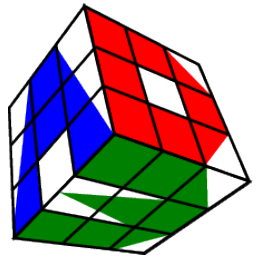


AD^3: A Fast Decoder for Structured Prediction

André Martins

priberam

instituto de telecomunicações

NIPS Workshop on Modern ML+NLP—Montréal, 12/12/14

Talking About Modern ML and NLP...

(Charlie Chaplin, "Modern Times," 1936)

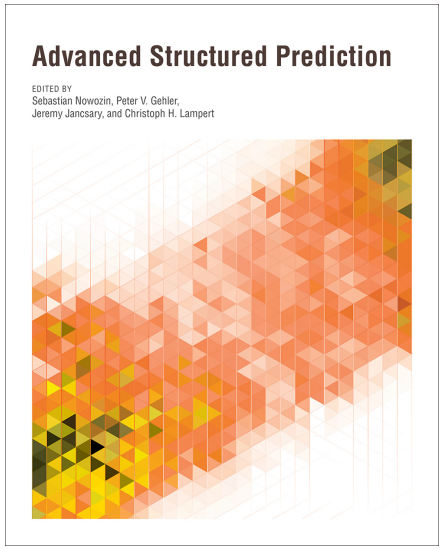
Talking About Modern ML and NLP...

(Charlie Chaplin, "Modern Times," 1936)

Key idea of this talk: connect small gears to build a larger machine.

Literature Pointers

- André F. T. Martins.
  "AD^3: A Fast Decoder for Structured Prediction."
  Book chapter of *Advanced Structured Prediction*, Sebastian Nowozin, Peter V. Gehler, Jeremy Jancsary, and Christoph H. Lampert (Editors), MIT Press, 2014.



- A. Martins, M. Figueiredo, P. Aguiar, N. Smith, E. Xing.
  "AD3: Alternating Directions Dual Decomposition for MAP Inference in Graphical Models."
  JMLR, 2015 (to appear).

More details: EMNLP 2014 tutorial on "LP Decoders for NLP."

Structured Prediction and NLP

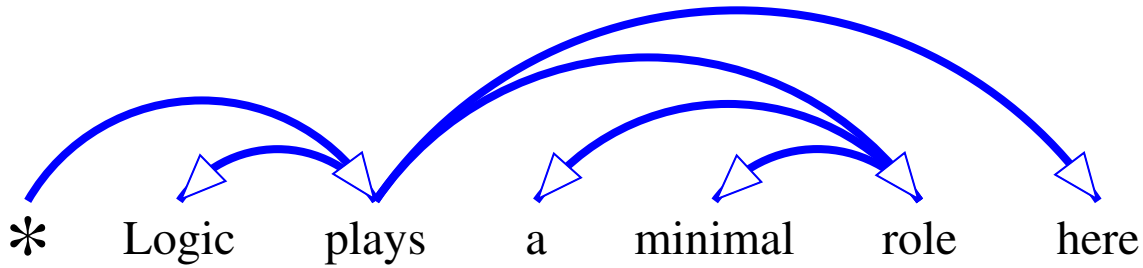
Structured prediction: a machine learning framework for predicting structured, constrained, and interdependent outputs

NLP deals with *structured* and *ambiguous* textual data:

- machine translation
- speech recognition
- syntactic parsing
- semantic parsing
- information extraction
- ...

Dependency Parsing

Map **sentences** to their **syntactic structure.**



- A lexicalized syntactic formalism
- Grammar functions represented as lexical relationships (dependencies)

(Eisner, 1996; McDonald et al., 2005; Nivre et al., 2006; Koo et al., 2007)

Multi-Document Summarization

Map a set of related **documents** to a brief **summary.**

CNN

Obama hopes for 'continued progress' in Myanmar

STORY HIGHLIGHTS

- Obama meets with pro-democracy icon Aung San Suu Kyi and Myanmar's president
- He's the first sitting U.S. president to visit Myanmar, also known as Burma
- Obama encourages the country to continue a "remarkable journey"
- He also visits Cambodia to meet the prime minister and attend the East Asia Summit

(CNN) -- Barack Obama met with Nobel Peace Prize winner Aung San Suu Kyi at her home in Myanmar on Monday, praising her "courage and determination" during a historic visit to the once repressive and secretive country.

The first sitting U.S. president to visit Myanmar, Obama urged its leaders, who have embarked on a series of far-reaching political and economic reforms since 2011, not to extinguish the "flickers of progress that we have seen."

Obama said that his visit to the lakeside villa where the pro-democracy icon spent years under house arrest marked a new chapter between the two countries.

"Here, through so many difficult years, is where she has displayed such unbreakable courage and determination," Obama told reporters, standing next to his fellow Nobel peace laureate. "It is here where she showed that human freedom and human dignity cannot be denied."

Myanmar Obama visit

The country, which is also known as Burma, was ruled by military leaders until early 2011 and for decades was politically and economically cut off from the rest of the world.

Suu Kyi acknowledged that Myanmar's opening up would be difficult.

The New York Times

YANGON, Myanmar — President Obama journeyed to this storied tropical outpost of pagodas and jungles on Monday to "extend the hand of friendship" as a land long tormented by repression and poverty begins to throw off military rule and emerge from decades of isolation.

The visit was intended to show support for the reforms put in place by Thein Sein's government since the end of military rule in November 2010.

Activists have warned that the visit may be too hasty - political prisoners remain behind bars and ethnic conflicts in border areas are unresolved.

BBC NEWS

Multi-Document Summarization

Map a set of related **documents** to a brief **summary.**

CNN

Obama hopes for 'continued progress' in Myanmar

STORY HIGHLIGHTS

- Obama meets with pro-democracy icon Aung San Suu Kyi and Myanmar's president
- He's the first sitting U.S. president to visit Myanmar, also known as Burma
- Obama encourages the country to continue a "remarkable journey"
- He also visits Cambodia to meet the prime minister and attend the East Asia Summit

(CNN) -- Barack Obama met with Nobel Peace Prize winner Aung San Suu Kyi at her home in Myanmar on Monday, praising her "courage and determination" during a historic visit to the once repressive and secretive country.

The first sitting U.S. president to visit Myanmar, Obama urged its leaders, who have embarked on a series of far-reaching political and economic reforms since 2011, not to extinguish the "flickers of progress that we have seen."

Obama said that his visit to the lakeside villa where the pro-democracy icon spent years under house arrest marked a new chapter between the two countries.

"Here, through so many difficult years, is where she has displayed such unbreakable courage and determination," Obama told reporters, standing next to his fellow Nobel peace laureate. "It is here where she showed that human freedom and human dignity cannot be denied."

Myanmar Obama visit

The country, which is also known as Burma, was ruled by military leaders until early 2011 and for decades was politically and economically cut off from the rest of the world.

Suu Kyi acknowledged that Myanmar's opening up would be difficult.

The New York Times

YANGON, Myanmar — President Obama journeyed to this storied tropical outpost of pagodas and jungles on Monday to "extend the hand of friendship" as a land long tormented by repression and poverty begins to throw off military rule and emerge from decades of isolation.

The visit was intended to show support for the reforms put in place by Thein Sein's government since the end of military rule in November 2010.

Activists have warned that the visit may be too hasty - political prisoners remain behind bars and ethnic conflicts in border areas are unresolved.

BBC NEWS

This talk: AD^3

A new dual decomposition algorithm, suitable for many scenarios encountered in NLP and IR.

This talk: AD^3

A new dual decomposition algorithm, suitable for many scenarios encountered in NLP and IR.

- Problems with factor graph representations

This talk: AD^3

A new dual decomposition algorithm, suitable for many scenarios encountered in NLP and IR.

- Problems with factor graph representations
- Statements in FOL

This talk: AD^3

A new dual decomposition algorithm, suitable for many scenarios encountered in NLP and IR.

- Problems with factor graph representations
- Statements in FOL
- Budget/knapsack constraints

This talk: AD^3

A new dual decomposition algorithm, suitable for many scenarios encountered in NLP and IR.

- Problems with factor graph representations
- Statements in FOL
- Budget/knapsack constraints
- Combination of structured models

This talk: AD^3

A new dual decomposition algorithm, suitable for many scenarios encountered in NLP and IR.

- Problems with factor graph representations
- Statements in FOL
- Budget/knapsack constraints
- Combination of structured models

We'll illustrate with two applications:

1. **Dependency Parsing (Syntax and Semantics)**
2. **Compressive Summarization**

This talk: AD^3

A new dual decomposition algorithm, suitable for many scenarios encountered in NLP and IR.

- Problems with factor graph representations
- Statements in FOL
- Budget/knapsack constraints
- Combination of structured models

We'll illustrate with two applications:

1. **Dependency Parsing (Syntax and Semantics)**
2. **Compressive Summarization**

Could be a great fit to many other applications!!

Outline

Outline

Structured Prediction

- For each $x \in \mathcal{X}$: a **large** set of candidate outputs $\mathcal{Y}(x)$
- **Decoding problem:**

$$\widehat{y} = \arg \max_{y \in \mathcal{Y}(x)} F_{\mathbf{w}}(x, y)$$

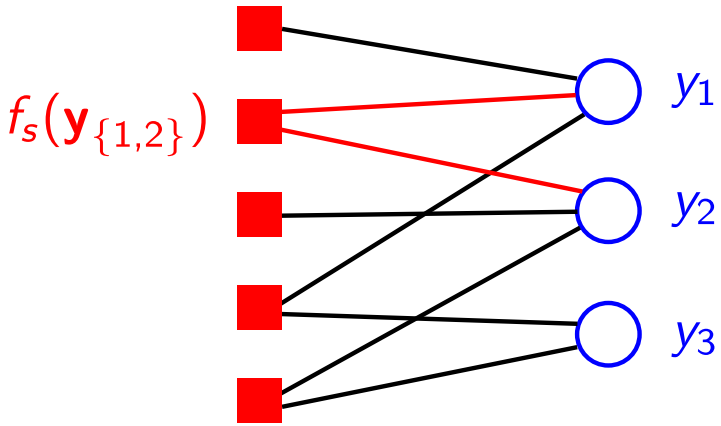
Structured Prediction

- For each $x \in \mathcal{X}$: a **large** set of candidate outputs $\mathcal{Y}(x)$
- **Decoding problem:**

$$\widehat{y} = \arg \max_{y \in \mathcal{Y}(x)} F_{\mathbf{w}}(x, y)$$

- **Key assumption:** $F_{\mathbf{w}}$ decomposes into (overlapping) *parts*

$$F_{\mathbf{w}}(x, y) := \sum_s f_s(\mathbf{y}_s)$$



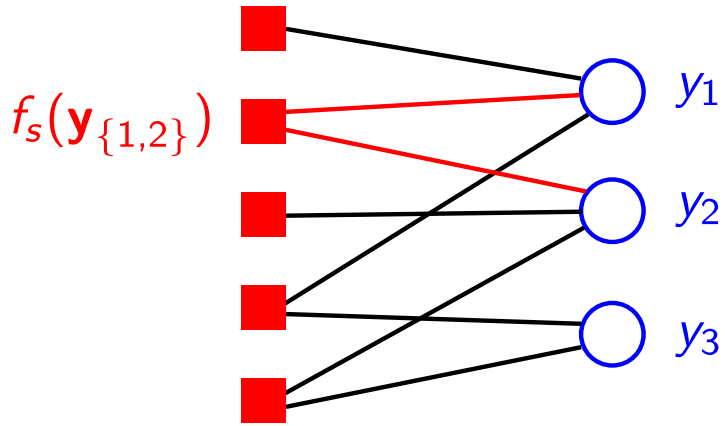
Structured Prediction

- For each $x \in \mathcal{X}$: a **large** set of candidate outputs $\mathcal{Y}(x)$
- **Decoding problem:**

$$\widehat{y} = \arg \max_{y \in \mathcal{Y}(x)} F_{\mathbf{w}}(x, y)$$

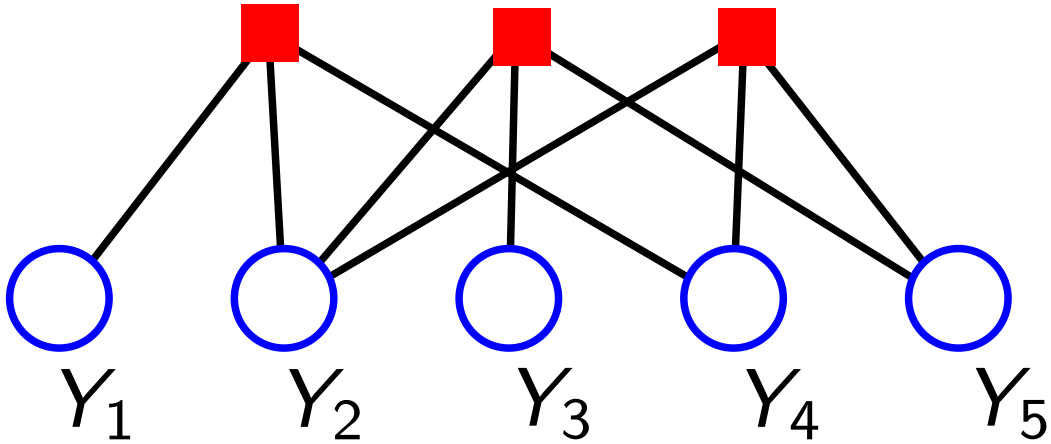
- **Key assumption:** $F_{\mathbf{w}}$ decomposes into (overlapping) *parts*

$$F_{\mathbf{w}}(x, y) := \sum_s f_s(\mathbf{y}_s)$$

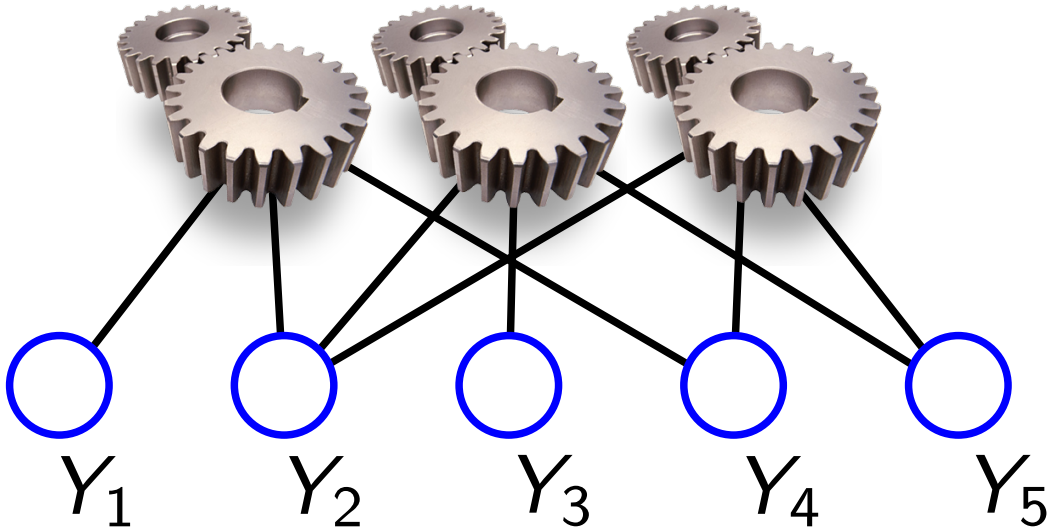


- **Examples:** HMMs, CRFs, PCFGs, general graphical models

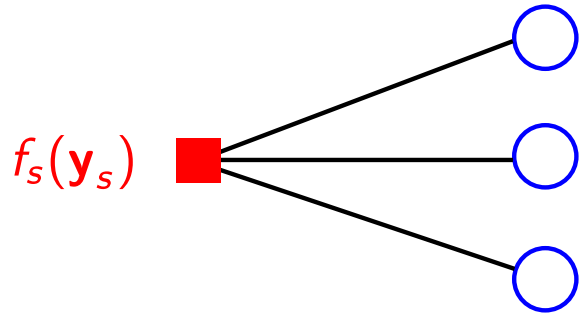
Factors as Machines



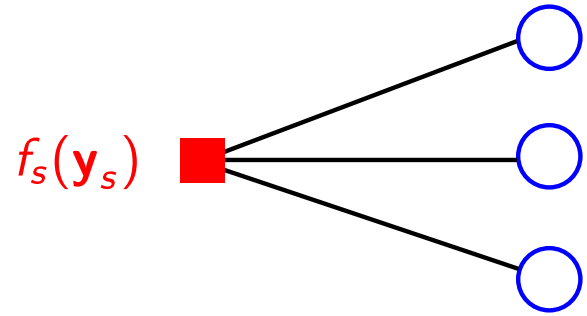
Factors as Machines



Three Kinds of Factors

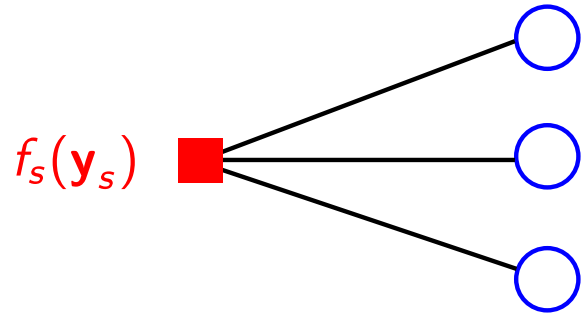


Three Kinds of Factors



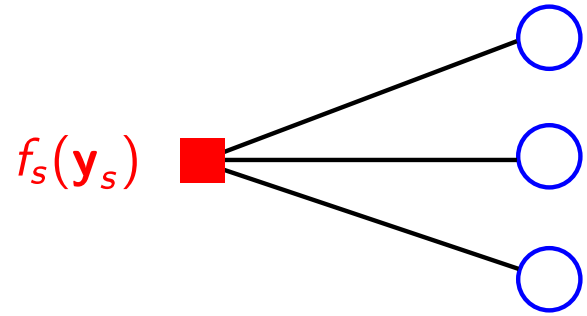
1. **Dense:** $O(\exp(|\partial(s)|))$ degrees of freedom

Three Kinds of Factors



1. **Dense:** $O(\exp(|\partial(s)|))$ degrees of freedom
2. **Structured factors:** $f_s(\mathbf{y}_s)$ is structured itself.
   E.g.: a sequence model, subtree, head automaton, PCFG (Smith and Eisner, 2008; Koo et al., 2010; Rush et al., 2010)

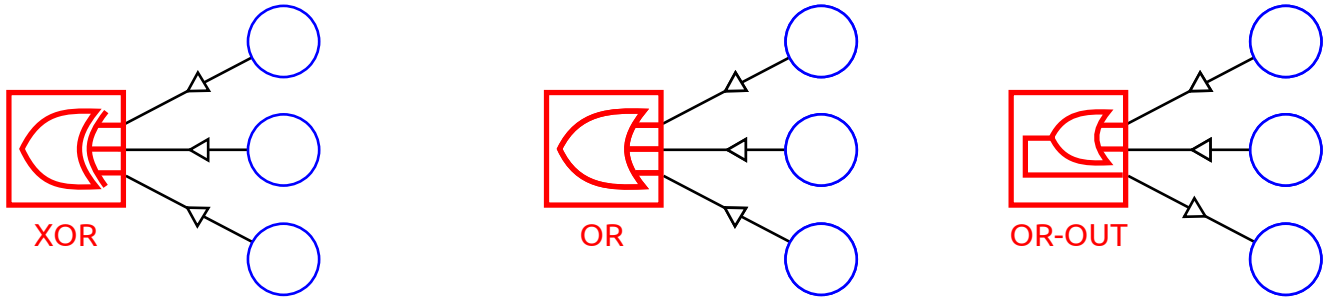
Three Kinds of Factors



1. **Dense:** $O(\exp(|\partial(s)|))$ degrees of freedom
2. **Structured factors:** $f_s(\mathbf{y}_s)$ is structured itself.
   E.g.: a sequence model, subtree, head automaton, PCFG (Smith and Eisner, 2008; Koo et al., 2010; Rush et al., 2010)
3. **Hard constraint factors:**

$$f_s(\mathbf{y}_s) := \begin{cases} 0, & \text{if } \mathbf{y}_s \in \mathcal{Y}_s \\ -\infty, & \text{otherwise.} \end{cases}$$

Example: Hard Constraint Factors

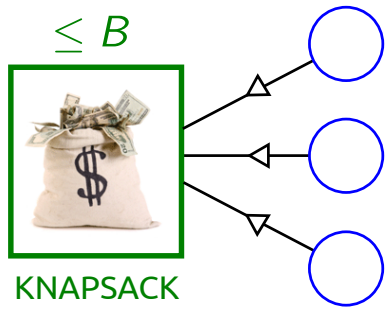


Logic factors: can express arbitrary FOL constraints

- Applications: Markov logic networks (Richardson and Domingos, 2006), constrained conditional models (Roth and Yih, 2004)

Knapsack factor: can express budget constraints

- Applications: summarization, diversity problems,...



(Martins et al., 2011b, 2015; Almeida and Martins, 2013)

Global/Local Decoding

Global decoding problem:

$$\widehat{y} = \arg \max_{y \in \mathcal{Y}(x)} \sum_{s} f_s(\mathbf{y}_s)$$

Solvable with dynamic programming when the graph is a tree...

Global/Local Decoding

Global decoding problem:

$$\widehat{y} = \arg \max_{y \in \mathcal{Y}(x)} \sum_{s} f_s(\mathbf{y}_s)$$

Solvable with dynamic programming when the graph is a tree...

Typically challenging otherwise:

- combination of dynamic programs **(blow-up the number of states)**
- non-projective parsing with higher-order features **(NP-hard)**
- summarization as relevance/redundancy trade-off **(NP-hard)**
- summarization as max-cover with knapsack constraints **(NP-hard)**

Global/Local Decoding

Global decoding problem:

$$\widehat{y} = \arg \max_{y \in \mathcal{Y}(x)} \sum_{s} f_s(\mathbf{y}_s)$$

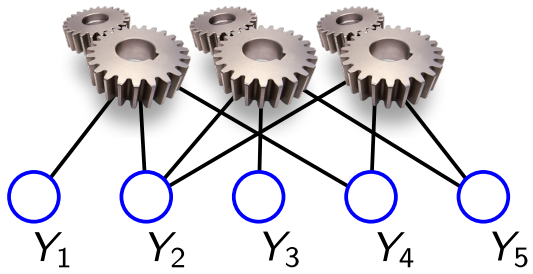
Solvable with dynamic programming when the graph is a tree...

Typically challenging otherwise:

- combination of dynamic programs **(blow-up the number of states)**
- non-projective parsing with higher-order features **(NP-hard)**
- summarization as relevance/redundancy trade-off **(NP-hard)**
- summarization as max-cover with knapsack constraints **(NP-hard)**

We'll build (approximate) global decoders given only local decoders.

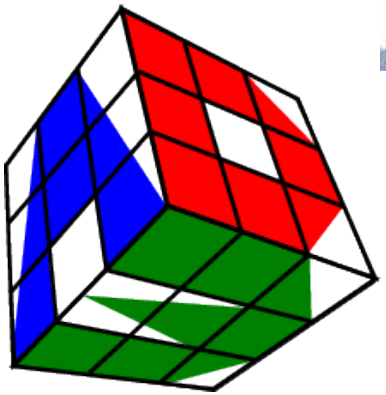
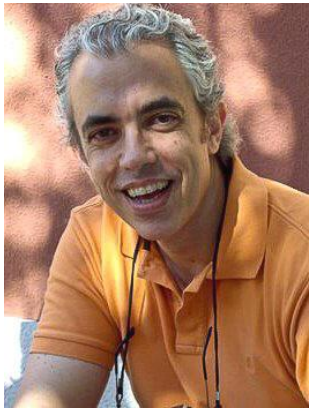
What Kind of Local Decoding Do We Need?



Algorithm	Local Operation
Sum-Prod. BP (Pearl, 1988)	marginals
TRBP (Wainwright et al., 2005)	marginals
Norm-Product BP (Hazan and Shashua, 2010)	marginals
Max-Prod. BP (Pearl, 1988)	max-marginals
TRW-S (Kolmogorov, 2006)	max-marginals
MPLP (Globerson and Jaakkola, 2008)	max-marginals
PSDD (Komodakis et al., 2007)	MAP
Accelerated DD (Jojic et al., 2010)	marginals
AD^3 (Martins et al., 2011a)	QP/MAP

Outline

AD^3: Alternating Directions Dual Decomposition


- A. Martins, M. Figueiredo, P. Aguiar, N. Smith, E. Xing. "An Augmented Lagrangian Approach to Constrained MAP Inference." ICML, Bellevue, USA, 2011.

Dual Decomposition (Komodakis et al., 2007; Rush et al., 2010)

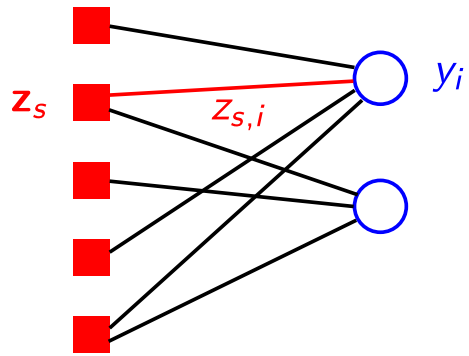
Dual Decomposition (Komodakis et al., 2007; Rush et al., 2010)

- Define *local* variables $\langle \mathbf{z}_s \rangle_{s=1}^{S}$ ("views")

Dual Decomposition (Komodakis et al., 2007; Rush et al., 2010)

- Define *local* variables $\langle \mathbf{z}_s \rangle_{s=1}^{S}$ ("views")
- Reformulate the problem by constraining views to agree on overlaps:

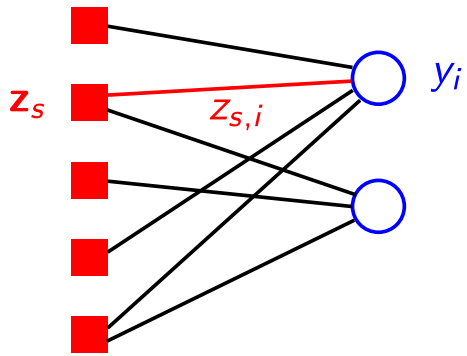
$$\begin{aligned} \text{maximize} \quad & \sum_s f_s(\mathbf{z}_s) \\ \text{w.r.t.} \quad & \mathbf{z}_s \in \mathcal{Y}_s, \ \forall s, \\ & \mathbf{y} \in \mathbb{R}^{|V|} \\ \text{s.t.} \quad & z_{s,i} = y_i, \ \forall s, i \end{aligned}$$



Dual Decomposition (Komodakis et al., 2007; Rush et al., 2010)

- Define *local* variables $\langle \mathbf{z}_s \rangle_{s=1}^{S}$ ("views")
- Reformulate the problem by constraining views to agree on overlaps:

$$\begin{aligned} \text{maximize} \quad & \sum_s f_s(\mathbf{z}_s) \\ \text{w.r.t.} \quad & \mathbf{z}_s \in \mathcal{Y}_s, \ \forall s, \\ & \mathbf{y} \in \mathbb{R}^{|V|} \\ \text{s.t.} \quad & z_{s,i} = y_i, \ \forall s, i \end{aligned}$$



- Solve a relaxation of the original problem.

Projected Subgradient (Komodakis et al., 2007)

initialize penalties to zero
repeat

until consensus (all $z_{s,i} = y_i$) or maximum number of iterations reached

Projected Subgradient (Komodakis et al., 2007)

initialize penalties to zero
repeat
  for each component $s = 1, \ldots, S$ **do**
    $\mathbf{z}_s \leftarrow \textsc{LocalDecode}(f_s(\mathbf{z}_s) + \text{penalty})$
  end for

until consensus (all $z_{s,i} = y_i$) or maximum number of iterations reached

Projected Subgradient (Komodakis et al., 2007)

initialize penalties to zero
repeat
  for each component $s = 1, \ldots, S$ **do**
    $\mathbf{z}_s \leftarrow \textsc{LocalDecode}(f_s(\mathbf{z}_s) + \text{penalty})$
  end for
  $y_i \leftarrow \textsc{Average}\,(z_{s,i} \mid i \in \partial(s))$

until consensus (all $z_{s,i} = y_i$) or maximum number of iterations reached

Projected Subgradient (Komodakis et al., 2007)

initialize penalties to zero
repeat
  for each component $s = 1, \ldots, S$ **do**
    $\mathbf{z}_s \leftarrow \textsc{LocalDecode}(f_s(\mathbf{z}_s) + \text{penalty})$
  end for
  $y_i \leftarrow \textsc{Average}\,(z_{s,i} \mid i \in \partial(s))$
  update penalties based on $z_{s,i} - y_i$
until consensus (all $z_{s,i} = y_i$) or maximum number of iterations reached

Projected Subgradient (Komodakis et al., 2007)

initialize penalties to zero
repeat
  for each component $s = 1, \ldots, S$ **do**
    $\mathbf{z}_s \leftarrow \textsc{LocalDecode}(f_s(\mathbf{z}_s) + \text{penalty})$
  end for
  $y_i \leftarrow \textsc{Average}\,(z_{s,i} \mid i \in \partial(s))$
  update penalties based on $z_{s,i} - y_i$
until consensus (all $z_{s,i} = y_i$) or maximum number of iterations reached

- **Problem:** too slow with many factors (Martins et al., 2011b)

Projected Subgradient (Komodakis et al., 2007)

initialize penalties to zero
repeat
  for each component $s = 1, \ldots, S$ **do**
    $\mathbf{z}_s \leftarrow \textsc{LocalDecode}(f_s(\mathbf{z}_s) + \text{penalty})$
  end for
  $y_i \leftarrow \textsc{Average}\,(z_{s,i} \mid i \in \partial(s))$
  update penalties based on $z_{s,i} - y_i$
until consensus (all $z_{s,i} = y_i$) or maximum number of iterations reached

- **Problem:** too slow with many factors (Martins et al., 2011b)
- **How to accelerate consensus?**

Projected Subgradient (Komodakis et al., 2007)

initialize penalties to zero
repeat
  for each component $s = 1, \ldots, S$ **do**
    $\mathbf{z}_s \leftarrow \textsc{LocalDecode}(f_s(\mathbf{z}_s) + \text{penalty})$
  end for
  $y_i \leftarrow \textsc{Average}\,(z_{s,i} \mid i \in \partial(s))$
  update penalties based on $z_{s,i} - y_i$
until consensus (all $z_{s,i} = y_i$) or maximum number of iterations reached

- **Problem:** too slow with many factors (Martins et al., 2011b)
- **How to accelerate consensus?**
- **Alternating Direction Method of Multipliers (ADMM)**

Projected Subgradient (Komodakis et al., 2007)

initialize penalties to zero
repeat
  for each component $s = 1, \ldots, S$ **do**
    $\mathbf{z}_s \leftarrow \textsc{LocalDecode}(f_s(\mathbf{z}_s) + \text{penalty})$
  end for
  $y_i \leftarrow \textsc{Average}\,(z_{s,i} \mid i \in \partial(s))$
  update penalties based on $z_{s,i} - y_i$
until consensus (all $z_{s,i} = y_i$) or maximum number of iterations reached

- **Problem:** too slow with many factors (Martins et al., 2011b)
- **How to accelerate consensus?**
- **Alternating Direction Method of Multipliers (ADMM)**
- *AD^3 = ADMM applied to graphical models*

From Subgradient to AD^3 (Martins et al., 2011a)

initialize penalties to zero
repeat
  for each component $s = 1, \ldots, S$ **do**
    $\mathbf{z}_s \leftarrow \textsc{LocalDecode}\,(f_s(\mathbf{z}_s) + \text{penalty})$
  end for
  $y_i \leftarrow \textsc{Average}\,(z_{s,i} \mid i \in \partial(s))$
  update penalties based on $z_{s,i} - y_i$
until consensus (all $z_{s,i} = y_i$) or maximum number of iterations reached

From Subgradient to AD^3 (Martins et al., 2011a)

initialize penalties to zero
repeat
  for each component $s = 1, \ldots, S$ **do**
    $\mathbf{z}_s \leftarrow \textsc{LocalDecode}\left(f_s(\mathbf{z}_s) + \text{penalty} + \sum_i \|z_{s,i} - y_i\|^2\right)$
  end for
  $y_i \leftarrow \textsc{Average}\left(z_{s,i} \mid i \in \partial(s)\right)$
  update penalties based on $z_{s,i} - y_i$
until consensus (all $z_{s,i} = y_i$) or maximum number of iterations reached

From Subgradient to AD^3 (Martins et al., 2011a)

initialize penalties to zero
repeat
  for each component $s = 1, \ldots, S$ **do**
    $\mathbf{z}_s \leftarrow \textsc{LocalDecode}\left(f_s(\mathbf{z}_s) + \text{penalty} + \sum_i \|z_{s,i} - y_i\|^2\right)$
  end for
  $y_i \leftarrow \textsc{Average}\left(z_{s,i} \mid i \in \partial(s)\right)$
  update penalties based on $z_{s,i} - y_i$
until consensus (all $z_{s,i} = y_i$) or maximum number of iterations reached

- **faster consensus:** regularize **z**-step towards average votes in **y**

From Subgradient to AD^3 (Martins et al., 2011a)

initialize penalties to zero
repeat
  for each component $s = 1, \ldots, S$ **do**
    $\mathbf{z}_s \leftarrow \textsc{LocalDecode}\left(f_s(\mathbf{z}_s) + \text{penalty} + \sum_i \|z_{s,i} - y_i\|^2\right)$
  end for
  $y_i \leftarrow \textsc{Average}\left(z_{s,i} \mid i \in \partial(s)\right)$
  update penalties based on $z_{s,i} - y_i$
until consensus (all $z_{s,i} = y_i$) or maximum number of iterations reached

- **faster consensus:** regularize **z**-step towards average votes in **y**
- **better stopping conditions:** keeps track of primal and dual residuals

Theoretical Guarantees of AD^3

Convergent in primal and dual (Glowinski and Le Tallec, 1989)

Iteration bound: $O(1/\epsilon)$ (cf. $O(1/\epsilon^2)$ for projected subgradient)

Inexact AD^3 subproblems: still convergent if residuals are summable (Eckstein and Bertsekas, 1992)

Always dual feasible: can compute upper bounds and embed in branch-and-bound toward *exact* decoding (Das et al., 2012)

Theoretical Guarantees of AD^3

Convergent in primal and dual (Glowinski and Le Tallec, 1989)

Iteration bound: $O(1/\epsilon)$ (cf. $O(1/\epsilon^2)$ for projected subgradient)

Inexact AD^3 subproblems: still convergent if residuals are summable (Eckstein and Bertsekas, 1992)

Always dual feasible: can compute upper bounds and embed in branch-and-bound toward *exact* decoding (Das et al., 2012)

But: AD^3 local subproblems are quadratic (more involved than in projected subgradient)

Theoretical Guarantees of AD^3

Convergent in primal and dual (Glowinski and Le Tallec, 1989)

Iteration bound: $O(1/\epsilon)$ (cf. $O(1/\epsilon^2)$ for projected subgradient)

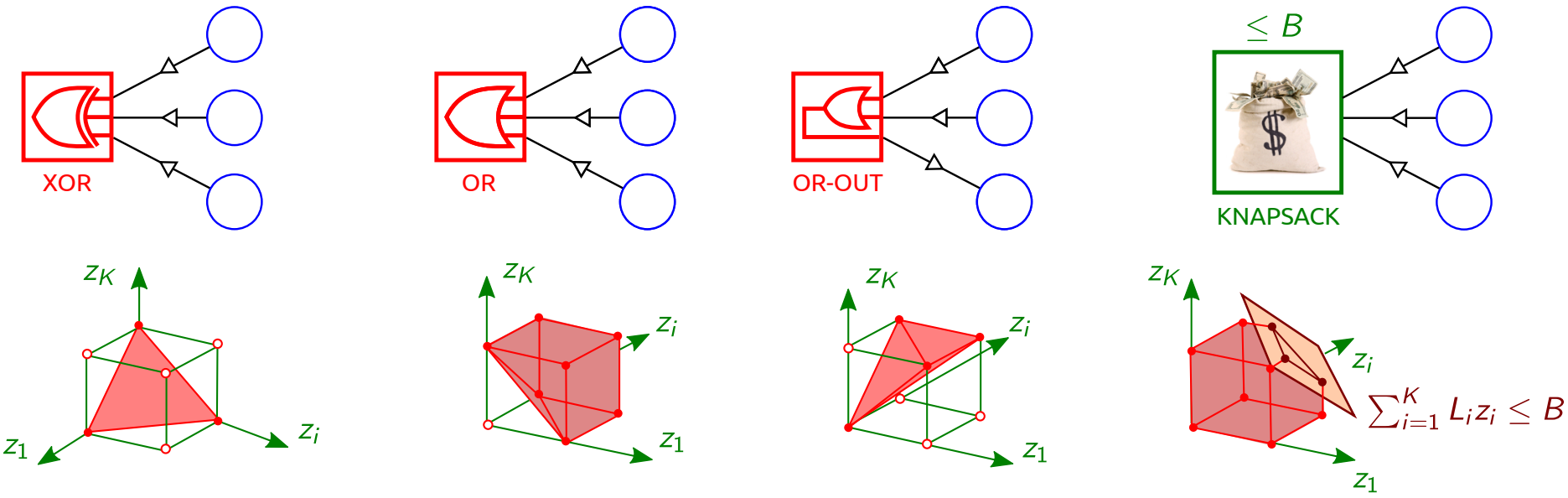
Inexact AD^3 subproblems: still convergent if residuals are summable (Eckstein and Bertsekas, 1992)

Always dual feasible: can compute upper bounds and embed in branch-and-bound toward *exact* decoding (Das et al., 2012)

But: AD^3 local subproblems are quadratic (more involved than in projected subgradient)

Still—very easy and efficient for logic and knapsack factors!

Projecting onto Hard Constraint Polytopes



- Martins et al. (2015): logic factors can be solved in $O(K)$ time
- **Almeida and Martins (2013): same for knapsack factors!**

Structured Factors

What about structured factors?

Structured Factors

What about structured factors?

Projected subgradient handles these quite well

- combinatorial machinery (Viterbi, Chu-Liu-Edmonds, Fulkerson-Ford, Floyd-Warshall,...)

We cannot solve the AD^3 subproblems with that machinery...

Structured Factors

What about structured factors?

Projected subgradient handles these quite well

- combinatorial machinery (Viterbi, Chu-Liu-Edmonds, Fulkerson-Ford, Floyd-Warshall,...)

We cannot solve the AD^3 subproblems with that machinery...

Or can we?

Structured Factors

What about structured factors?

Projected subgradient handles these quite well

- combinatorial machinery (Viterbi, Chu-Liu-Edmonds, Fulkerson-Ford, Floyd-Warshall,...)

We cannot solve the AD^3 subproblems with that machinery...

Or can we?

Active set method: seek the support of the solution by adding/removing components; very suitable for warm-starting (Nocedal and Wright, 1999)

Only requirement is a local-max oracle (as in projected subgradient)

Structured Factors

What about structured factors?

Projected subgradient handles these quite well

- combinatorial machinery (Viterbi, Chu-Liu-Edmonds, Fulkerson-Ford, Floyd-Warshall,...)

We cannot solve the AD^3 subproblems with that machinery...

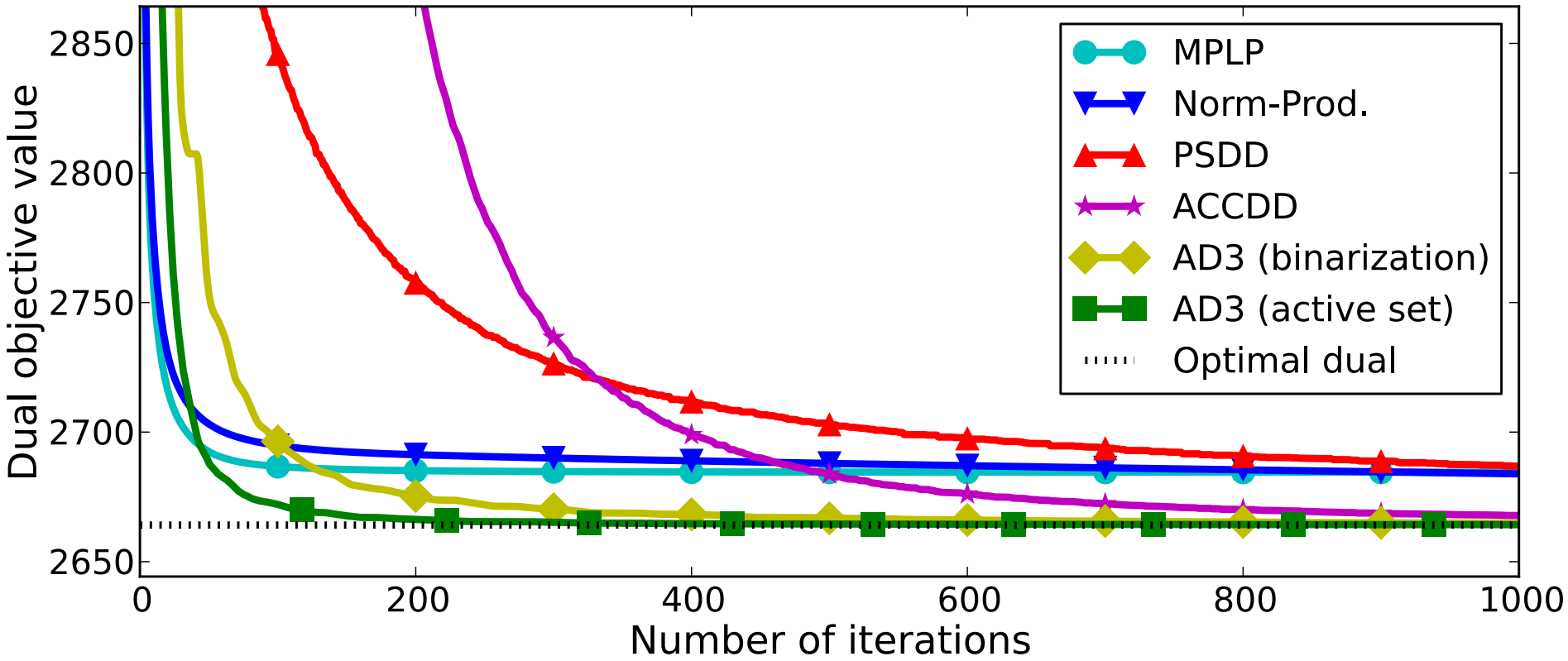
Or can we?

Active set method: seek the support of the solution by adding/removing components; very suitable for warm-starting (Nocedal and Wright, 1999)

Only requirement is a local-max oracle (as in projected subgradient)

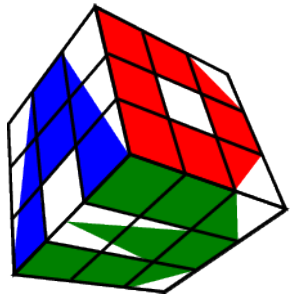
More info: Martins et al. (2015)

Example: Potts Grid (20×20, 8 states)



- A. Martins, M. Figueiredo, P. Aguiar, N. Smith, E. Xing (2015).
  AD^3: Alternating Directions Dual Decomposition for MAP Inference in Graphical Models.
  Journal of Machine Learning Research (to appear).

Try It Yourself: AD^3 Toolkit

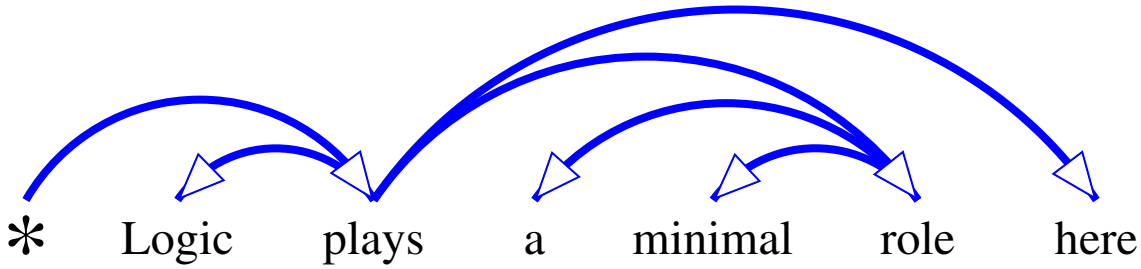


- Freely available at: *http://www.ark.cs.cmu.edu/AD3*
- Implemented in C++, includes a Python wrapper (thanks to Andy Mueller)
- Many built-in factors: logic, knapsack, dense, and some structured factors
- You can implement your own factor (only need to write a local MAP decoder!)
- Toy examples included (parsing, coreference, Potts models)

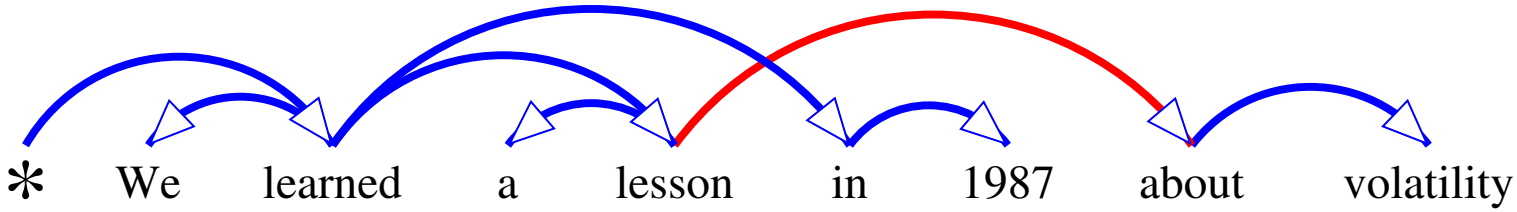
Outline

An Important Distinction

- A projective tree:

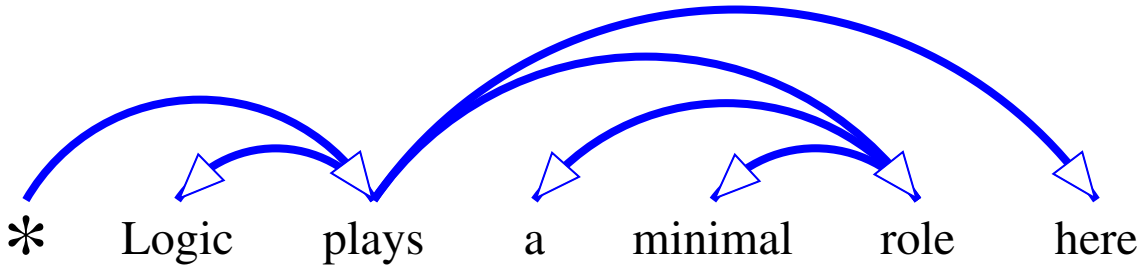


- A non-projective tree:

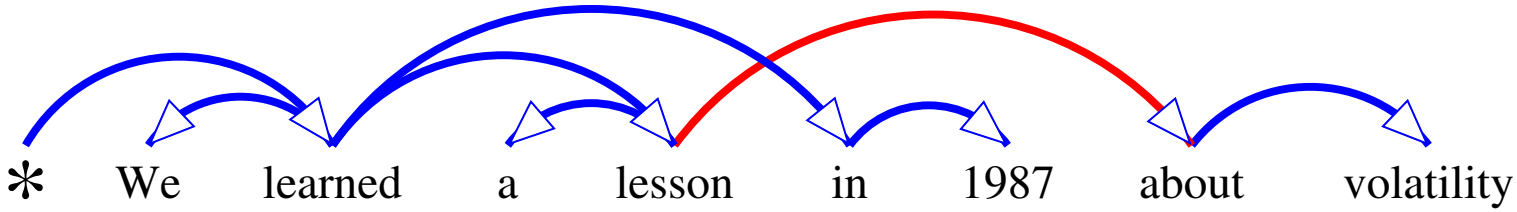


An Important Distinction

- A projective tree:



- A non-projective tree:

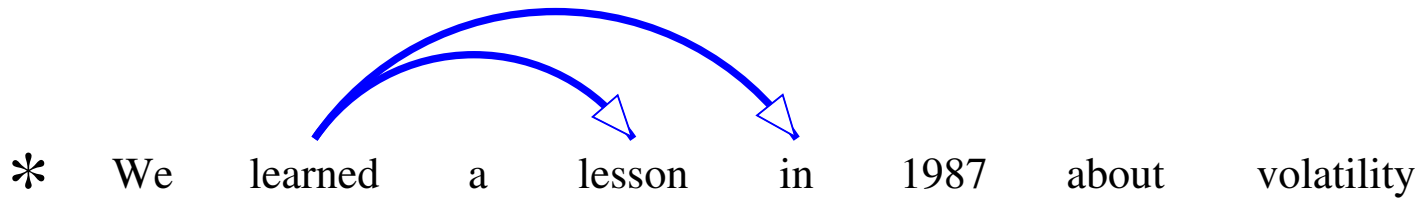


This talk: we allow non-projective trees.

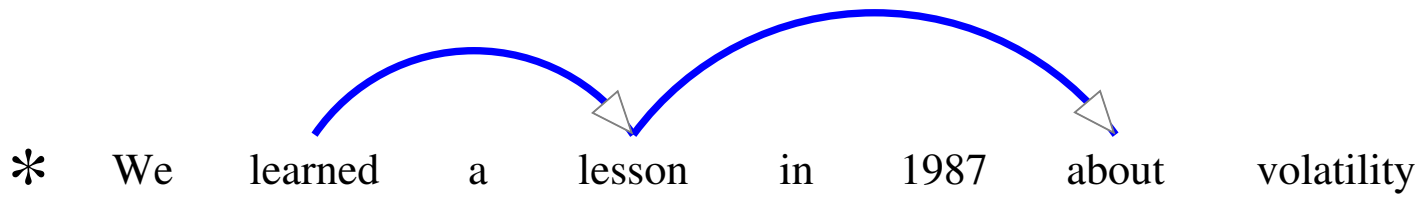
Suitable for languages with flexible word order (Dutch, German, Czech,...)

First-Order Scores for Arcs

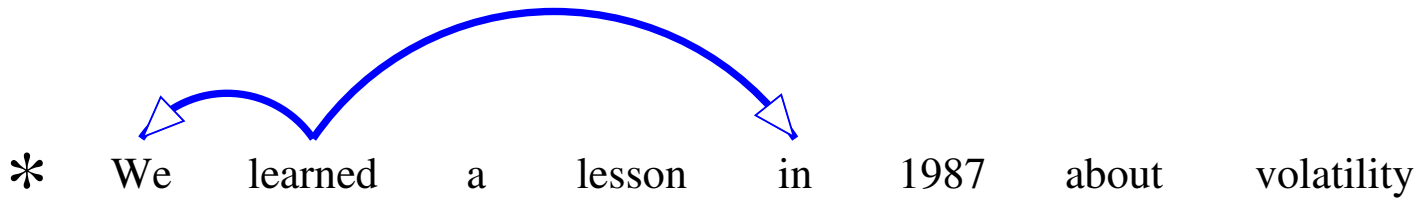
Second-Order Scores for Consecutive Siblings



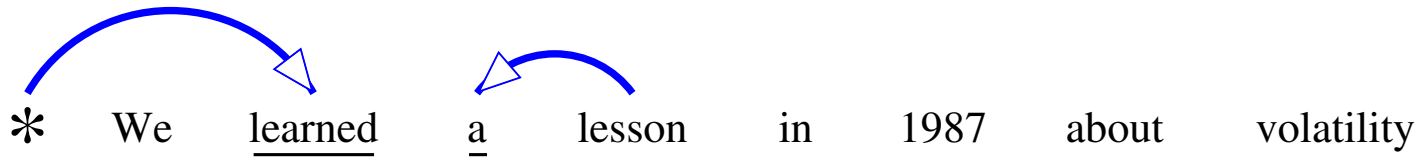
Second-Order Scores for Grandparents



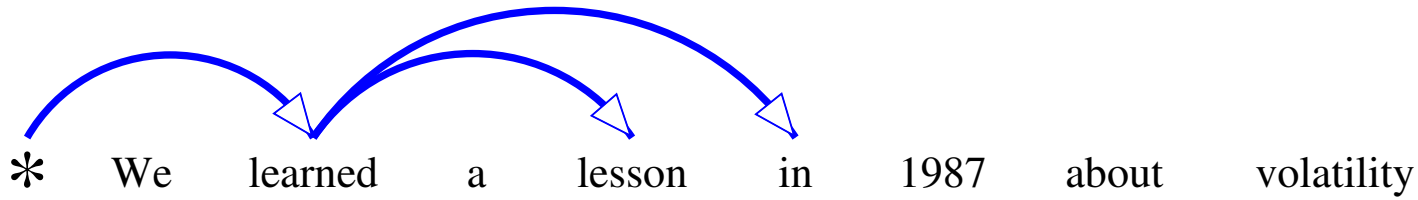
Scores for Arbitrary Siblings



Scores for Head Bigrams

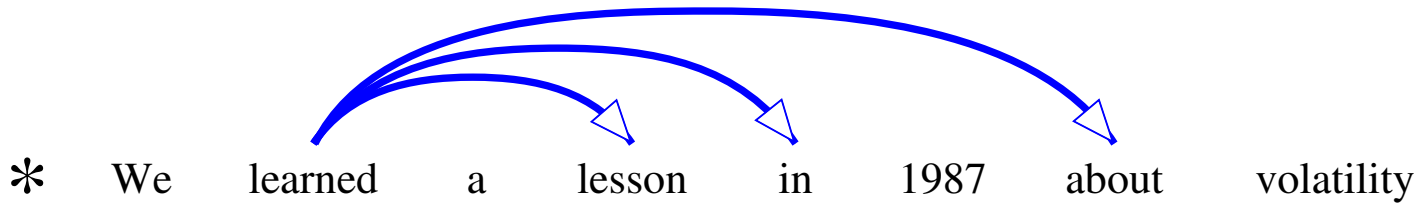


Third-Order Scores for Grand-siblings



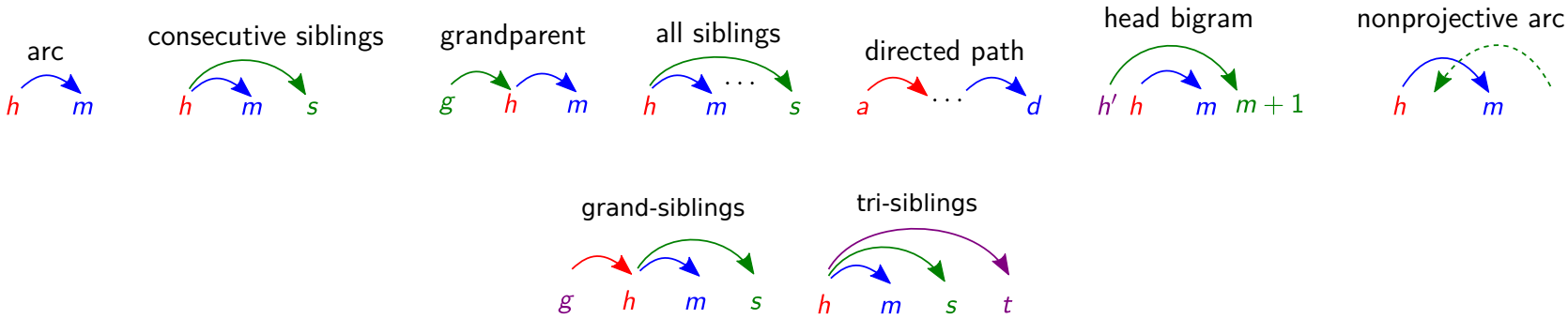
Used by Koo and Collins (2010) for *projective* parsing.

Third-Order Scores for Tri-siblings



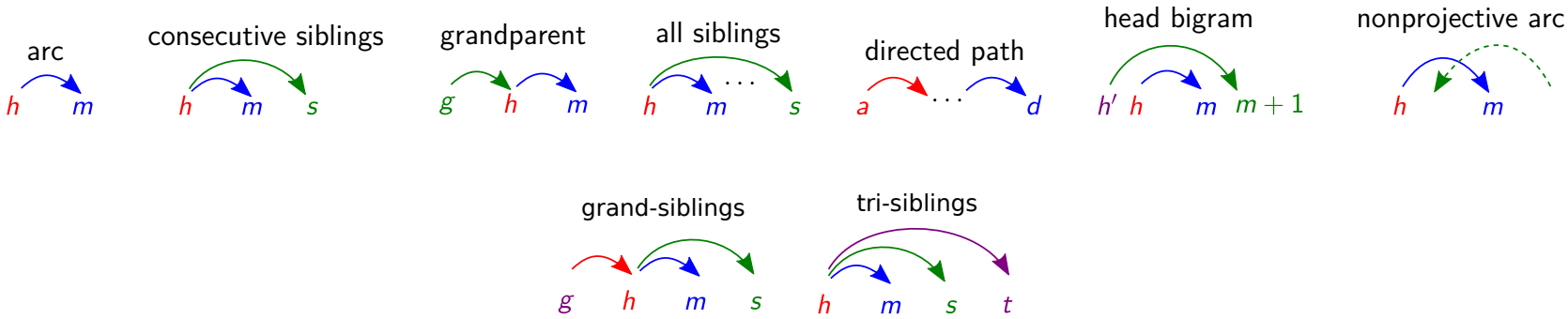
Used by Koo and Collins (2010) for *projective* parsing.

Decoding



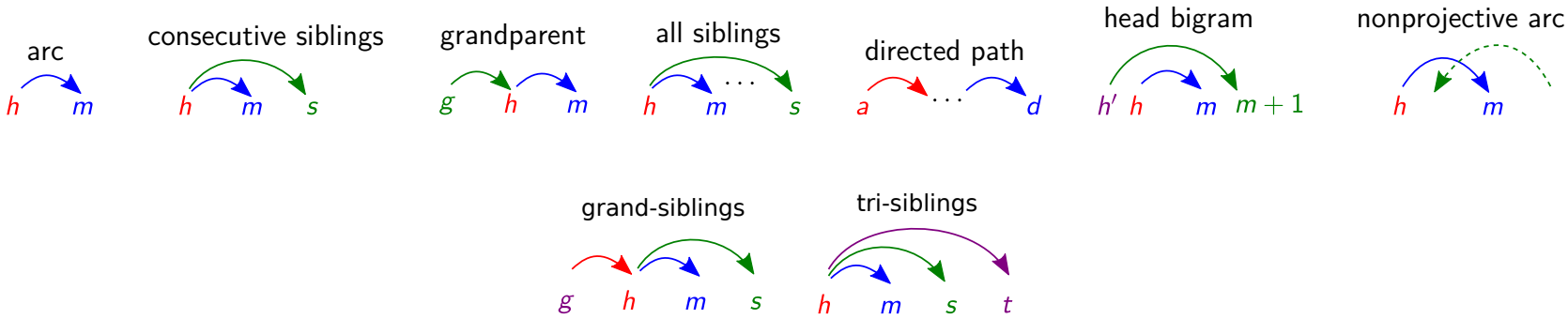
- How to deal with all these parts?

Decoding



- How to deal with all these parts?
- Beyond arc-factored models, non-projective parsing is **NP-hard** (McDonald and Satta, 2007)—**need to embrace approximations!**

Decoding



- How to deal with all these parts?
- Beyond arc-factored models, non-projective parsing is **NP-hard** (McDonald and Satta, 2007)—**need to embrace approximations!**

	parser	AF	CS	G	AS	DP	HB	NPA	GS	TS
McDonald et al. (2006)	projective + greedy	✓	✓							
Smith et al. (2008)	loopy BP	✓	✓	✓	✓					
Martins et al. (2010)	LP solver	✓		✓	✓			✓		
Koo et al. (2010)	dual decomp.	✓	✓							
Martins et al. (2011)	AD^3	✓	✓	✓	✓	✓	✓	✓		
Martins et al. (2013)	AD^3 & active set	✓	✓	✓	✓		✓		✓	✓

Factor Graph Representation

$O(L^2)$ binary variable nodes (one per possible dependency arc), linked to a mix of **dense**, **structured** and **hard-constraint factors**:

- **Pairwise factors** for arbitrary siblings
- **Head automata** for consecutive siblings, grandparents, tri-siblings, and grand-siblings (as Smith and Eisner (2008); Koo et al. (2010))
- **Sequence model** for head bigrams
- A **tree constraint** to make sure we have a valid tree at the end

Factor Graph Representation

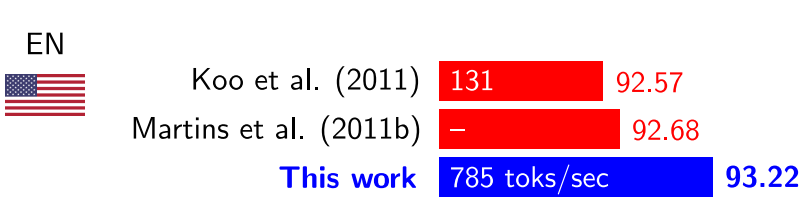
$O(L^2)$ binary variable nodes (one per possible dependency arc), linked to a mix of **dense**, **structured** and **hard-constraint factors**:

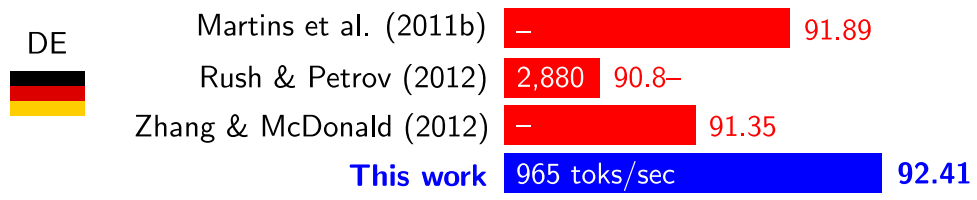
- **Pairwise factors** for arbitrary siblings
- **Head automata** for consecutive siblings, grandparents, tri-siblings, and grand-siblings (as Smith and Eisner (2008); Koo et al. (2010))
- **Sequence model** for head bigrams
- A **tree constraint** to make sure we have a valid tree at the end

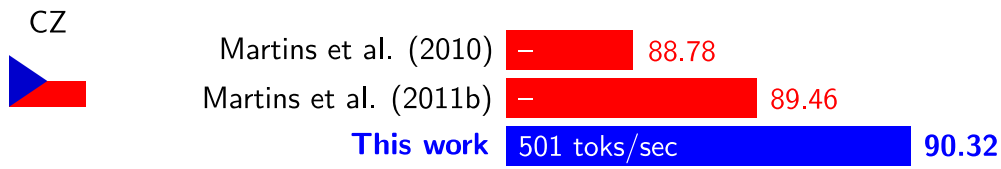
We decode with AD^3 followed by simple rounding.

Parsing Accuracies/Runtimes

SOTA accuracies for the largest non-projective datasets (CoNLL-2006 and CoNLL-2008):

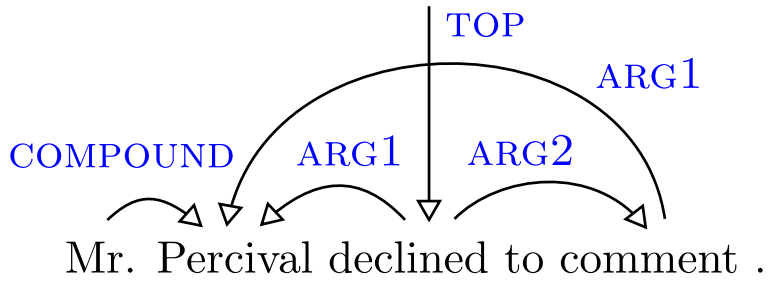






Extension: Broad-Coverage Semantic Parsing

Same idea applied to **semantic role labeling**.

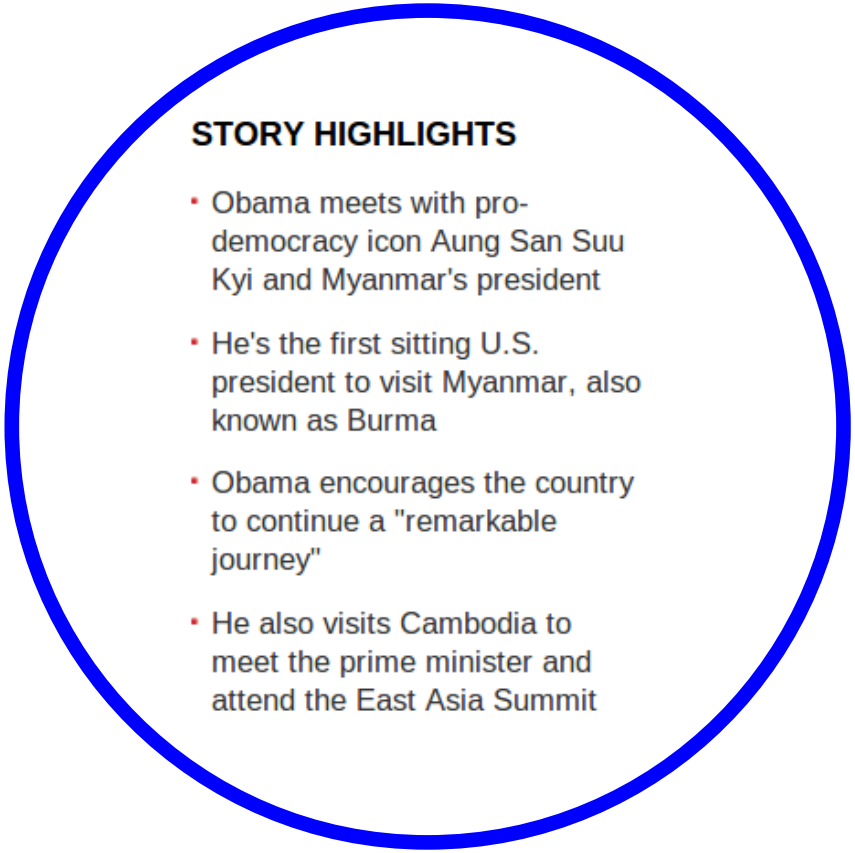


Best results in the SemEval 2014 shared task:

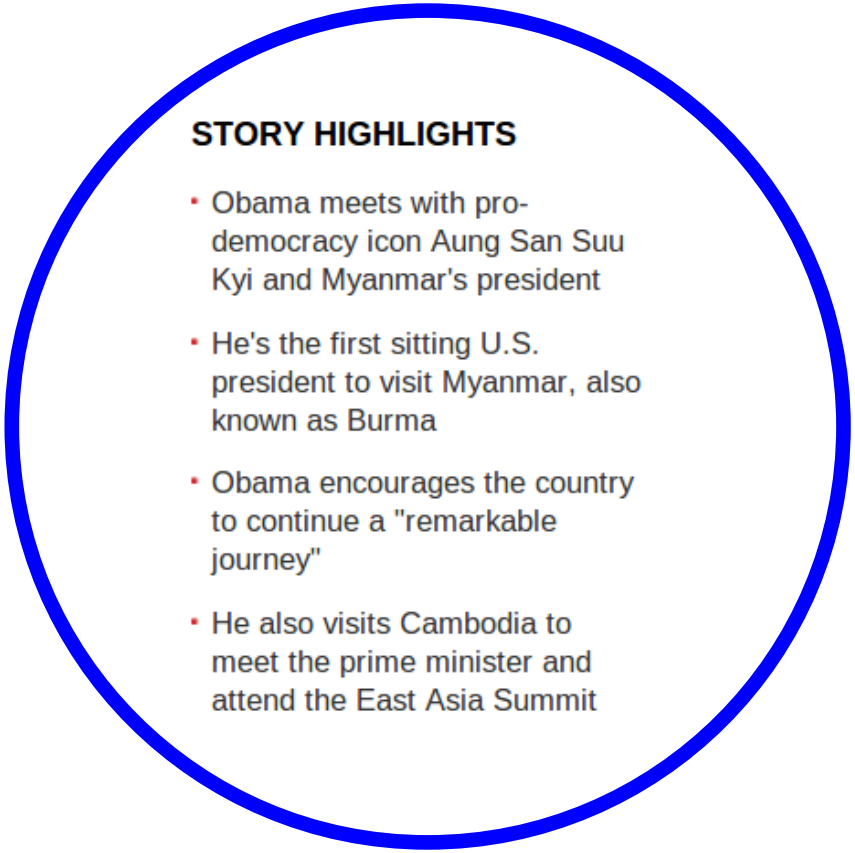
- André F. T. Martins and Mariana S. C. Almeida. "Priberam: A Turbo Semantic Parser with Second Order Features." SemEval 2014.

Outline

What Makes a Good Summary?

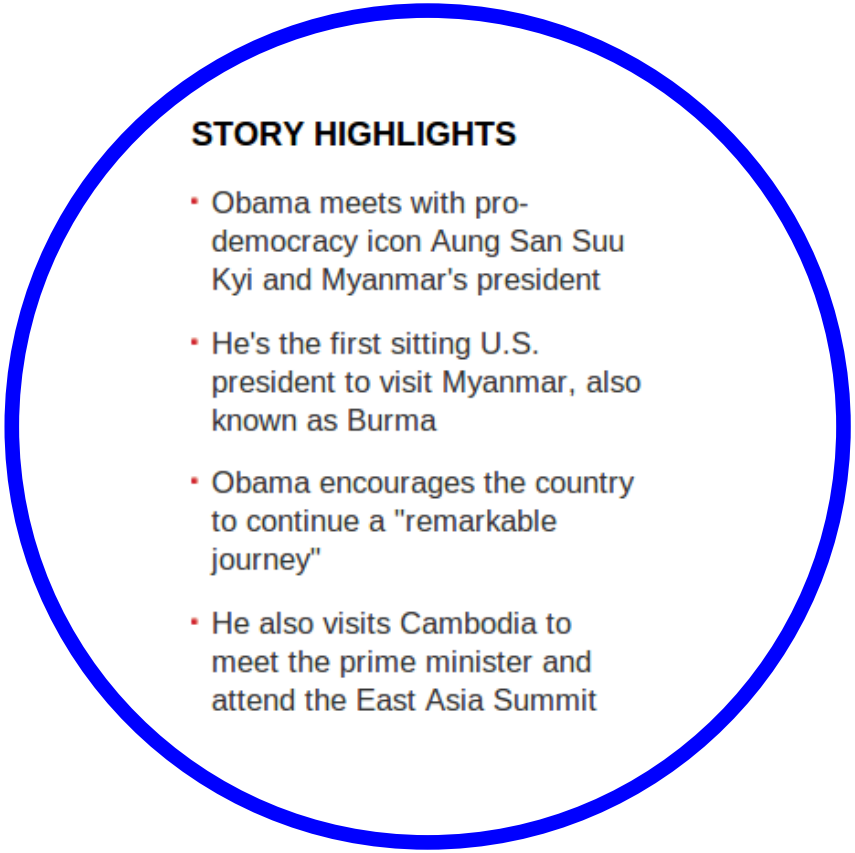


What Makes a Good Summary?



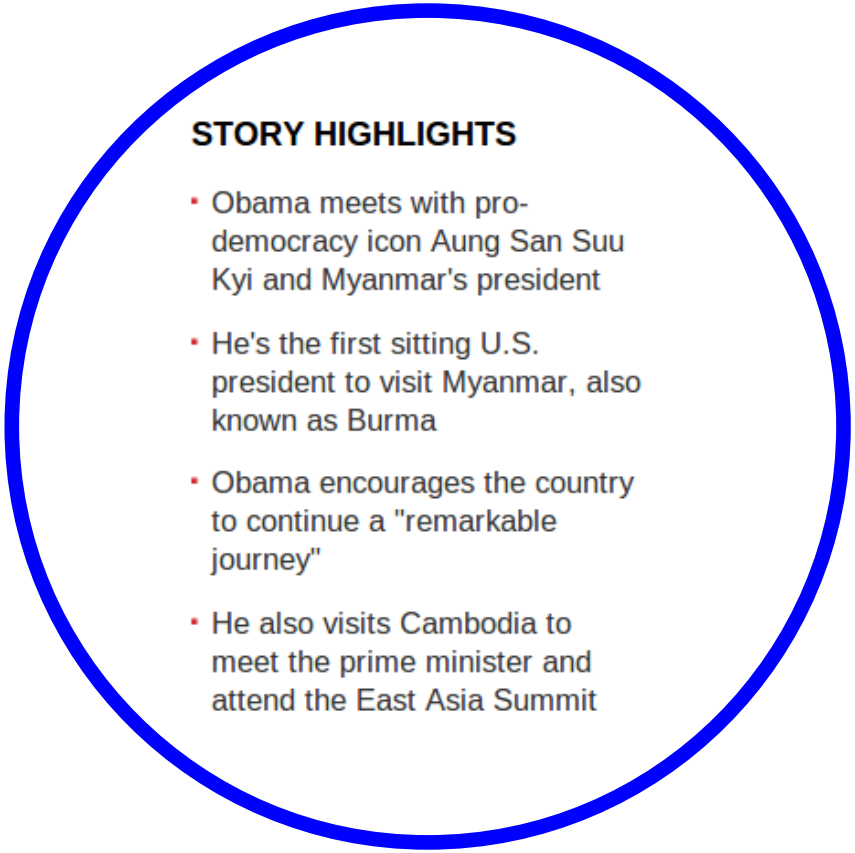
1 **conciseness**

What Makes a Good Summary?



1. **conciseness**
2. **informativeness**

What Makes a Good Summary?



1. conciseness
2. informativeness
3. grammaticality

Extractive Summarization

Just **extract** the most salient sentences.

Extractive Summarization

Just **extract** the most salient sentences.

CNN

Obama hopes for 'continued progress' in Myanmar

STORY HIGHLIGHTS

- Obama meets with pro-democracy icon Aung San Suu Kyi and Myanmar's president
- He's the first sitting U.S. president to visit Myanmar, also known as Burma
- Obama encourages the country to continue a "remarkable journey"
- He also visits Cambodia to meet the prime minister and attend the East Asia Summit

(CNN) -- Barack Obama met with Nobel Peace Prize winner Aung San Suu Kyi at her home in Myanmar on Monday, praising her "courage and determination" during a historic visit to the once repressive and secretive country.

The first sitting U.S. president to visit Myanmar, Obama urged its leaders, who have embarked on a series of far-reaching political and economic reforms since 2011, not to extinguish the "flickers of progress that we have seen."

Obama said that his visit to the lakeside villa where the pro-democracy icon spent years under house arrest marked a new chapter between the two countries.

"Here, through so many difficult years, is where she has displayed such unbreakable courage and determination," Obama told reporters, standing next to his fellow Nobel peace laureate. "It is here where she showed that human freedom and human dignity cannot be denied."


Myanmar Obama visit

The country, which is also known as Burma, was ruled by military leaders until early 2011 and for decades was politically and economically cut off from the rest of the world.

Suu Kyi acknowledged that Myanmar's opening up would be difficult.

The New York Times

YANGON, Myanmar — President Obama journeyed to this storied tropical outpost of pagodas and jungles on Monday to "extend the hand of friendship" as a land long tormented by repression and poverty begins to throw off military rule and emerge from decades of isolation.

The visit was intended to show support for the reforms put in place by Thein Sein's government since the end of military rule in November 2010.

Activists have warned that the visit may be too hasty - political prisoners remain behind bars and ethnic conflicts in border areas are unresolved.

BBC NEWS

Extractive Summarization

Just **extract** the most salient sentences.

CNN

Obama hopes for 'continued progress' in Myanmar

STORY HIGHLIGHTS

- Obama meets with pro-democracy icon Aung San Suu Kyi and Myanmar's president
- He's the first sitting U.S. president to visit Myanmar, also known as Burma
- Obama encourages the country to continue a "remarkable journey"
- He also visits Cambodia to meet the prime minister and attend the East Asia Summit

(CNN) -- Barack Obama met with Nobel Peace Prize winner Aung San Suu Kyi at her home in Myanmar on Monday, praising her "courage and determination" during a historic visit to the once repressive and secretive country.

The first sitting U.S. president to visit Myanmar, Obama urged its leaders, who have embarked on a series of far-reaching political and economic reforms since 2011, not to extinguish the "flickers of progress that we have seen."

Obama said that his visit to the lakeside villa where the pro-democracy icon spent years under house arrest marked a new chapter between the two countries.

"Here, through so many difficult years, is where she has displayed such unbreakable courage and determination," Obama told reporters, standing next to his fellow Nobel peace laureate. "It is here where she showed that human freedom and human dignity cannot be denied."

Myanmar Obama visit

The country, which is also known as Burma, was ruled by military leaders until early 2011 and for decades was politically and economically cut off from the rest of the world.

Suu Kyi acknowledged that Myanmar's opening up would be difficult.

The New York Times

YANGON, Myanmar — President Obama journeyed to this storied tropical outpost of pagodas and jungles on Monday to "extend the hand of friendship" as a land long tormented by repression and poverty begins to throw off military rule and emerge from decades of isolation.

The visit was intended to show support for the reforms put in place by Thein Sein's government since the end of military rule in November 2010.

Activists have warned that the visit may be too hasty - political prisoners remain behind bars and ethnic conflicts in border areas are unresolved.

BBC NEWS

What We Do: Compressive Summarization (Almeida and Martins, 2013)

Jointly **extract** and **compress** sentences.

What We Do: Compressive Summarization (Almeida and Martins, 2013)

Jointly **extract** and **compress** sentences.

CNN

Obama hopes for 'continued progress' in Myanmar

STORY HIGHLIGHTS

- Obama meets with pro-democracy icon Aung San Suu Kyi and Myanmar's president
- He's the first sitting U.S. president to visit Myanmar, also known as Burma
- Obama encourages the country to continue a "remarkable journey"
- He also visits Cambodia to meet the prime minister and attend the East Asia Summit

(CNN) -- Barack Obama met with Nobel Peace Prize winner Aung San Suu Kyi at her home in Myanmar on Monday, praising her "courage and determination" during a historic visit to the once repressive and secretive country.

The first sitting U.S. president to visit Myanmar, Obama urged its leaders, who have embarked on a series of far-reaching political and economic reforms since 2011, not to extinguish the "flickers of progress that we have seen."

Obama said that his visit to the lakeside villa where the pro-democracy icon spent years under house arrest marked a new chapter between the two countries.

"Here, through so many difficult years, is where she has displayed such unbreakable courage and determination," Obama told reporters, standing next to his fellow Nobel peace laureate. "It is here where she showed that human freedom and human dignity cannot be denied."

Myanmar Obama visit

The country, which is also known as Burma, was ruled by military leaders until early 2011 and for decades was politically and economically cut off from the rest of the world.

Suu Kyi acknowledged that Myanmar's opening up would be difficult.

The New York Times

YANGON, Myanmar — President Obama journeyed to this storied tropical outpost of pagodas and jungles on Monday to "extend the hand of friendship" as a land long tormented by repression and poverty begins to throw off military rule and emerge from decades of isolation.

The visit was intended to show support for the reforms put in place by Thein Sein's government since the end of military rule in November 2010.

Activists have warned that the visit may be too hasty - political prisoners remain behind bars and ethnic conflicts in border areas are unresolved.

BBC NEWS

Compressive Summarization as Global Optimization

- Indicator variables for every word of the nth sentence, $\mathbf{z}_n := \langle z_{n,\ell} \rangle_{\ell=1}^{L_n}$

Compressive Summarization as Global Optimization

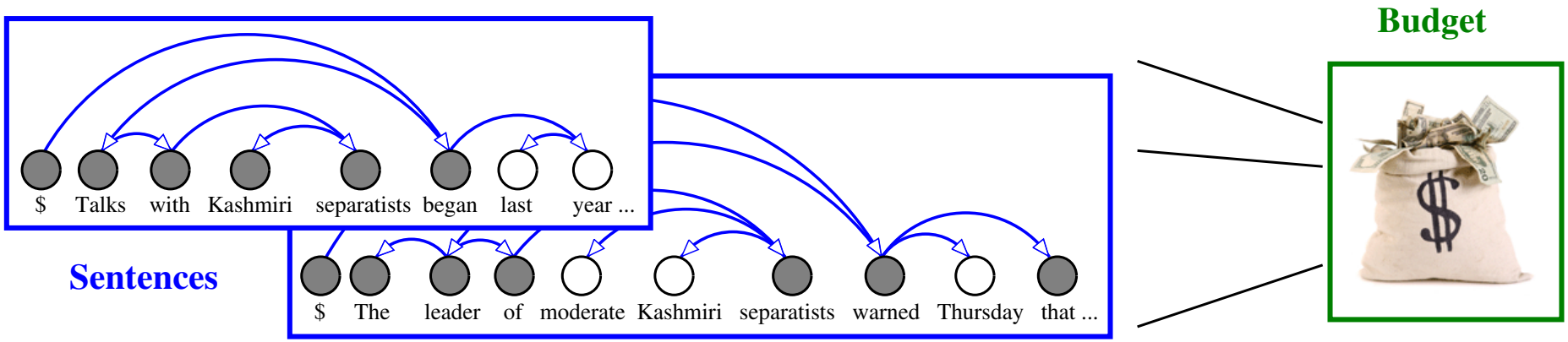
- Indicator variables for every word of the nth sentence, $\mathbf{z}_n := \langle z_{n,\ell} \rangle_{\ell=1}^{L_n}$
- Summary length must not exceed the **budget** (B words)
- Quality function rewards *global* **informativeness** (through $g(\mathbf{z})$)...
- ... but also *local* **grammaticality** (through $h_n(\mathbf{z}_n)$):

$$\begin{aligned} \text{maximize} \quad & g(\mathbf{z}) + \sum_{n=1}^{N} h_n(\mathbf{z}_n) \\ \text{s.t.} \quad & \sum_{n=1}^{N} \sum_{\ell=1}^{L_n} z_{n,\ell} \leq B. \end{aligned}$$

Graphical Model for Our Compressive Summarizer

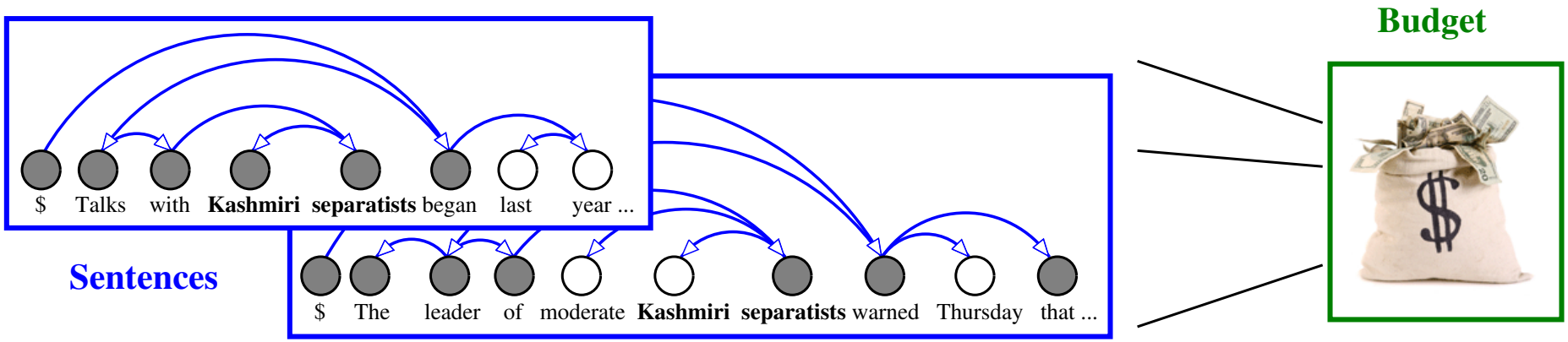
Budget

Graphical Model for Our Compressive Summarizer



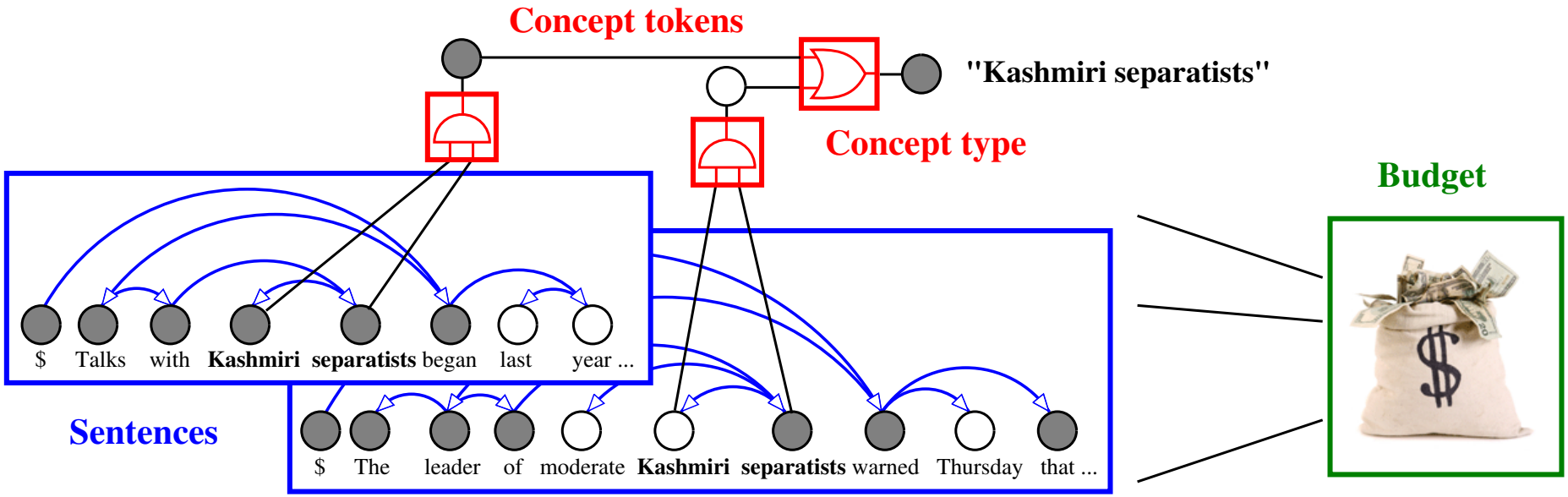
- **Grammaticality** via sentence compression models (structured factors)

Graphical Model for Our Compressive Summarizer



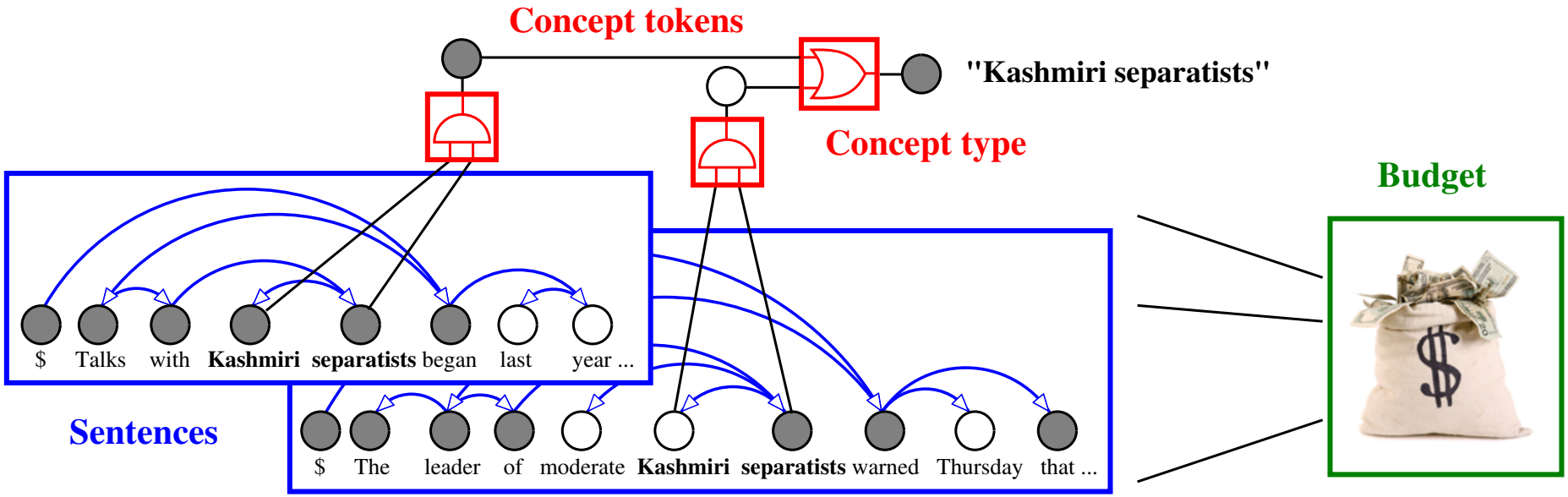
- **Grammaticality** via sentence compression models (structured factors)
- **Informativeness** via concept coverage (logic factors)

Graphical Model for Our Compressive Summarizer



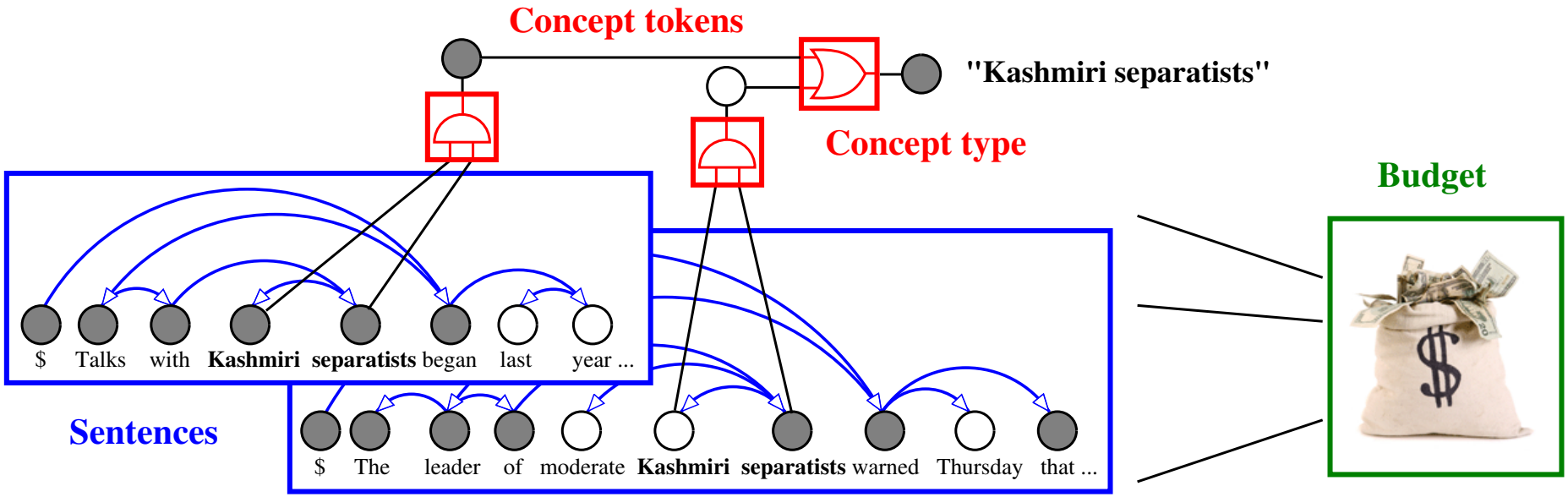
- **Grammaticality** via sentence compression models (structured factors)
- **Informativeness** via concept coverage (logic factors)

Graphical Model for Our Compressive Summarizer



- **Grammaticality** via sentence compression models (structured factors)
- **Informativeness** via concept coverage (logic factors)
- We decode with AD^3 followed by a fast rounding procedure

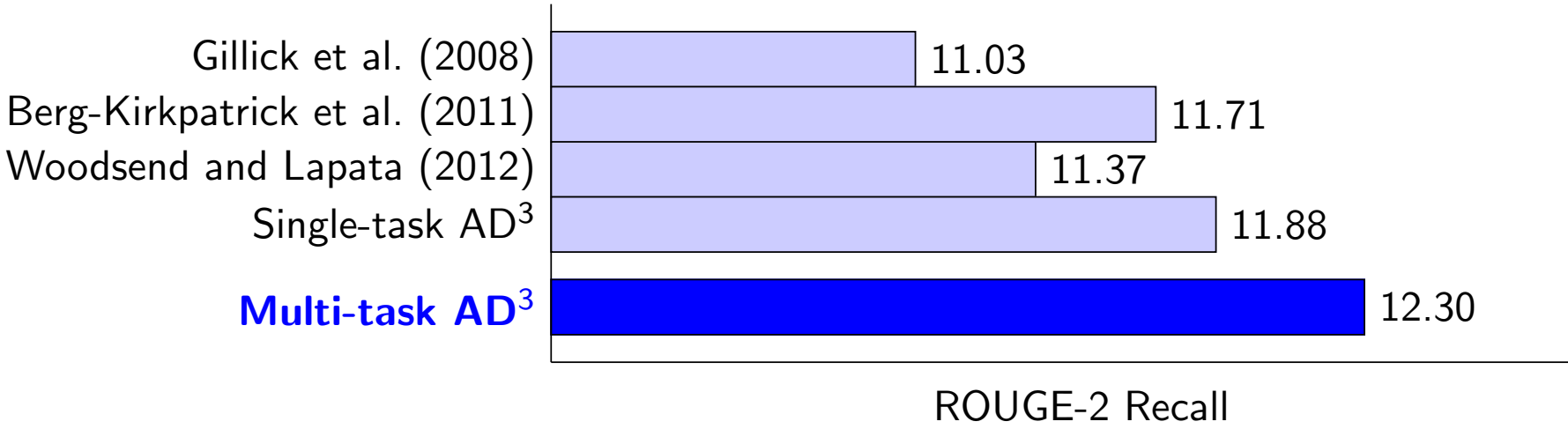
Graphical Model for Our Compressive Summarizer



- **Grammaticality** via sentence compression models (structured factors)
- **Informativeness** via concept coverage (logic factors)
- We decode with AD^3 followed by a fast rounding procedure
- We train as multi-task by mixing Simple English Wikipedia, manual abstracts, and compressive summaries.

Results on TAC-2008 Dataset

- Better informativeness (without sacrificing grammaticality):



- Averaged runtimes per summarization problem (10 documents):

Solver	Runtime (sec.)	ROUGE-2
ILP Exact, GLPK	10.394	12.40
LP-Relax., GLPK	2.265	12.38
AD^3 (1,000 its.)	**0.406**	**12.30**
Extractive (ILP)	0.265	11.16

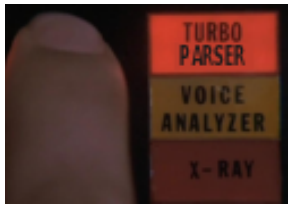
Outline

Conclusions

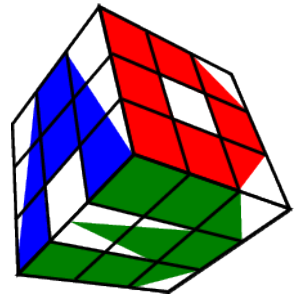
- Many structured problems in NLP are NP-hard or expensive (constrained models, diversity, combination of structured models)
- **AD^3** is a new decoder based on dual decomposition and ADMM
- Faster than the subgradient algorithm both in theory and in practice
- AD^3 subproblems can be solved in linear time for **logic**, **budget**, and **knapsack constraints**
- An **active set method** to tackle structured factors assuming only a local decoder
- Two applications with SOTA results: **turbo parsing** and **compressive summarization**
- **Could be a great fit to many other applications!!**

Thank you!

The syntactic/semantic parser and AD^3 are freely available at:



http://www.ark.cs.cmu.edu/TurboParser
http://www.ark.cs.cmu.edu/AD3



ARK lti Carnegie Mellon

Next: Coffee Break

Acknowledgments

- Fundação para a Ciência e Tecnologia, grants PEst-OE/EEI/LA0008/2011 and PTDC/EEI-SII/2312/2012.
- Fundação para a Ciência e Tecnologia and Information and Communication Technologies Institute (Portugal/USA), through the CMU-Portugal Program.
- Priberam: QREN/POR Lisboa (Portugal), EU/FEDER programme, Discooperio project, contract 2011/18501.
- Priberam: QREN/POR Lisboa (Portugal), EU/FEDER programme, Intelligo project, contract 2012/24803.

References I

Almeida, M. B. and Martins, A. F. T. (2013). Fast and robust compressive summarization with dual decomposition and multi-task learning. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*.

Berg-Kirkpatrick, T., Gillick, D., and Klein, D. (2011). Jointly learning to extract and compress. In *Proc. of Annual Meeting of the Association for Computational Linguistics*.

Berrou, C., Glavieux, A., and Thitimajshima, P. (1993). Near Shannon limit error-correcting coding and decoding. In *Proc. of International Conference on Communications*, volume 93, pages 1064–1070.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*. Now Publishers.

Das, D., Martins, A. F. T., and Smith, N. A. (2012). An Exact Dual Decomposition Algorithm for Shallow Semantic Parsing with Constraints. In *Proc. of First Joint Conference on Lexical and Computational Semantics (*SEM)*.

Eckstein, J. and Bertsekas, D. (1992). On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1):293–318.

Eisner, J. (1996). Three new probabilistic models for dependency parsing: An exploration. In *Proc. of International Conference on Computational Linguistics*, pages 340–345.

Filatova, E. and Hatzivassiloglou, V. (2004). A formal model for information selection in multi-sentence text extraction. In *Proc. of International Conference on Computational Linguistics*.

Gabay, D. and Mercier, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers and Mathematics with Applications*, 2(1):17–40.

Gillick, D., Favre, B., and Hakkani-Tur, D. (2008). The icsi summarization system at tac 2008. In *Proc. of Text Understanding Conference*.

Globerson, A. and Jaakkola, T. (2008). Fixing max-product: Convergent message passing algorithms for MAP LP-relaxations. *Neural Information Processing Systems*, 20.

Glowinski, R. and Le Tallec, P. (1989). *Augmented Lagrangian and operator-splitting methods in nonlinear mechanics*. Society for Industrial Mathematics.

Glowinski, R. and Marroco, A. (1975). Sur l'approximation, par éléments finis d'ordre un, et la résolution, par penalisation-dualité, d'une classe de problèmes de Dirichlet non linéaires. *Rev. Franc. Automat. Inform. Rech. Operat.*, 9:41–76.

References II

Hazan, T. and Shashua, A. (2010). Norm-product belief propagation: Primal-dual message-passing for approximate inference. *IEEE Transactions on Information Theory*, 56(12):6294–6316.

Jojic, V., Gould, S., and Koller, D. (2010). Accelerated dual decomposition for MAP inference. In *International Conference of Machine Learning*.

Knight, K. and Marcu, D. (2000). Statistics-based summarization—step one: Sentence compression. In *AAAI/IAAI*.

Kolmogorov, V. (2006). Convergent tree-reweighted message passing for energy minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1568–1583.

Komodakis, N., Paragios, N., and Tziritas, G. (2007). MRF optimization via dual decomposition: Message-passing revisited. In *Proc. of International Conference on Computer Vision*.

Koo, T. and Collins, M. (2010). Efficient third-order dependency parsers. In *Proc. of Annual Meeting of the Association for Computational Linguistics*, pages 1–11.

Koo, T., Globerson, A., Carreras, X., and Collins, M. (2007). Structured prediction models via the matrix-tree theorem. In *Empirical Methods for Natural Language Processing*.

Koo, T., Rush, A. M., Collins, M., Jaakkola, T., and Sontag, D. (2010). Dual decomposition for parsing with non-projective head automata. In *Proc. of Empirical Methods for Natural Language Processing*.

Lin, H. and Bilmes, J. (2010). Multi-document summarization via budgeted maximization of submodular functions. In *Proc. of Annual Meeting of the North American chapter of the Association for Computational Linguistics*.

Martins, A. F. T., Almeida, M. B., and Smith, N. A. (2013). Turning on the turbo: Fast third-order non-projective turbo parsers. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*.

Martins, A. F. T., Figueiredo, M. A. T., Aguiar, P. M. Q., Smith, N. A., and Xing, E. P. (2011a). An Augmented Lagrangian Approach to Constrained MAP Inference. In *Proc. of International Conference on Machine Learning*.

Martins, A. F. T., Figueiredo, M. A. T., Aguiar, P. M. Q., Smith, N. A., and Xing, E. P. (2015). AD^3: Alternating Directions Dual Decomposition for MAP Inference in Graphical Models. *Journal of Machine Learning Research (to appear)*.

Martins, A. F. T., Smith, N. A., Aguiar, P. M. Q., and Figueiredo, M. A. T. (2011b). Dual Decomposition with Many Overlapping Components. In *Proc. of Empirical Methods for Natural Language Processing*.

References III

Martins, A. F. T., Smith, N. A., Xing, E. P., Aguiar, P. M. Q., and Figueiredo, M. A. T. (2010a). Augmented Dual Decomposition for MAP Inference. In *Neural Information Processing Systems: Workshop in Optimization for Machine Learning*.

Martins, A. F. T., Smith, N. A., Xing, E. P., Figueiredo, M. A. T., and Aguiar, P. M. Q. (2010b). Turbo Parsers: Dependency Parsing by Approximate Variational Inference. In *Proc. of Empirical Methods for Natural Language Processing*.

McDonald, R. and Satta, G. (2007). On the complexity of non-projective data-driven dependency parsing. In *Proc. of International Conference on Parsing Technologies*.

McDonald, R. T., Pereira, F., Ribarov, K., and Hajic, J. (2005). Non-projective dependency parsing using spanning tree algorithms. In *Proc. of Empirical Methods for Natural Language Processing*.

Nivre, J., Hall, J., Nilsson, J., Eryiğit, G., and Marinov, S. (2006). Labeled pseudo-projective dependency parsing with support vector machines. In *Procs. of International Conference on Natural Language Learning*.

Nocedal, J. and Wright, S. (1999). *Numerical optimization*. Springer verlag.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Richardson, M. and Domingos, P. (2006). Markov logic networks. *Machine Learning*, 62(1):107–136.

Roth, D. and Yih, W. (2004). A linear programming formulation for global inference in natural language tasks. In *International Conference on Natural Language Learning*.

Rush, A., Sontag, D., Collins, M., and Jaakkola, T. (2010). On dual decomposition and linear programming relaxations for natural language processing. In *Proc. of Empirical Methods for Natural Language Processing*.

Smith, D. and Eisner, J. (2008). Dependency parsing by belief propagation. In *Proc. of Empirical Methods for Natural Language Processing*.

Wainwright, M. J., Jaakkola, T., and Willsky, A. (2005). A new class of upper bounds on the log partition function. *IEEE Transactions on Information Theory*, 51(7):2313–2335.

Woodsend, K. and Lapata, M. (2012). Multiple aspect summarization using integer linear programming. In *Proc. of Empirical Methods in Natural Language Processing*.

Yih, W.-t., Goodman, J., Vanderwende, L., and Suzuki, H. (2007). Multi-document summarization by maximizing informative content-words. In *Proc. of International Joint Conference on Artifical Intelligence*.