

## Exploiting Syntactic, Semantic and Lexical Regularities in Language Modeling via Directed Markov Random Fields

Shaojun Wang

Department of Computing Science  
University of Alberta

1

## Language modeling

“Accurately calculating the probability of naturally occurring word sequences in human natural language.”

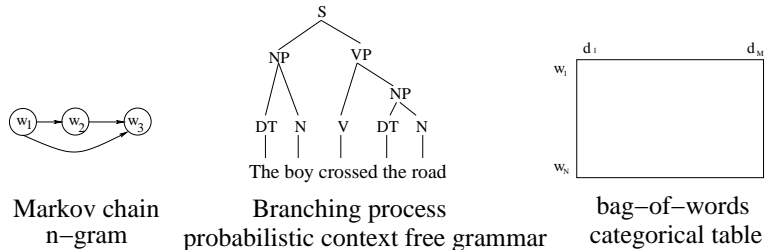
### Applications:

- Automatic speech recognition
- Machine translation
- Information retrieval
- Optical character recognition
- Spelling correction
- Bioinformatics

2

## Statistical models of natural languages

- Lots of models to represent natural language



- But ...

– Lack of a unified probabilistic framework to encode natural language

⇒ this work addresses that issue and introduces one of 2 approaches for language modeling

\* Directed Markov random fields

3

## How to combine statistical models

- Linear interpolation
  1. *Good:* easy to use
  2. *Bad:* makes suboptimal use of components; limited improvement
- Jaynes' maximum entropy principle
  1. *Good:* no data fragmentation; no independence assumption; automatic training
  2. *Major weakness:* only handles explicit features, but for natural language there are hidden structures we do not observe directly.

4

## Desired properties of a good model of language

- Can incorporate various aspect of the language (lexical, syntactic, semantic models)
- Can represent hidden interactions between these different aspects of the language
- Can be trained from data

We propose 2 approaches satisfying these requirements

- Latent maximum entropy principle (LME): allows relationships over hidden features to be effectively captured in a unified model
- Directed Markov random fields: encodes syntactic structure into semantic n-gram language model with a tractable parameter estimation algorithm

5

## Combining n-gram + PLSA by LME?

- See our paper  
“Combining Statistical Language Models via the Latent Maximum Entropy Principle” *Machine Learning Journal: Special Issue on Learning in Speech and Language Technologies*, 60(1):229-250, September 2005

## Combining n-gram + PLSA + PCFG by LME?

- It is *intractable* due to global normalization factor over infinitely large configuration space
- Have to use MCMC

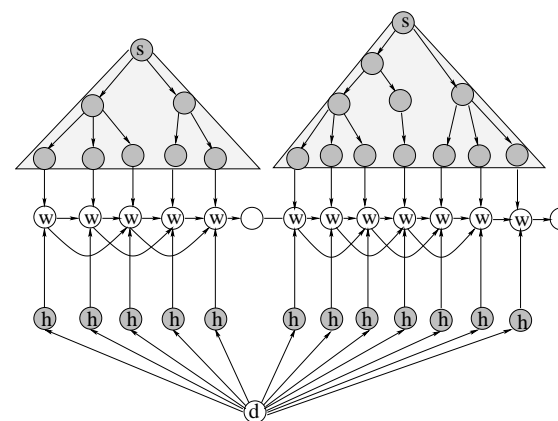
6

## Outline

- Motivation: language modeling
- **Directed Markov random fields**
  - Language modeling via directed Markov random fields

7

## Combining n-gram models, PCFGs and PLSA via directed Markov random fields



A composite chain/tree/table model, light nodes: observed information and dark nodes/triangles: hidden information.

8

## Undirected MRFs vs. directed MRFs?

- Undirected Markov random fields

$$p_\lambda(x) = \exp(\langle \lambda, f(x) \rangle - \log(\Phi_\lambda))$$

ONE global normalization factor:  $\Phi_\lambda$

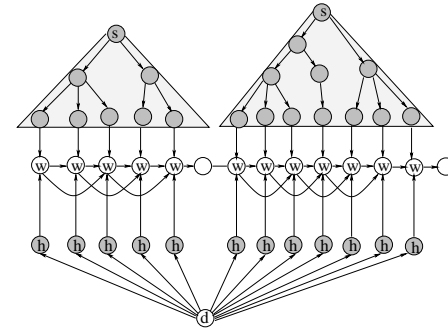
- Directed Markov random fields

$$p_\lambda(x) = \prod_j \exp(\langle \lambda_j, f(x_j, \pi_j) \rangle - \log(\Phi_{\lambda_j}(\pi_j)))$$

MANY local normalization factors:  $\Phi_{\lambda_j}(\pi_j)$

9

## Hard parameter estimation problems!!!

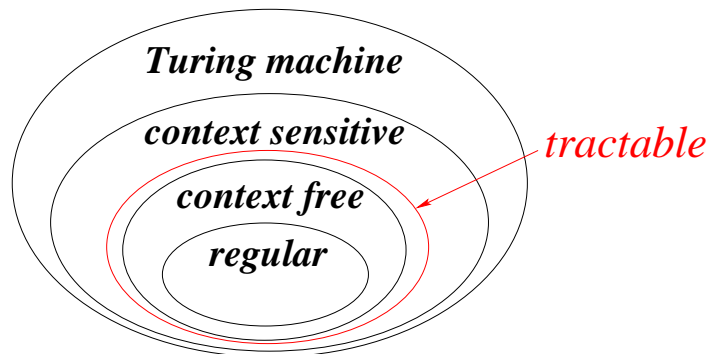


- Context sensitive grammars: NP-hard problems in general
- Potentially exponential number of loops
  - Loopy belief propagation/variational approximation or MCMC have to be used? (No)

10

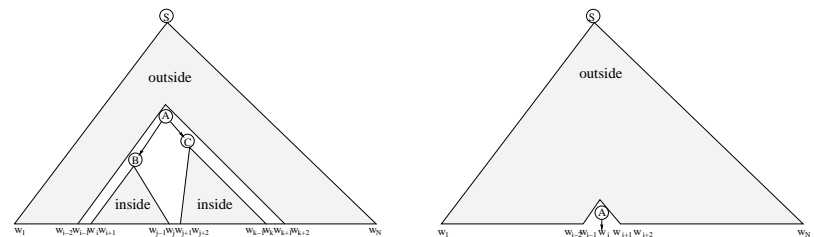
## Surprising result

- Cubic time **exact** recursive EM iterative procedure:  
*generalized inside-outside algorithm*



11

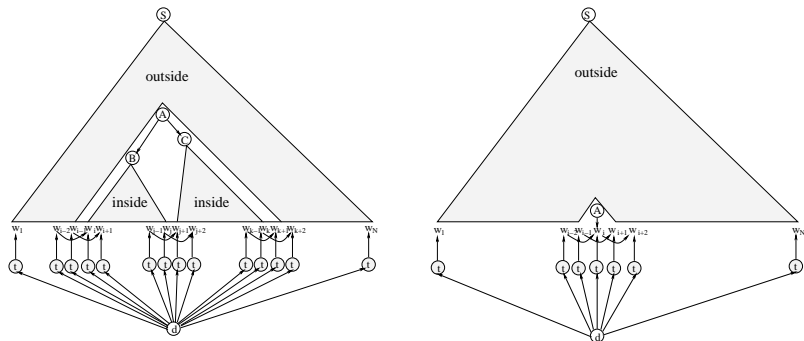
## Inside-outside algorithm (Baker 1979)



Inside and outside probabilities for the PCFG model.

12

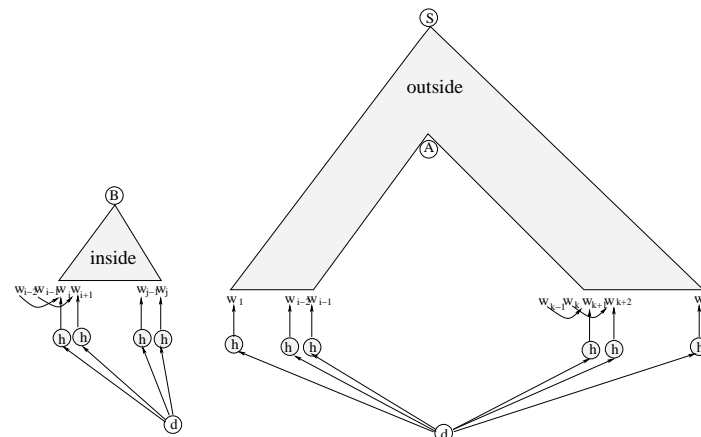
## Generalized inside outside algorithm



Inside and outside probabilities for the composite model, each component is influenced by the injected trigram and PLSA nodes.

13

## Inside and outside probabilities



Inside probability  $\alpha_{ij}(B)$  and outside probability  $\beta_{ik}(A)$  in the composite trigram/syntactic/semantic model.

14

## Recursively calculate inside and outside probabilities: one line change

$$\alpha_{ij}(A; W_l \text{ in } d) = \sum_{BC} \sum_{i \leq k \leq j} \theta(A \rightarrow BC) \alpha_{ik}(A; W_l \text{ in } d) \alpha_{k+1j}(C; W_l \text{ in } d)$$

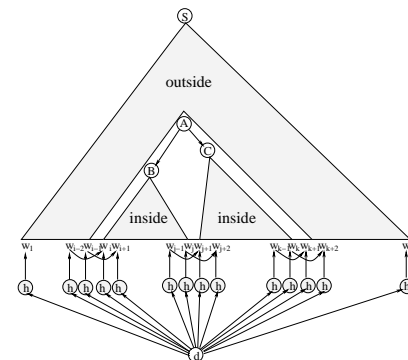
$$\alpha_{ii}(A; W_l \text{ in } d) = \sum_h \theta(d \rightarrow h) \theta(w_{i-2} w_{i-1} A h \rightarrow w_i)$$

$$\beta_{ij}(A; W_l \text{ in } d) = \sum_{B,C} \sum_{k < i} \theta(B \rightarrow CA) \alpha_{ki-1}(C; W_l \text{ in } d) \beta_{kj}(B; W_l \text{ in } d) + \sum_{B,C} \sum_{k > j} \theta(B \rightarrow AC) \alpha_{j+1k}(C; W_l \text{ in } d) \beta_{ik}(B; W_l \text{ in } d)$$

$$\beta_{1N}(A; W_l \text{ in } d) = \delta_S(A; W_l \text{ in } d)$$

15

## Why tractable? Factorization property!



$$\begin{aligned} & p_\theta(S \rightarrow W_l \text{ in } d; \text{ using } A \rightarrow BC \text{ in position } (i, j, k)) \\ &= \theta(A \rightarrow BC) p_\theta(B \Rightarrow w_i \cdots w_j; W_l \text{ in } d) \\ & \quad p_\theta(C \Rightarrow w_{j+1} \cdots w_k; W_l \text{ in } d) \\ & \quad p_\theta(S \Rightarrow w_1 \cdots w_{i-1} A w_{k+1} \cdots w_N; W_l \text{ in } d) \end{aligned}$$

16

### Proper distribution?

- Let  $\Omega$  be the set of finite parse trees,  $\hat{p}$  be any intermediate iteration of the EM procedure, the generalized inside-outside algorithm. Then  $\hat{p}(\Omega) = 1$ .

### Smoothing?

- Various smoothing techniques (Chen and Goodman, 1999) to alleviate the sparseness of trigram counts even though there exist hidden variables  $A$  and  $h$ .

17

### Left-to-right probability?

- Analogous algorithm (Jelinek and Lafferty, 1991) to calculate the probability of initial subsequence of a sentence generated by the composite language model.

### Parsing?

- Viterbi algorithm to find the most likely parser of a sentence generated by the composite language model.

18

### WSJ data sets statistics

	NO. OF ARTICLES	NO. OF SENTENCES	NO. OF WORDS
TRAIN	150,981	1,611,571	41,780,924
TEST	378	6904	157,312

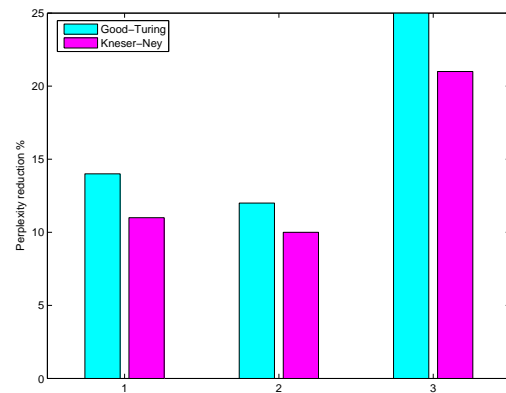
19

### Perplexity results for the composite syntactic semantic trigram model

LANGUAGE MODEL	PERPLEXITY	PERPLEXITY
	GOOD-TURING	KNESER-NEY
TRIGRAM (BASELINE)	109	103
PCFG	678	
LINEAR PCFG & TRIGRAM	109	102
PLSA	1487	
LINEAR PLSA & TRIGRAM	109	103
LINEAR PLSA, PCFG & TRIGRAM	108	102
PCFG+TRIGRAM	94	90
PLSA+TRIGRAM	96	91
PCFG+PLSA+TRIGRAM	82	79

20

## Perplexity reductions over baseline trigram



1. PCFG+trigram, 2. PLSA+trigram, 3. PCFG+PLSA+trigram

21

## Conclusions

- Generalized inside-outside algorithm for parameter estimation with cubic time complexity.
- The composite syntactic semantic trigram model induces significant perplexity reductions over baseline trigram with Good-Turing and Kneser-Ney smoothing techniques.

22

## Acknowledgements

- Work done jointly with Shaomin Wang, Russell Greiner, Dale Schuurmans and Li Cheng.
- Thanks to Olivier Siohan for his comments to improve the presentation of this material.

23