we annotate examples?

3. Add them to the labeled training set
4. Train a classifier on the expanded training set
5. Optional: Repeat

Makes sense only if the classifier is used at stages 1 or 2

- Basic self-training
  - Use the classifier itself to label examples (and, often, its confidence to choose examples at stage 1)
  - Does not produce noticeable improvement for SRL [He and Gildea, 2006]

Need a better method for choosing and annotating unlabeled examples

[Furstenau and Lapata, 2009]

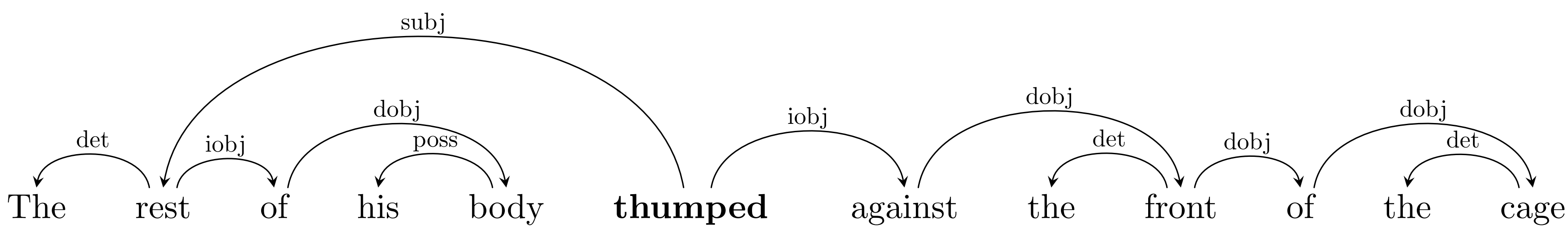
Monolingual projection: an idea

- Assumptions: sentences similar in their lexical material and syntactic structure are likely to share the same frame-semantic structure
- An example:
  - Labeled sentence: [*His back*]$_{Impactor}$ [***thudded***]$_{\textbf{Impact}}$ [*against the wall*]$_{Impactee}$
  - Unlabeled sentence: *The rest of his body thumped against the front of the cage*
- An Implementation (roughly):
  - Choose labeled examples which are *similar* to an unlabeled example (compute scored alignments between them, select pairs with high scores)
  - Use alignments to project semantic role information to the unlabeled sentences

How do we compute these alignments?

Monolingual projection: alignment

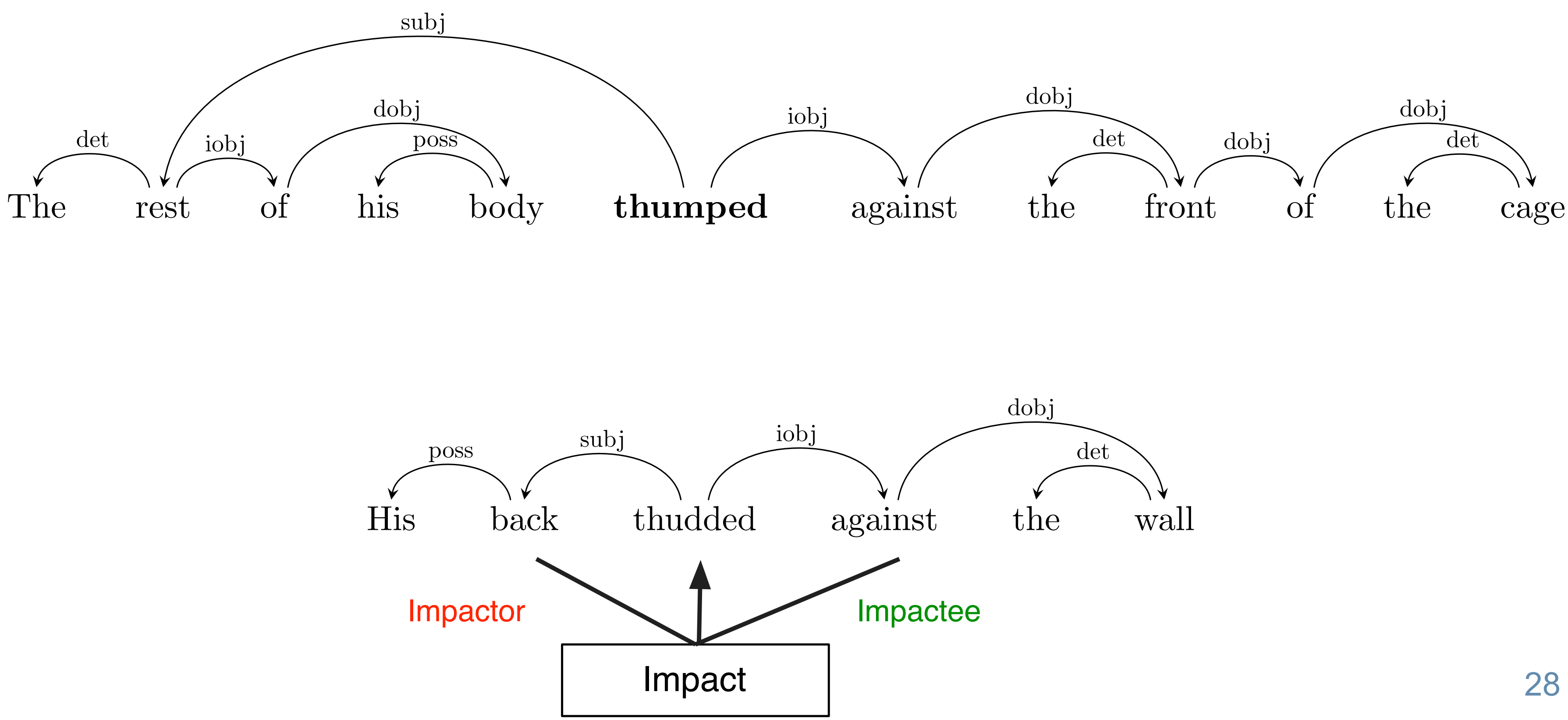
- Start with an unlabeled sentence, and a target predicate



[Furstenau and Lapata, 2009]

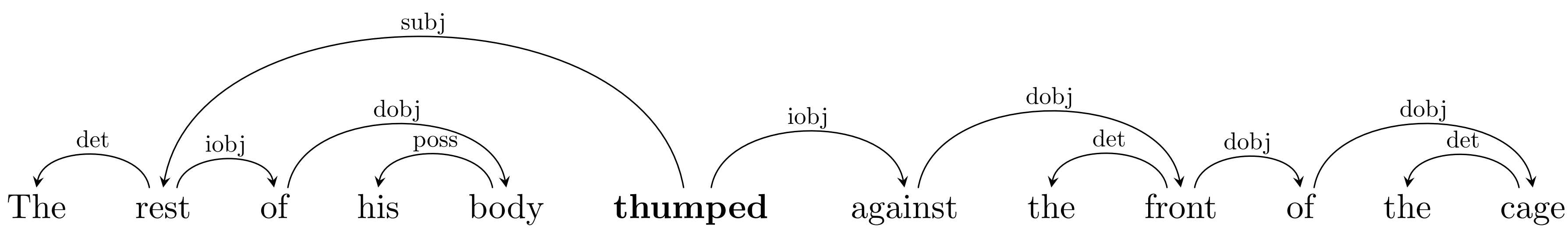
Monolingual projection: alignment

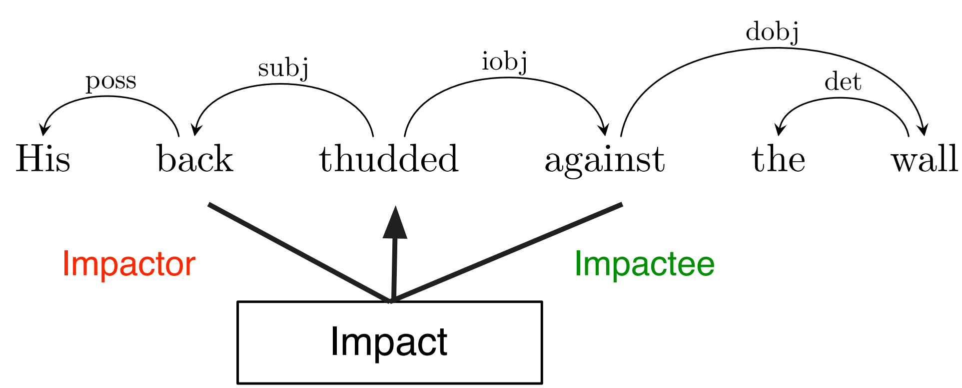
- Start with an unlabeled sentence, and a target predicate
- Check a labeled sentence (one by one)



Monolingual projection: alignment

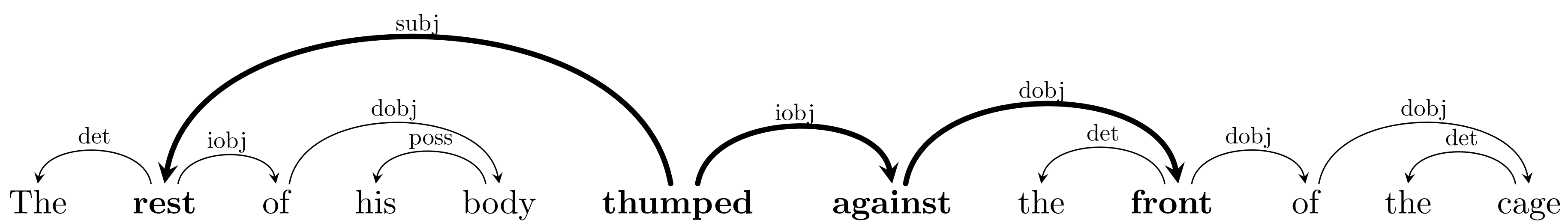
- Start with an unlabeled sentence, and a target predicate
- Check a labeled sentence (one by one)
- Find the best alignment

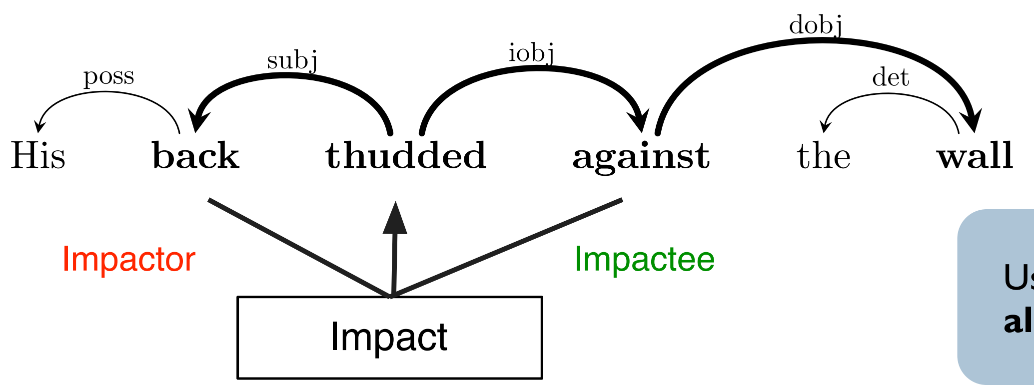




Monolingual projection: alignment

- Start with an unlabeled sentence, and a target predicate
- Check a labeled sentence (one by one)
- Find the best alignment



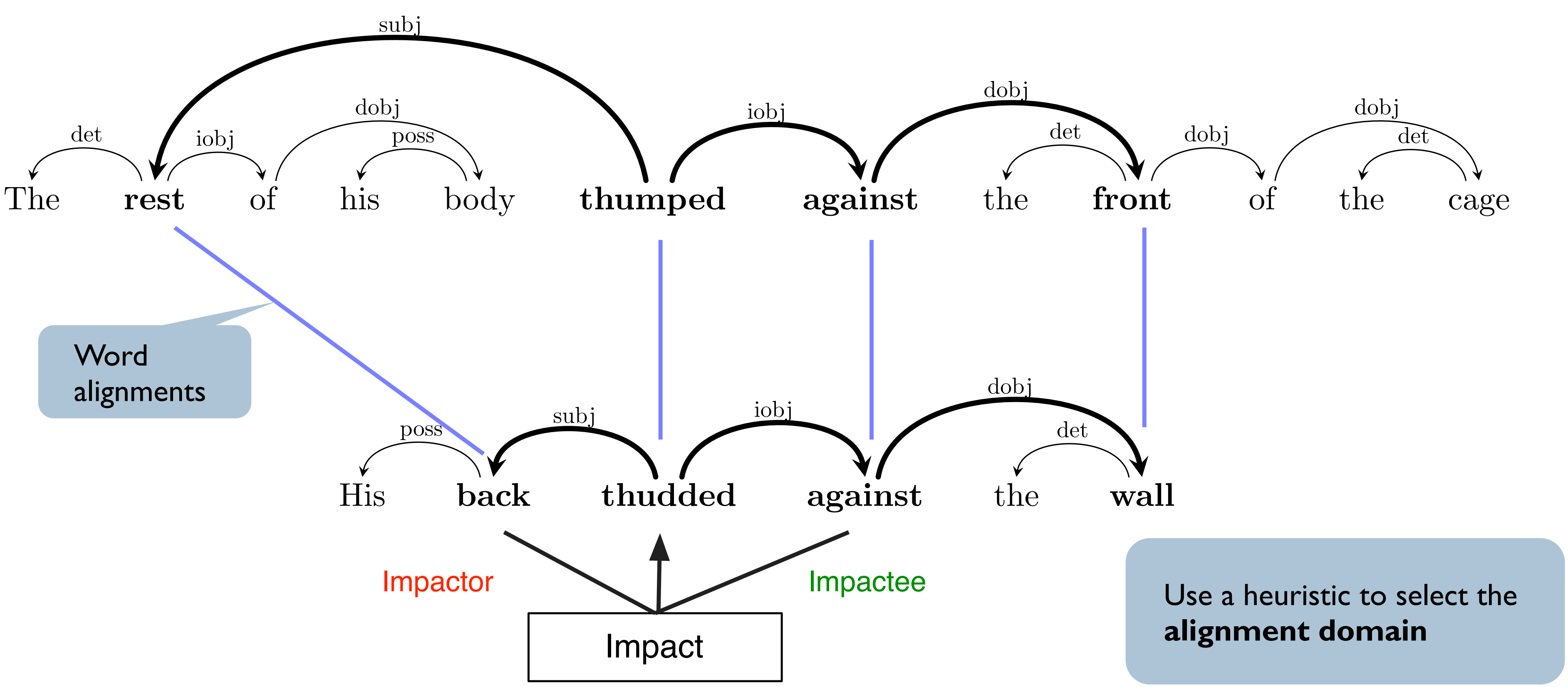


Use a heuristic to select the **alignment domain**

[Furstenau and Lapata, 2009]

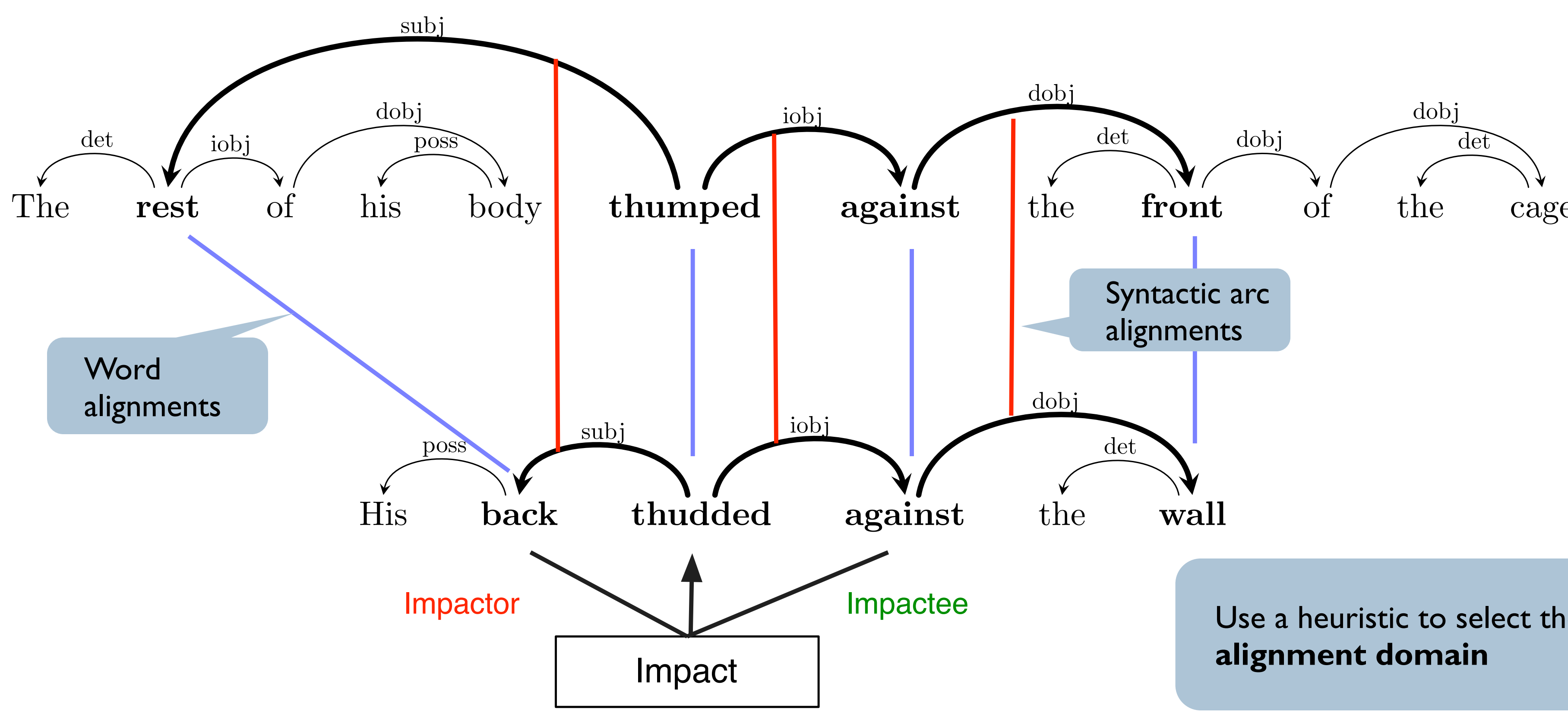
Monolingual projection: alignment

- Start with an unlabeled sentence, and a target predicate
- Check a labeled sentence (one by one)
- Find the best alignment



Monolingual projection: alignment

- Start with an unlabeled sentence, and a target predicate
- Check a labeled sentence (one by one)
- Find the best alignment



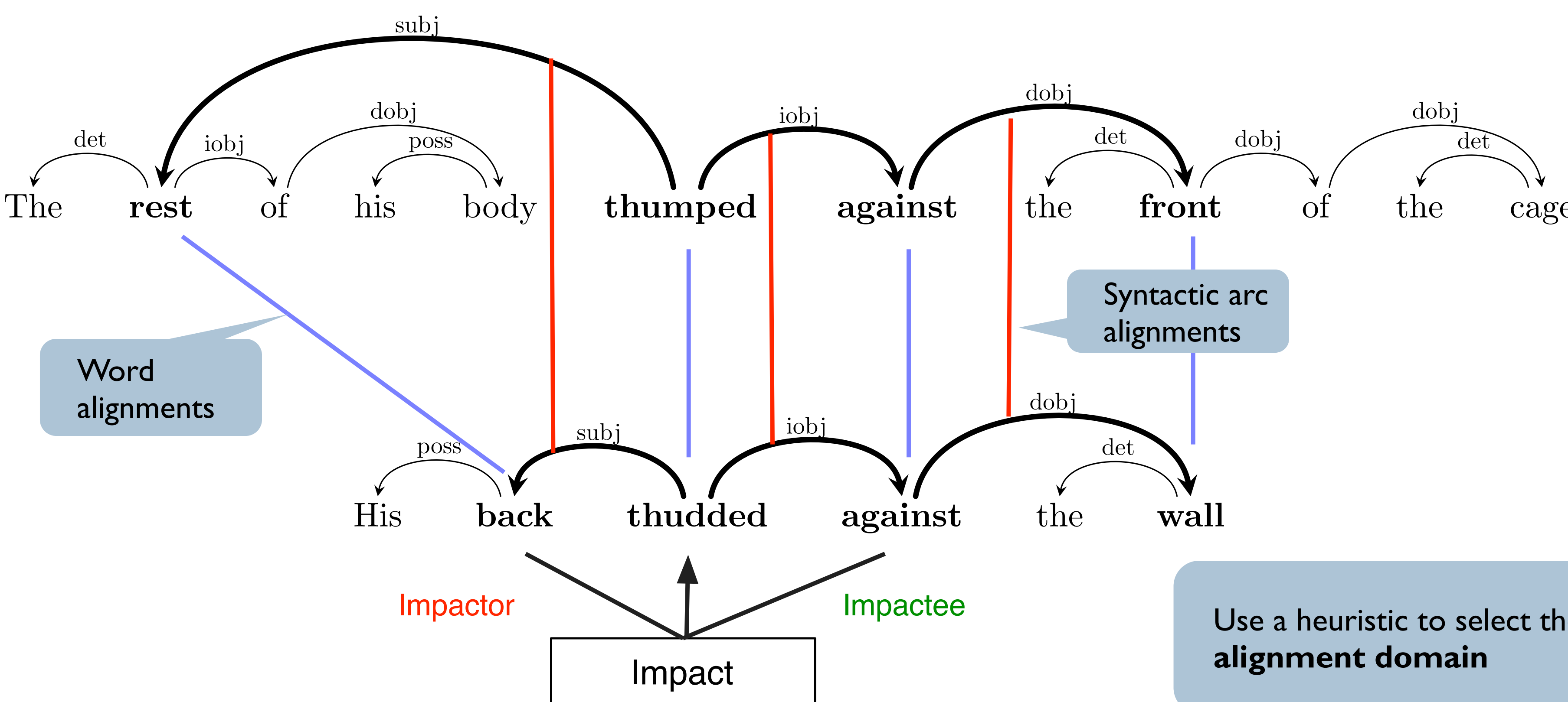
Use a heuristic to select the **alignment domain**

[Furstenau and Lapata, 2009]

Monolingual projection: alignment

- Start with an unlabeled sentence, and a target predicate
- Check a labeled sentence (one by one)
- Find the best alignment, with **Score = Lexical Score + Syntactic Score**

Using integer linear programming



Use a heuristic to select the **alignment domain**

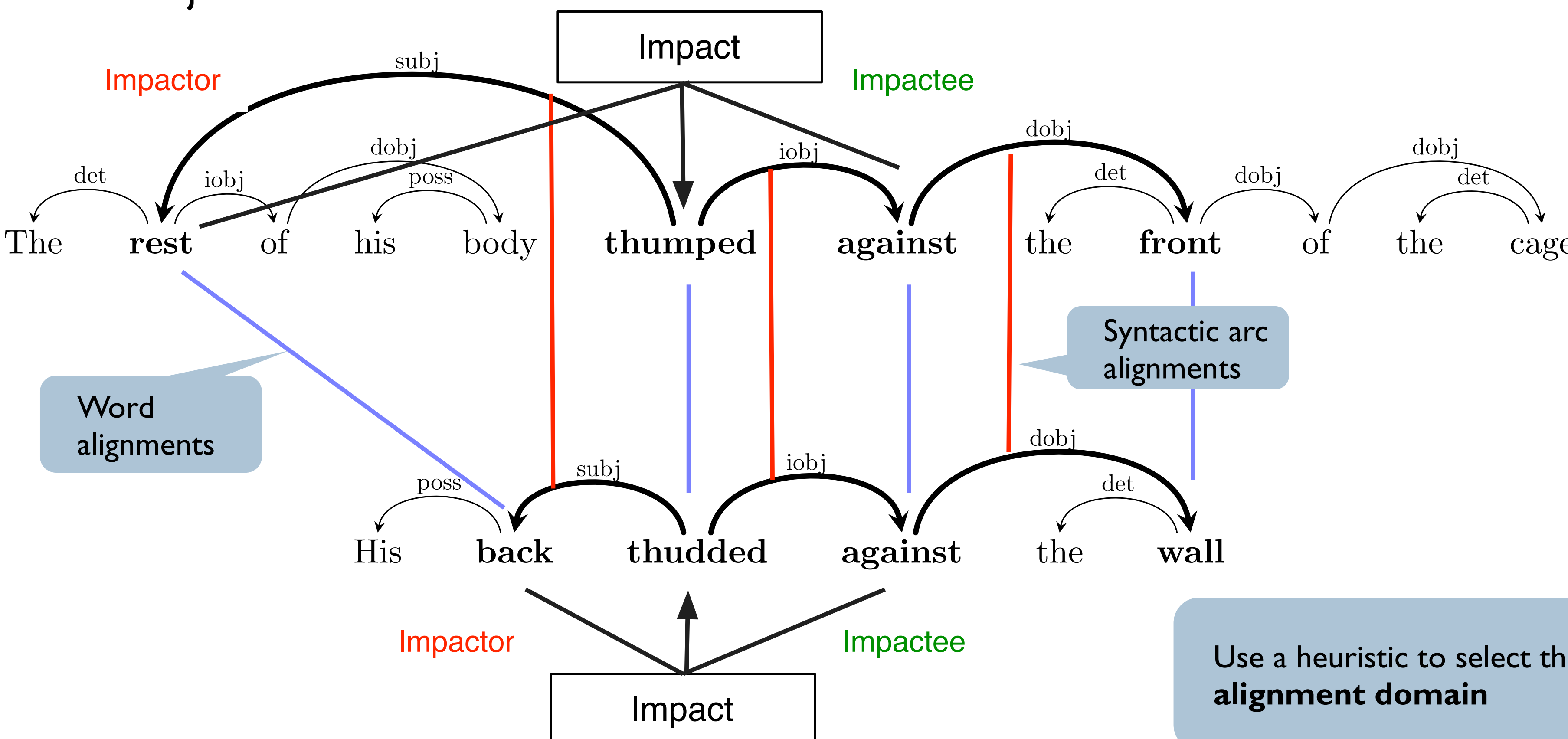
[Furstenau and Lapata, 2009]

Monolingual projection: alignment

- Start with an unlabeled sentence, and a target predicate
- Check a labeled sentence (one by one)
- Find the best alignment, with **Score = Lexical Score + Syntactic Score**
- Project annotation

Using integer linear programming

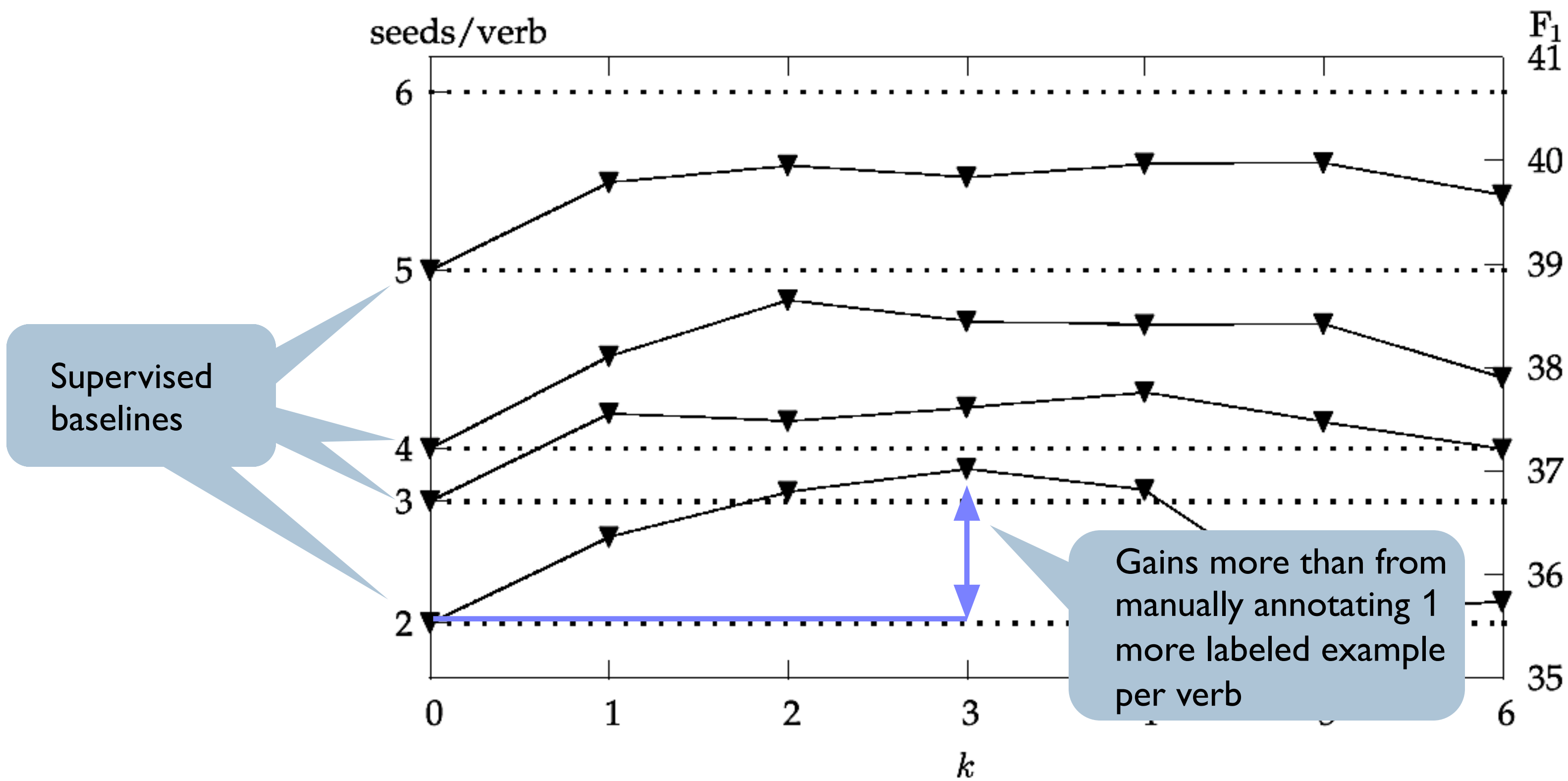
From the one or few *closest labeled sentences*



Use a heuristic to select the **alignment domain**

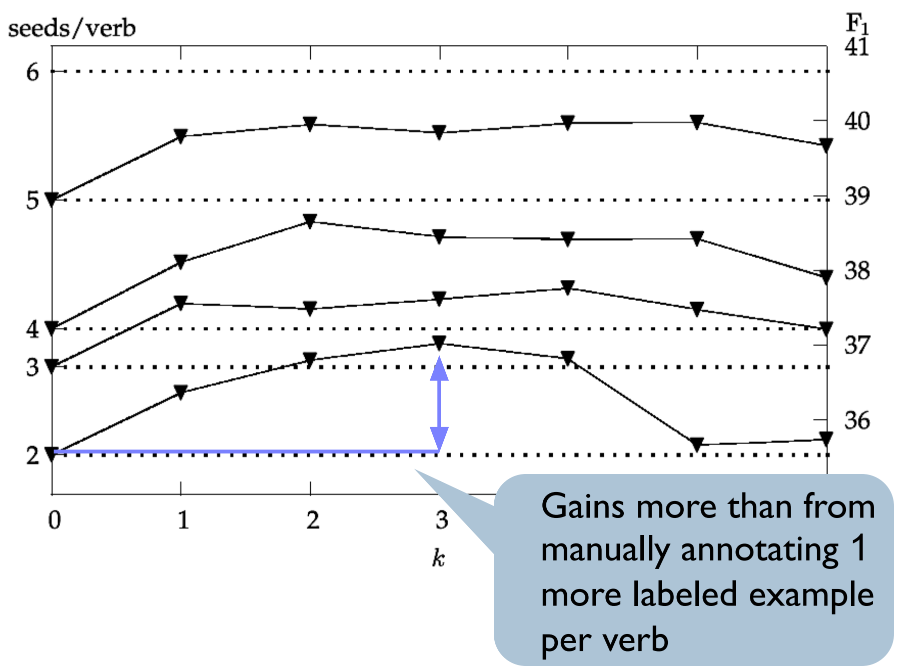
Evaluation

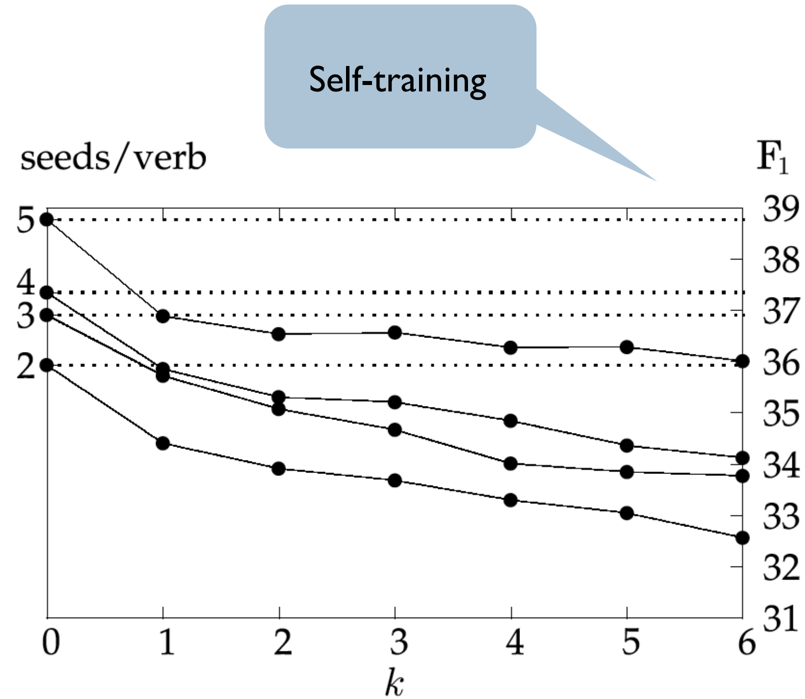
- Evaluation scenario:
  - For a verb, we observe in the labeled training set a few seed examples
  - The seed corpora is expanded by selecting *k* closest unlabeled examples, projecting annotation to them and adding them to training data



Evaluation

- Evaluation scenario:
  - For a verb, we observe in the labeled training set a few seed examples
  - The seed corpora is expanded by selecting *k* closest unlabeled examples, projecting annotation to them and adding them to training data





Outline

- Crosslingual annotation and model transfer
- Semi-supervised learning
  - methods creating surrogate supervision
  - parameter sharing methods
- Unsupervised learning

Reducing sparsity of word representations

- Lexical features are crucial for accurate semantic role labeling
  - However, they are problematic as they are sparse

Especially, if one considers 2nd or higher order features

- Less sparse features capturing lexical information are needed
- Representations can be learnt from unlabeled data in the context of the *language model task*, for example:
  - Brown clusters [Brown et al., 1992]
  - Distributed word representations [Bengio et al., 2003]

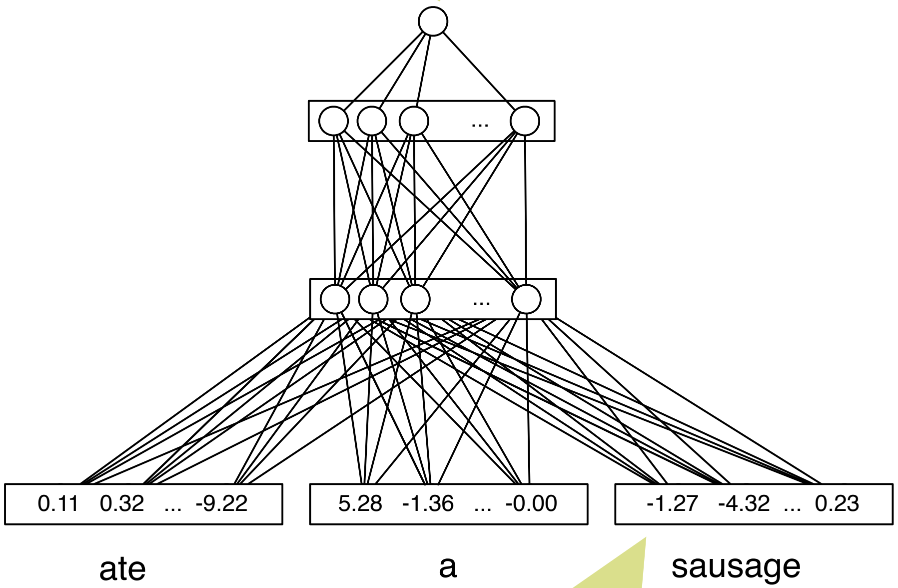
and then used as features in SRL systems

Challenge: they might not capture phenomena relevant to SRL or not have needed granularity.

[Collobert et al., 2011]

Learning lexical representations

Predict if an ngram belongs to the language

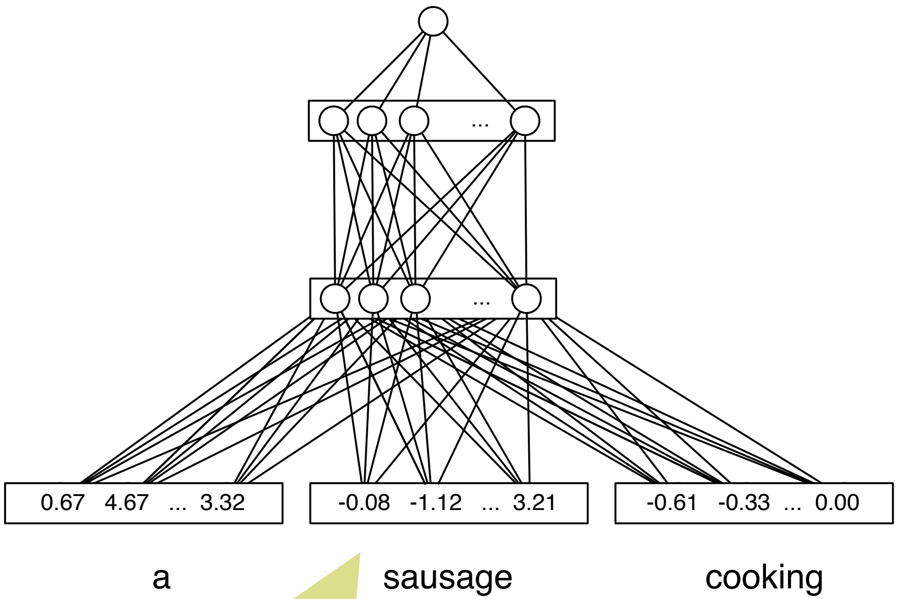


Distributed word representations

Can be trained on large unlabeled texts

Learning lexical representations

Predict a semantic role for the middle word

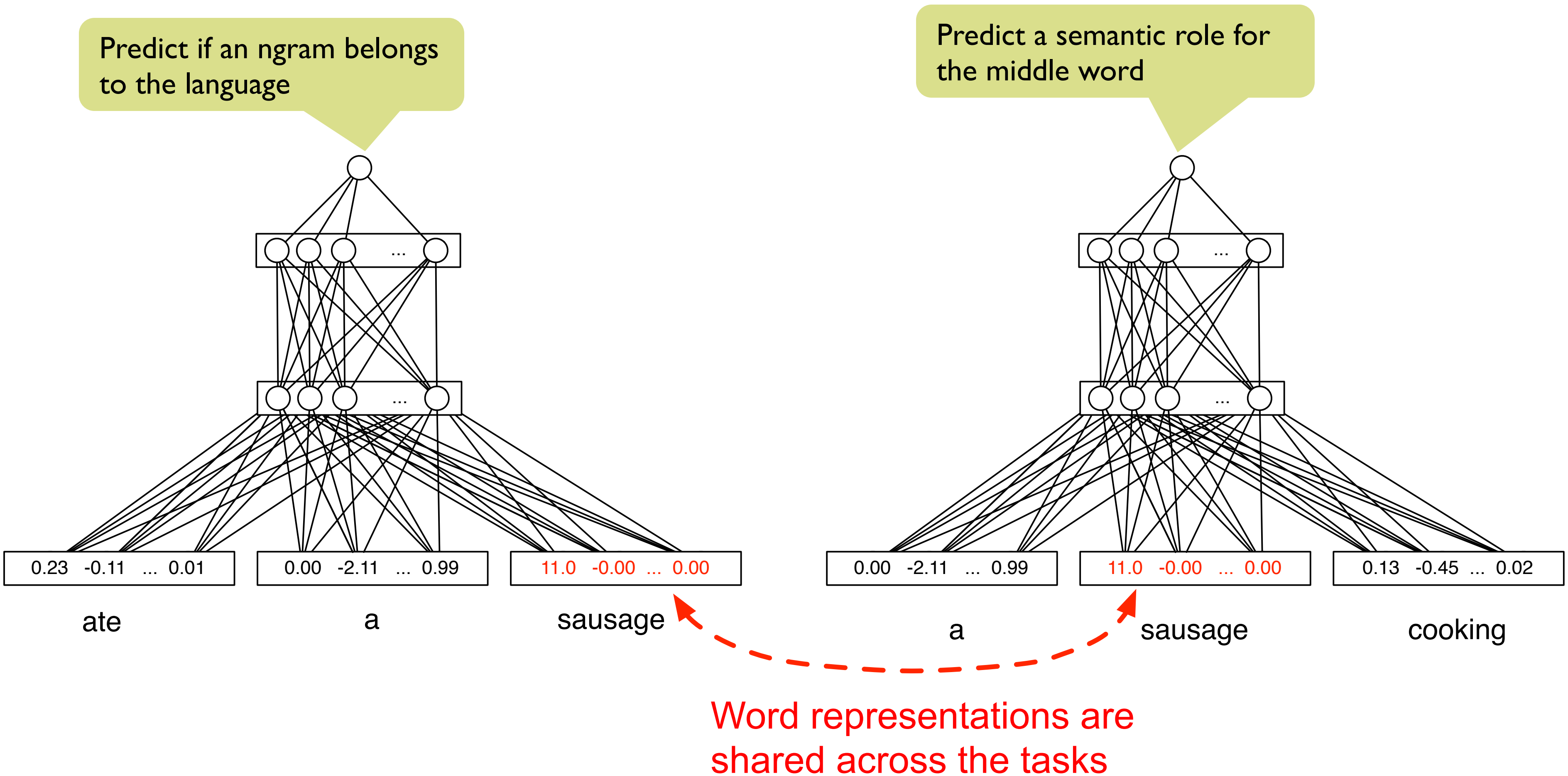


Distributed word representations

Can be trained only on semantically annotated texts

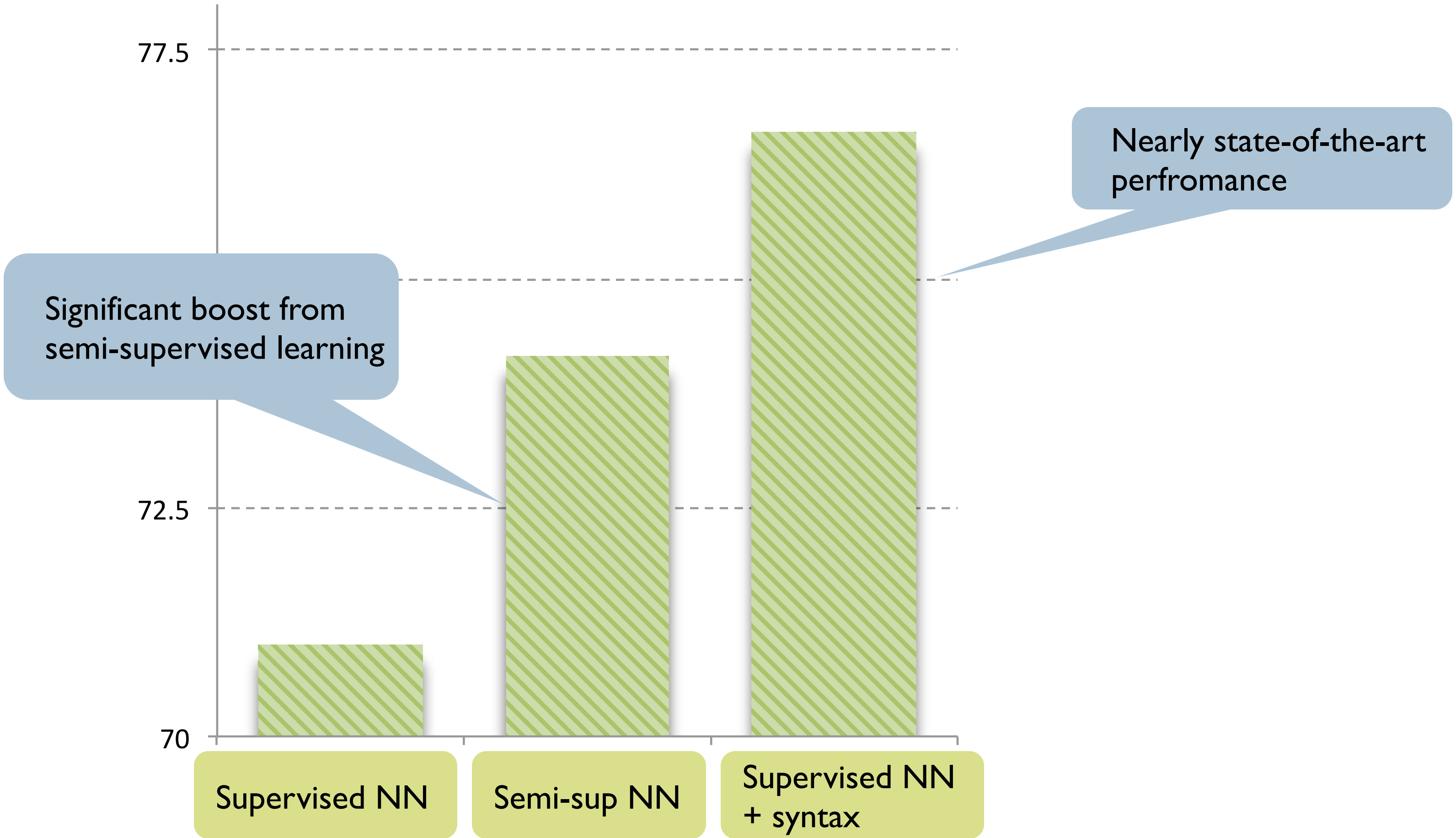
[Collobert et al., 2011]

Learning lexical representations



Share words representations across tasks and learn simultaneously to be useful for both tasks

Evaluation on PropBank



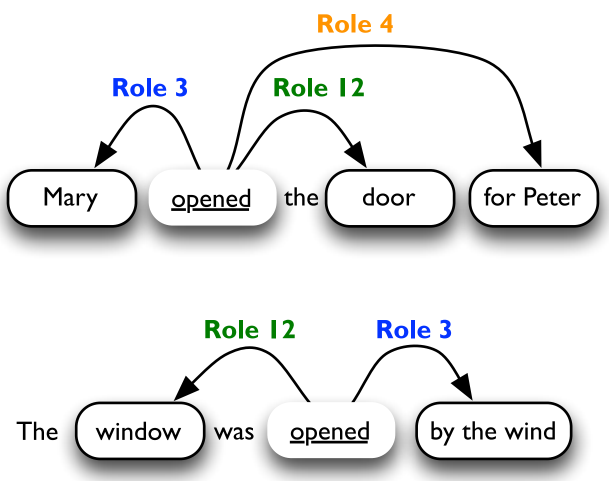
Outline

- Crosslingual annotation and model transfer
- Semi-supervised learning
- Unsupervised learning
  - agglomerative clustering
  - generative modeling

Defining Unsupervised SRL

- Semantic role labeling is typically divided into two sub-tasks:
  - Identification: identification of predicate arguments
  - Labeling: assignment of their sematic roles

Arguably, the easier sub-task, can be handled with heuristics, e.g. [Lang and Lapata, 2010]



Goal: induce semantic roles automatically from unannotated texts

- Equivalent to clustering of argument occurrences (or "coloring" them)

Evaluating Unsupervised SRL

- Before we begin, a note about evaluating unsupervised SRL
- We do not have labels for clusters, so we use standard clustering metrics instead
  - Purity (PU) measures the degree to which each induced role contains arguments sharing the same gold ("true") role

$$PU = \frac{1}{N} \sum_i \max_j |\underbrace{G_j}_{\text{Gold role}} \cap \underbrace{C_i}_{\text{Induced role}}|$$

  - Collocation (CO) 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

Outline

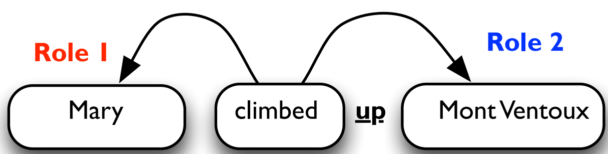
- Crosslingual annotation and model transfer
- Semi-supervised learning
- Unsupervised learning
  - agglomerative clustering [Lang and Lapata, 2011b]
  - generative modeling [Titov and Klementev 2012]

Eariler methods [Swier and Stevenson, 2004; Grenager and Manning 2006] relied on strong linguistic priors / resources for the language in question

[Lang and Lapata, 2011b]

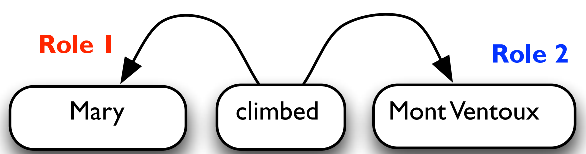
Role Labeling as Clustering of Argument Keys

- 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

Role Labeling via "Split-Merge" Clustering

- Agglomerative clustering of arguments
  - Start with each argument key in its own cluster (high purity, low collocation)
  - Merge clusters together to improve collocation
- For a pair of clusters score
  - whether a pair contains lexically similar arguments
  - whether arguments have similar parts of speech
  - whether the constraint that arguments in a clause should be in different roles is satisfied

    John <u>taught</u> students math

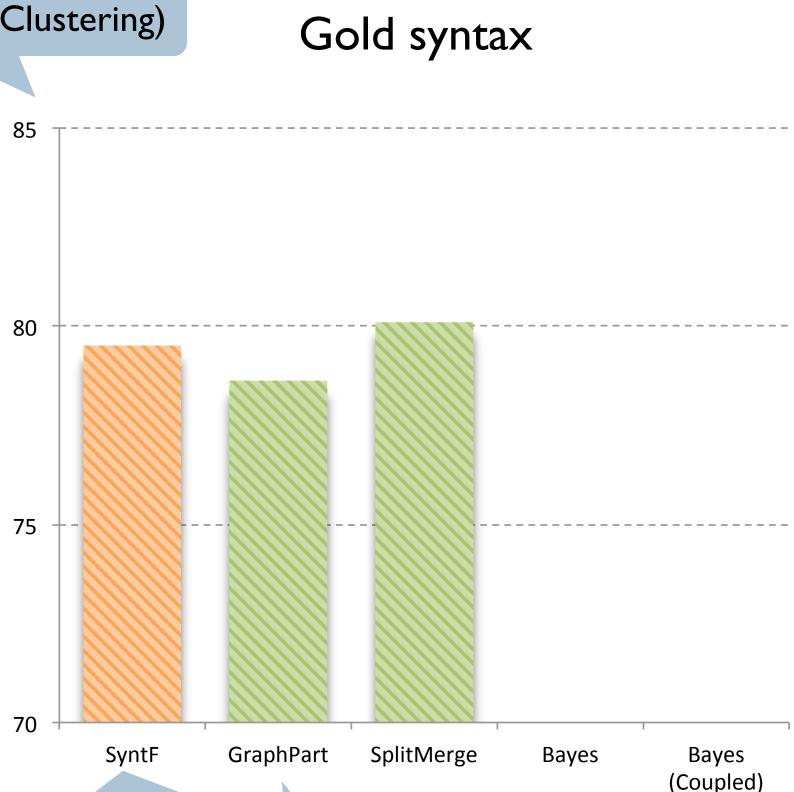
More important clustering decisions are done early

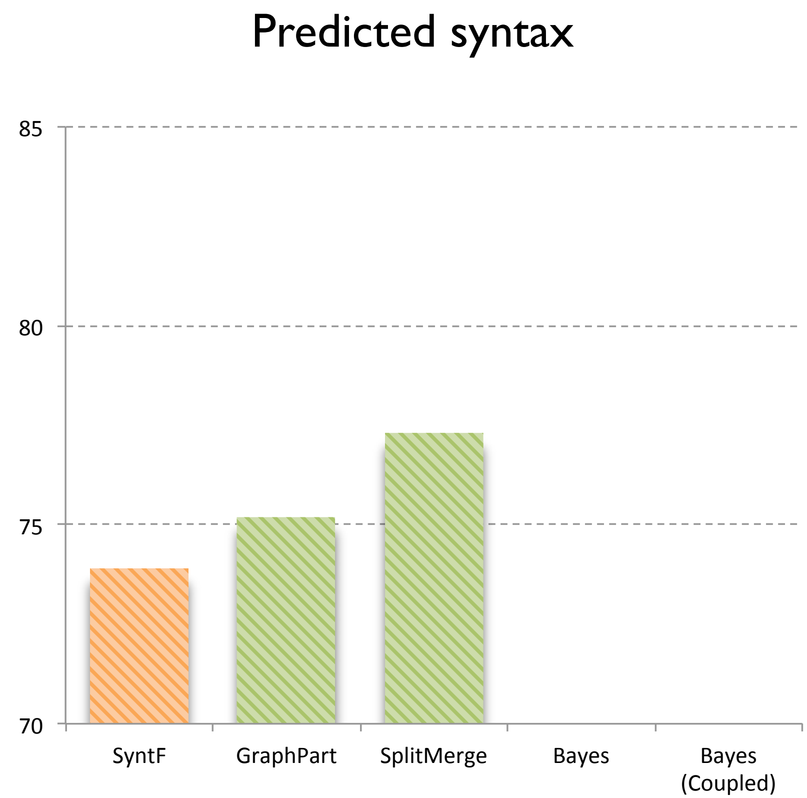
- Prioritization
  - Instead of greedily choosing the highest scoring pair at each step, start with larger clusters and select best match for each of them

[Lang and Lapata, 2011b]

PropBank (CoNLL 08)

F1 (Clustering)





Syntactic baseline

A graph-based method (Lang and Lapata, 2011a)

Outline

- Crosslingual annotation and model transfer
- Semi-supervised learning
- Unsupervised learning
  - agglomerative clustering
  - generative modeling

A Bayesian model for role labeling

- Idea: propose a generative model for inducing argument clusters
  - As before, clusters are of argument keys, not argument occurrences
- Learning signals are similar to Lang and Lapata (2011a, 2011b), e.g.
  - Selection preferences

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

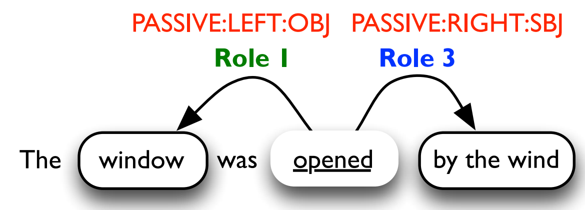
  - Duplicate roles are unlikely to occur. E.g. this clustering is a bad idea:

    John taught students math

    GB-criterion

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

A Bayesian model for role labeling



At least one argument

Draw first argument

Continue generation

Draw more arguments

Decide on arg key clustering

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

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

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

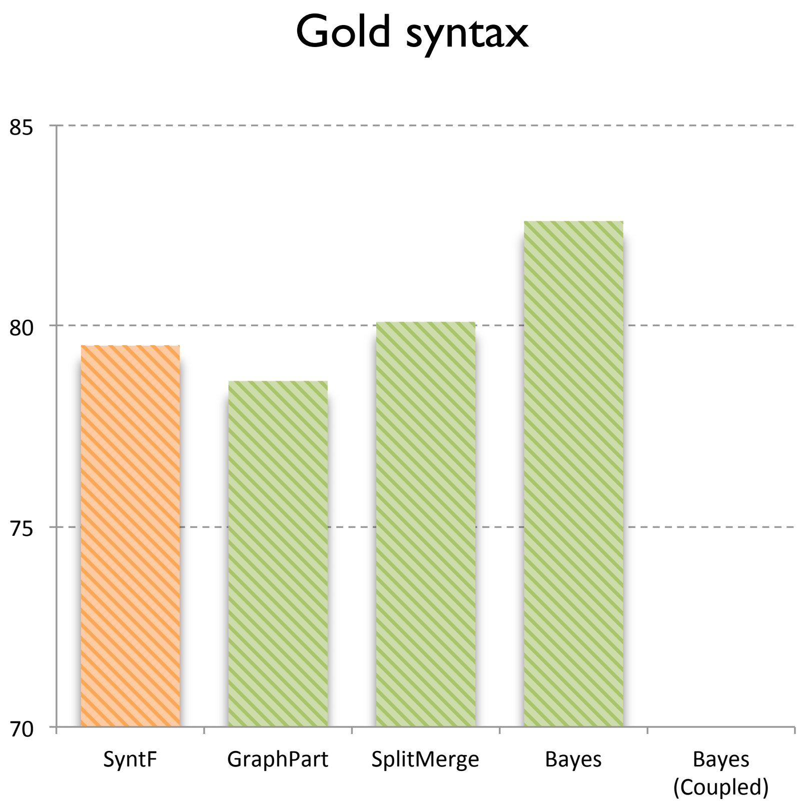
Draw argument key

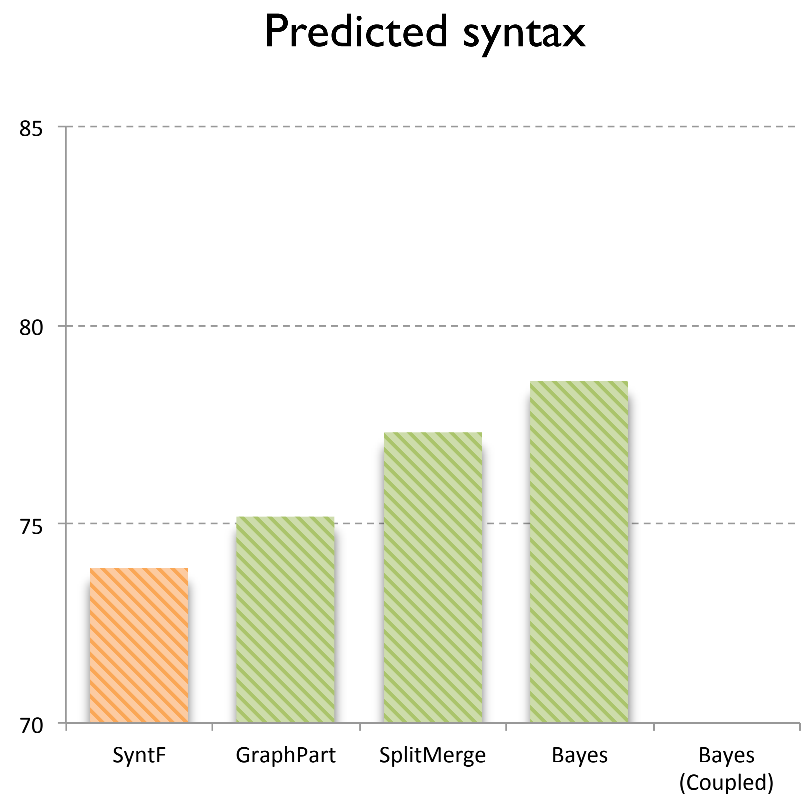
Draw argument filler

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

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)





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

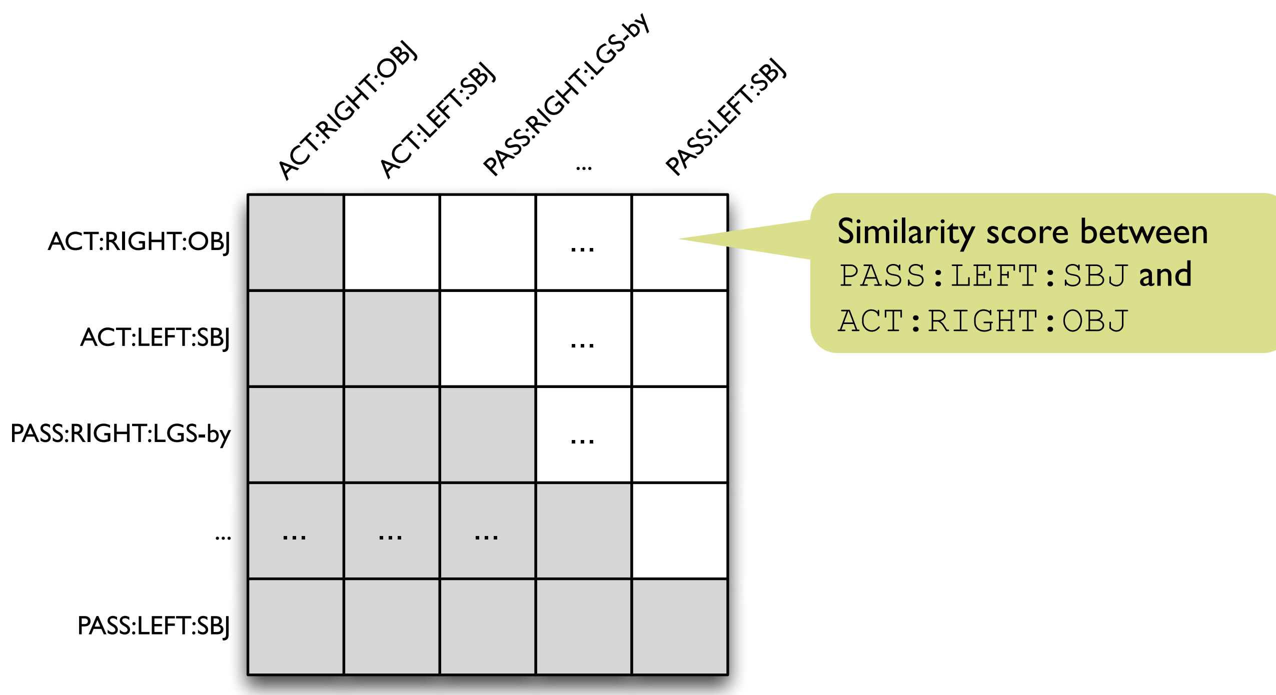
John gave the book to Mary	*vs*	John gave Mary the book
Mike threw the ball to me	*vs*	Mike threw me the ball

Dative alternation

- Can we share this information across verbs?

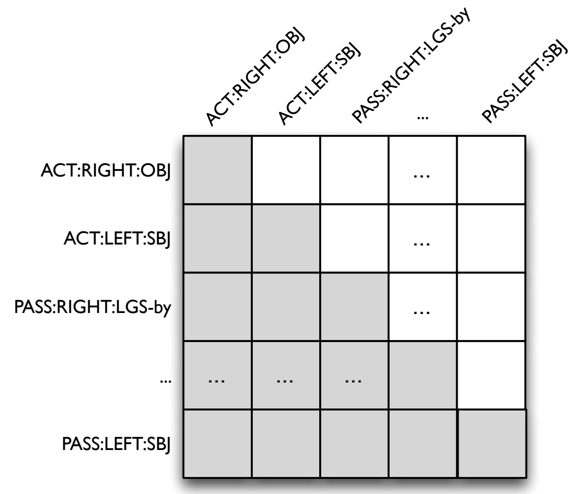
A Bayesian model for role labeling

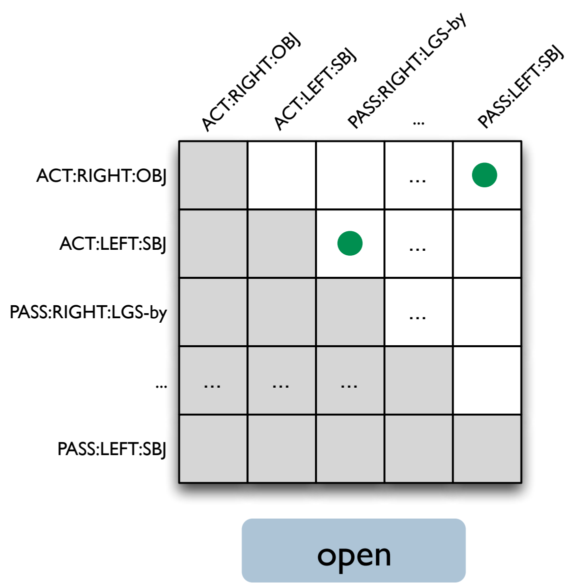
- Idea: keep track of how likely a pair of argument keys should be clustered
  - Define a similarity matrix (or similarity graph)

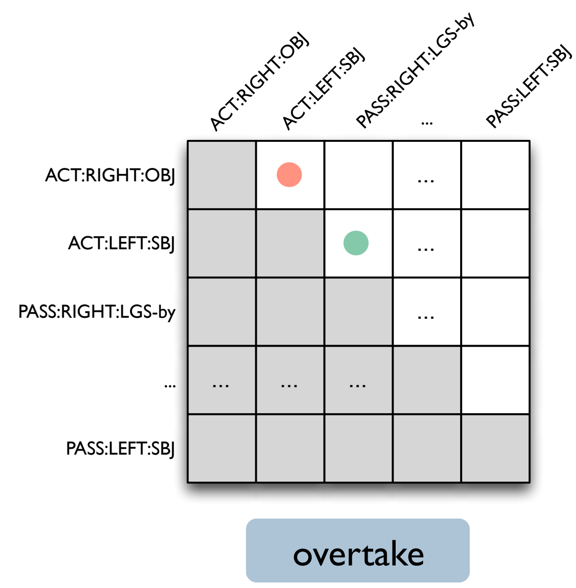


[Titov and Klementiev, 2012a]

A Bayesian model for role labeling

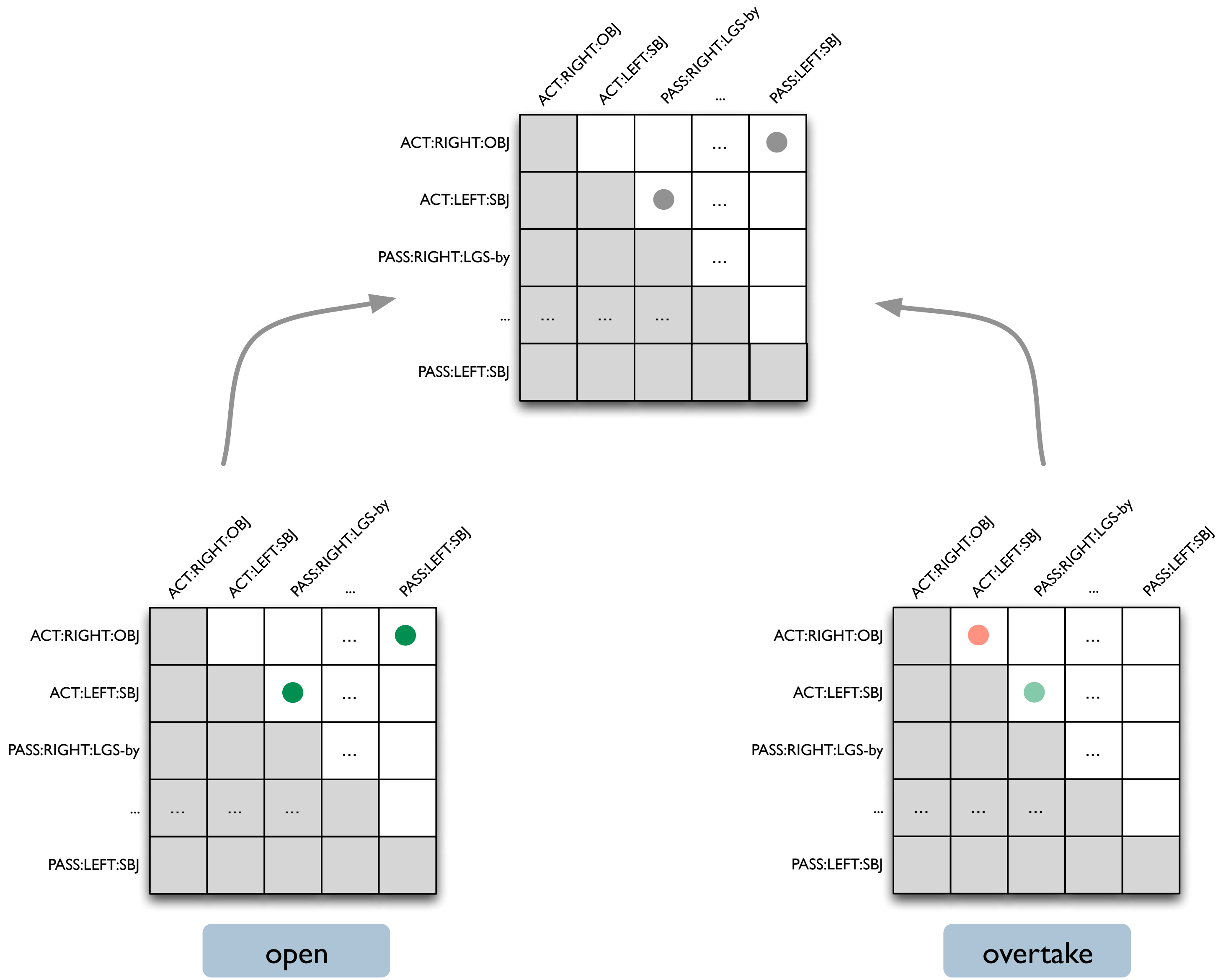






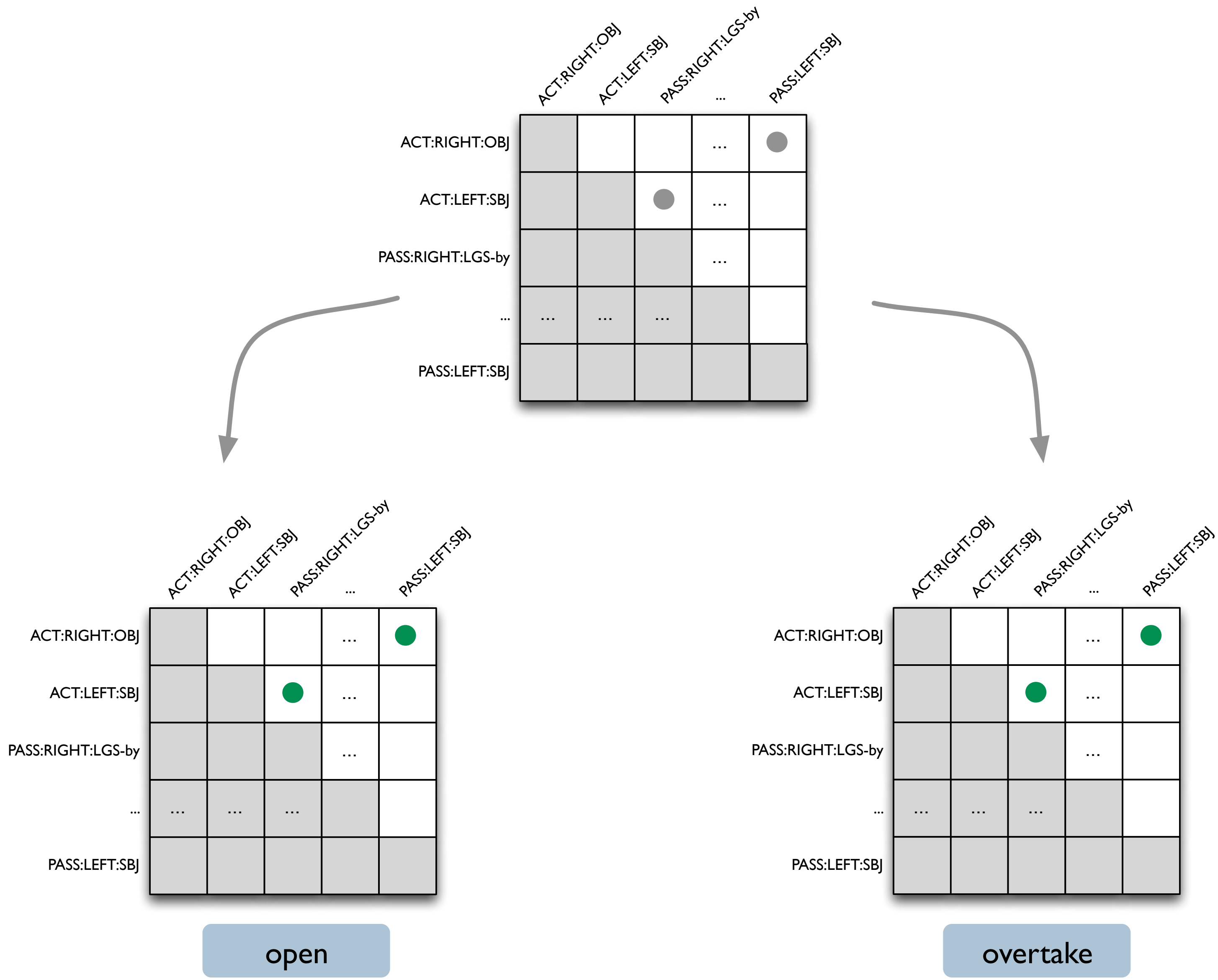
[Titov and Klementiev, 2012a]

A Bayesian model for role labeling



[Titov and Klementiev, 2012a]

A Bayesian model for role labeling



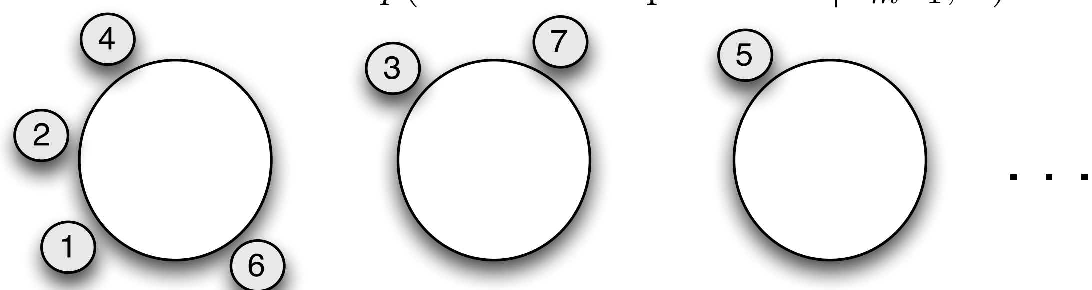
A formal way to encode this: dd-CRP

- 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

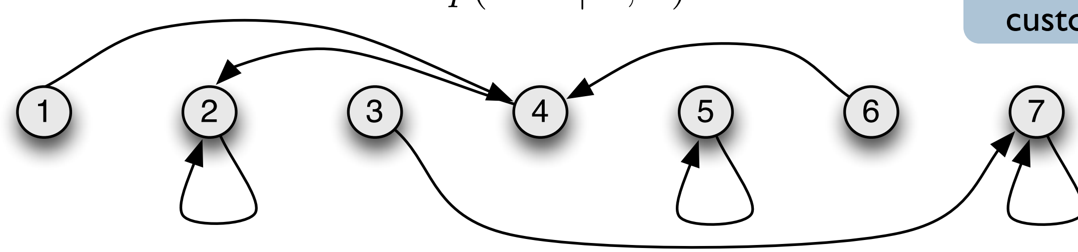


- An extension is distance-dependent CRP (dd-CRP):
  - m-th customer chooses a *customer* to sit with according to:

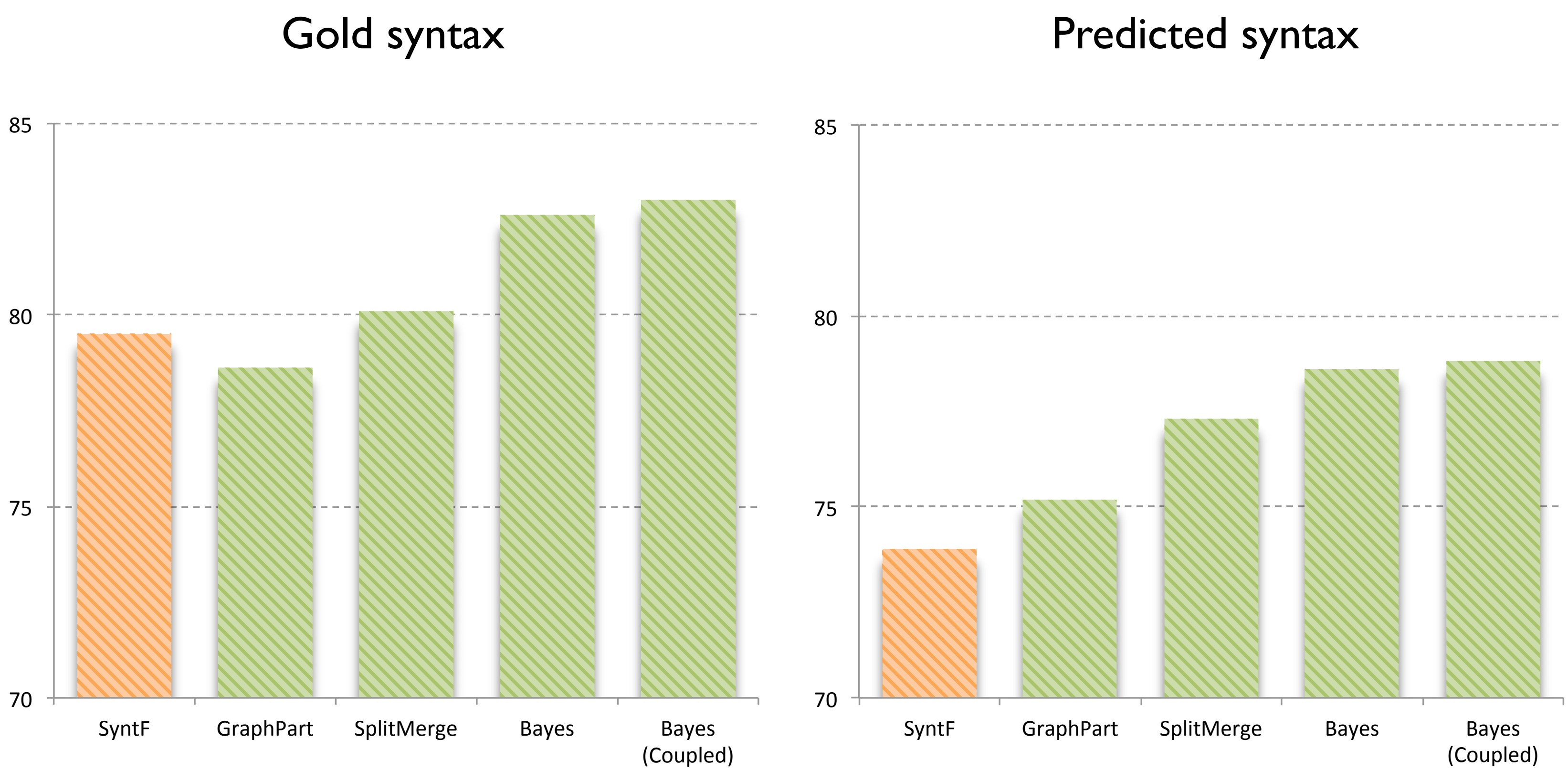
Entire similarity graph

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

Similarity between customers m and j



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

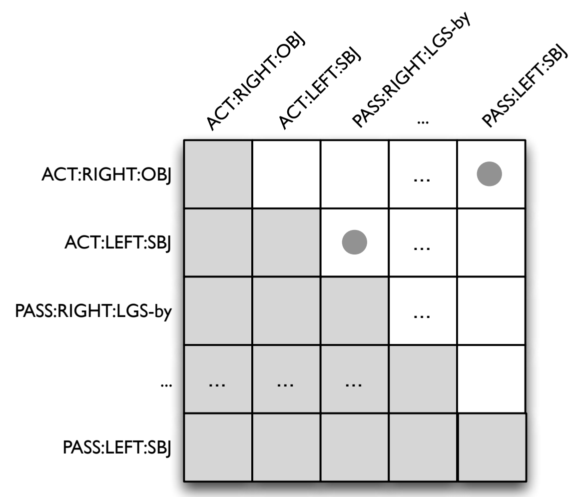
  Martha carved the baby a doll
- Dativization

  I gave the book to Mary

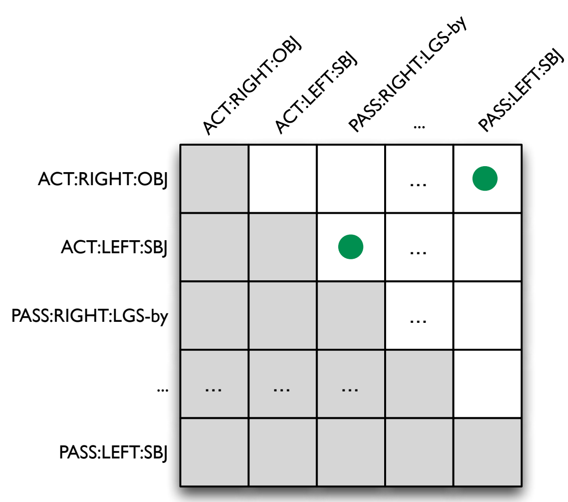
  I gave Mary the book
- Recovery of unnecessary splits introduced by argument keys

[Titov and Klementiev, 2012a]

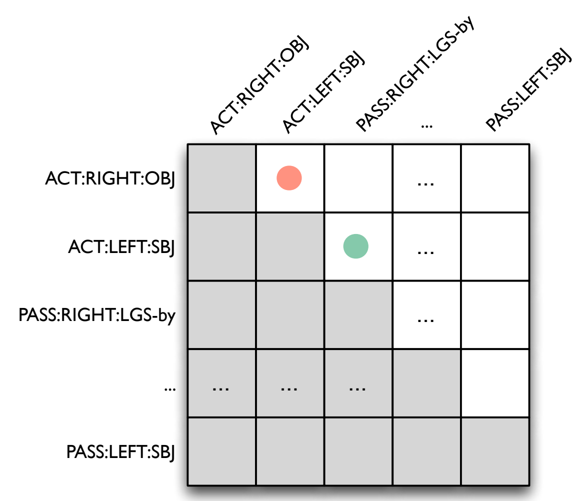
A Bayesian model for role labeling



supervised data



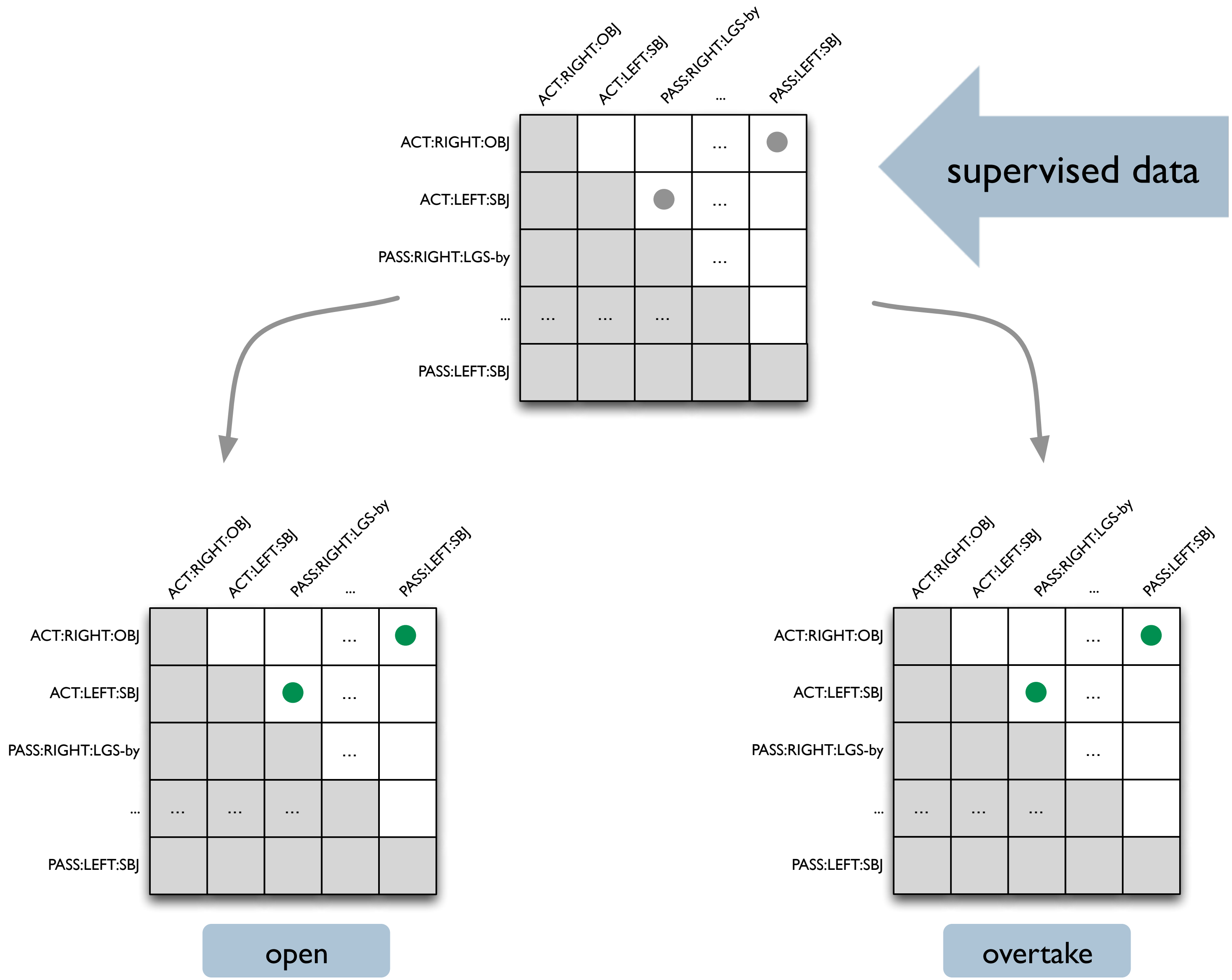
open



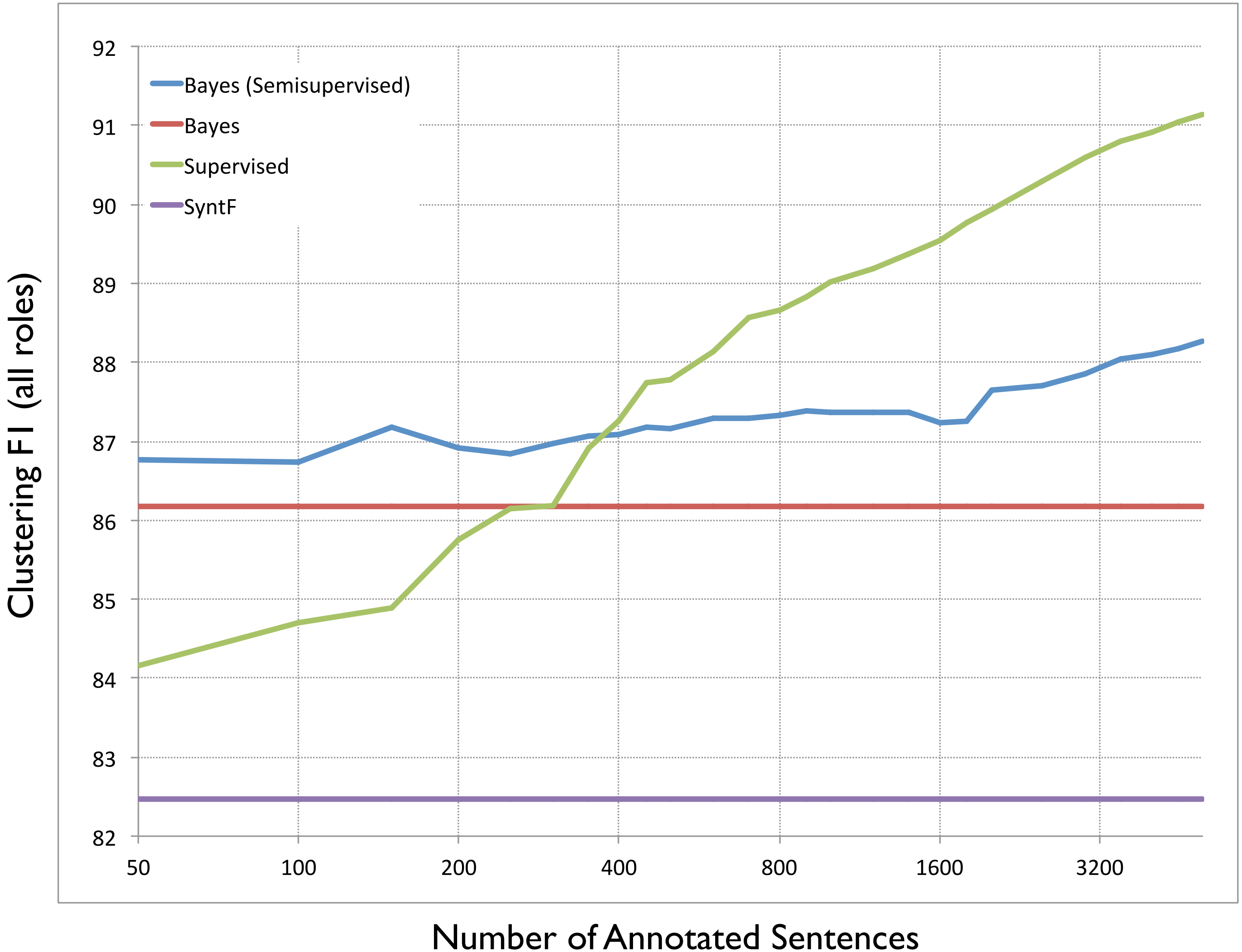
overtake

[Titov and Klementiev, 2012b]

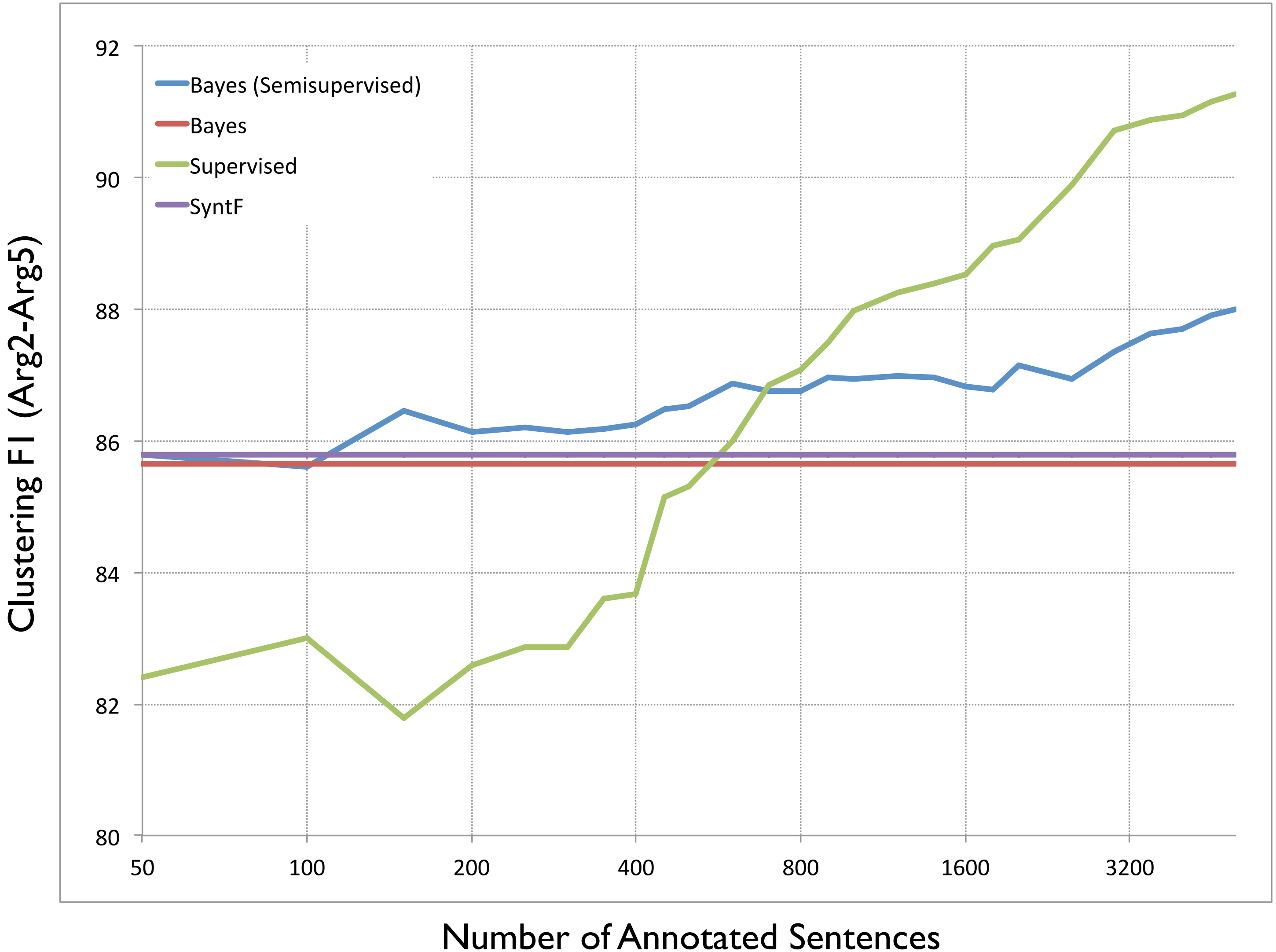
A Bayesian model for role labeling



PropBank (CoNLL 09)

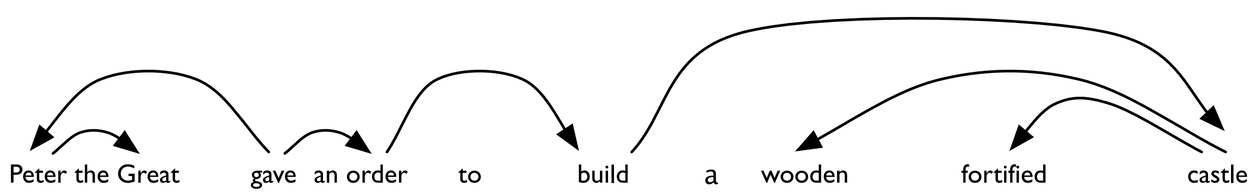


PropBank (CoNLL 09)



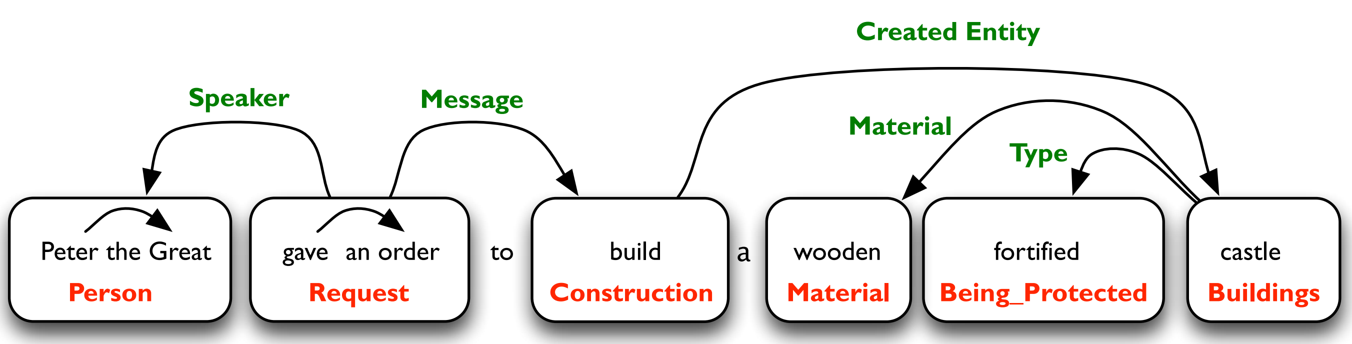
Generalization of the role induction model

- 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)



Conclusions

- We looked in examples of key directions in exploiting unlabeled data and cross-lingual correspondences
  - a lot of relevant recent work has not been covered
- Still a new direction with a lot of ongoing work
  - research in the related area of information extraction should also closely watched
- Many thanks to Alex Klementiev, Hagen Furstenau, Sebastian Pado for their help

References

Y. Bengio, R. Ducharme, 2003. P. Vincent and C. Janvin. A neural probabilistic language model. JMLR.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa. 2011. Natural Language Processing (Almost) from Scracth. JMLR.

P.F. Brown, R.L.Mercer, V.J. Della Pietra, and J.C. Lai. 2009. Class-based n-gram models of natural language. Computational Linguistics, 1992.

H. Fürstenau and M. Lapata. Semisupervised semantic role labeling. EACL.

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

S. He and D. Gildea. 2006. Self-training and Co-training for Semantic Role Labeling: Primary Report", TR 891 , Department of Computer Science, University of Rochester.

R. Johansson and P. Nugues. 2006. A FrameNet-based Semantic Role Labeler for Swedish. COLING/ACL.

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.

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.

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

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. COLING/ACL.

S. Pado and G. Pitel. 2007. Annotation précise du français en sémantique de rôles par projection cross-linguistique. TALN.

References

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.

S. Petrov, D. Das, and R. McDonald. 2012. A universal part-of-speech tagset. LREC.

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

S. Pradhan, W. Ward, and J. Martin, 2008. Towards robust semantic role labeling. Computational Linguistics.

R. Swier and S. Stevenson. Unsupervised Semantic Role Labeling. EMNLP.

O. Tackstrom, R. McDonald, and J. Uszkoreit, 2012. Cross-lingual word clusters for direct transfer of linguistic structure. NAACL.

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.

D. Zeman and P. Resnik. 2008 Crosslanguage parser adaptation between related languages. IJCNLP workshop.