PCFGs, Topic Models, Adaptor Grammars and Learning Topical Collocations and the Structure of Proper Names

Mark Johnson
Department of Computing
Macquarie University
Mark.Johnson@mq.edu.au

July 4, 2010

MACQUARIE UNIVERSITY

Outline

LDA topic models as PCFGs

Adaptor grammars

Finding topic-specific collocations

Learning the structure of proper nouns

Conclusion

LDA topic models

- LDA topic models are generative models for documents
  - identifies documents about similar topics
  - identifies words characteristic of topics
- Each topic i is a distribution over words ϕ_i
- Each document j has a *distribution* $\boldsymbol{\theta}_j$ over topics
- To generate document j:
  - for each word position in document:
    - choose a topic z according to $\boldsymbol{\theta}_j$, and then
    - choose a word belonging to that topic according to ϕ_z
- "Sparse priors" on ϕ and $\boldsymbol{\theta}$

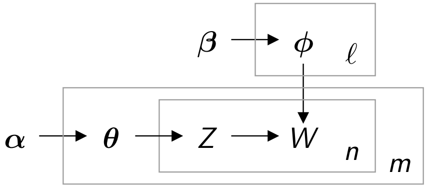
  $\Rightarrow$ most documents have few topics

  $\Rightarrow$ most topics have few words
- Bayesian inference (Gibbs sampling, Variational Bayes)

  See: Blei, Ng and Jordan (2002), Griffiths and Steyvers (2004)

LDA topic models: formal description

$$
\begin{aligned}
\boldsymbol{\phi}_i &\sim \mathrm{Dir}(\boldsymbol{\beta}) & i &= 1, \ldots, \ell = \text{number of topics} \\
\boldsymbol{\theta}_j &\sim \mathrm{Dir}(\boldsymbol{\alpha}) & j &= 1, \ldots, m = \text{number of documents} \\
z_{j,k} &\sim \boldsymbol{\theta}_j & j &= 1, \ldots, m \\
& & k &= 1, \ldots, n = \text{number of words in a document} \\
w_{j,k} &\sim \boldsymbol{\phi}_{z_{j,k}} & j &= 1, \ldots, m \\
& & k &= 1, \ldots, n
\end{aligned}
$$



Context-Free Grammars

- A CFG (N, W, R, S) defines *sets of trees* $\mathcal{T}_X$ for each $X \in N \cup W$:
  - if $X \in W$ then $\mathcal{T}_X = \{X\}$ (the 1-node tree labelled X)
  - if $X \in N$ then:

$$\mathcal{T}_X = \bigcup_{X \to B_1 \ldots B_n \in R_X} \textsc{Tree}_X(\mathcal{T}_{B_1}, \ldots, \mathcal{T}_{B_n})$$

where $R_A = \{A \to \beta : A \to \beta \in R\}$ for each $A \in N$, and

$$\textsc{Tree}_X(\mathcal{T}_{B_1}, \ldots, \mathcal{T}_{B_n}) = \left\{ \begin{array}{c} X \\ \overbrace{t_1 \ \ldots \ t_n} \end{array} : \begin{array}{l} t_i \in \mathcal{T}_{B_i}, \\ i = 1, \ldots, n \end{array} \right\}$$

That is, $\textsc{Tree}_X(\mathcal{T}_{B_1}, \ldots, \mathcal{T}_{B_n})$ consists of the set of trees with whose root node is labelled X and whose ith child is a member of $\mathcal{T}_{B_i}$.

Probabilistic Context-Free Grammars

- A PCFG is a CFG (N, W, R, S) and multinomials $\boldsymbol{\theta}_X$ over R_X for each $X \in N$
  - $\theta_{X \to \beta}$ is the probability of X expanding to β
- A PCFG associates each $X \in N \cup W$ with a *distribution G_X over trees $\mathcal{T}_X$*
  - if $X \in W$ then $G_X(X) = 1$
  - if $X \in N$ then:

$$G_X(t) = \sum_{X \to B_1 \ldots B_n \in R_X} \theta_{X \to B_1 \ldots B_n} \mathrm{TD}_X(G_{B_1}, \ldots, G_{B_n})(t) \quad (1)$$

where:

$$\mathrm{TD}_A(G_1, \ldots, G_n)\left(\begin{array}{c} X \\ \overbrace{t_1 \;\ldots\; t_n} \end{array} \right) = \prod_{i=1}^{n} G_i(t_i).$$

That is, $\mathrm{TD}_A(G_1, \ldots, G_n)$ is a distribution over $\mathcal{T}_A$ where each subtree t_i is generated independently from G_i.

Bayesian PCFGs

- Place Dirichlet priors $\mathrm{Dir}(\boldsymbol{\alpha}_X)$ on each rule probability multinomial $\boldsymbol{\theta}_X$ for each $X \in N$

$$\boldsymbol{\theta}_X \sim \mathrm{Dir}(\boldsymbol{\alpha}_X) \quad X \in N$$

- "Sparse priors" $\Rightarrow$ prefer to use as few rules as possible
- Unsupervised Bayesian inference for PCFGs from strings:
  - MCMC sampling
  - Variational Bayes

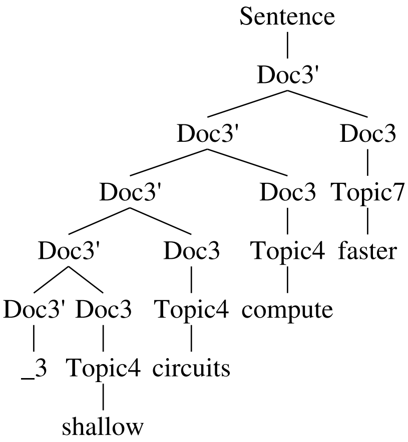
See: Kurihara and Sato (2006), Johnson et al (2007)

MACQUARIE UNIVERSITY

LDA topic models as PCFGs (1)

- Prefix strings from document j with a *document identifier* "$\_j$"

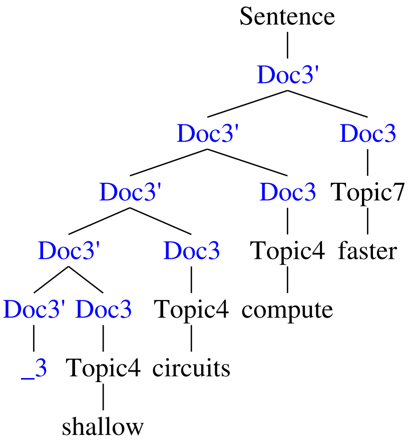
$$
\begin{array}{lll}
\text{Sentence} \rightarrow & \text{Doc}'_j & j \in 1, \ldots, m \\
\text{Doc}'_j \rightarrow & \_j & j \in 1, \ldots, m \\
\text{Doc}'_j \rightarrow & \text{Doc}'_j \; \text{Doc}_j & j \in 1, \ldots, m \\
\text{Doc}_j \rightarrow & \text{Topic}_i & i \in 1, \ldots, \ell \\
 & & j \in 1, \ldots, m \\
\text{Topic}_i \rightarrow & w & i \in 1, \ldots, \ell \\
 & & w \in \mathcal{V}
\end{array}
$$



LDA topic models as PCFGs (2)

- Spine *propagates document id up through tree*

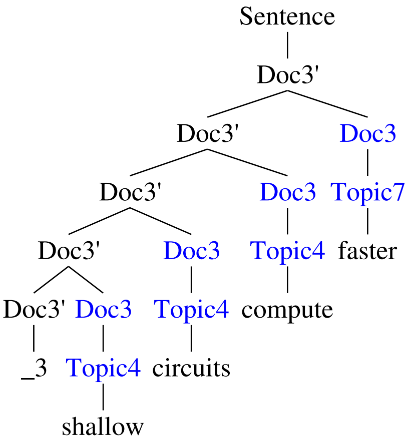
$$
\begin{array}{lll}
\mathrm{Sentence} \to & \mathrm{Doc}'_j & j \in 1, \ldots, m \\
\mathrm{Doc}'_j \to & \_j & j \in 1, \ldots, m \\
\mathrm{Doc}'_j \to & \mathrm{Doc}'_j \; \mathrm{Doc}_j & j \in 1, \ldots, m \\
\mathrm{Doc}_j \to & \mathrm{Topic}_i & i \in 1, \ldots, \ell \\
 & & j \in 1, \ldots, m \\
\mathrm{Topic}_i \to & w & i \in 1, \ldots, \ell \\
 & & w \in \mathcal{V}
\end{array}
$$



LDA topic models as PCFGs (3)

- $\mathrm{Doc}_j \to \mathrm{Topic}_i$ rules map *documents to topics*

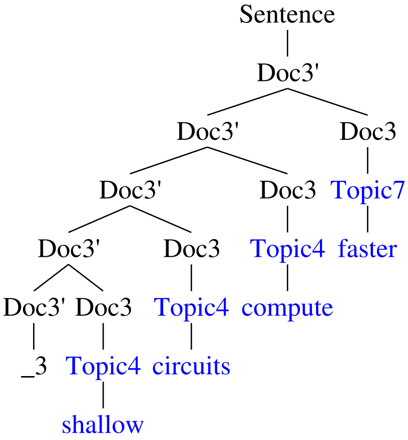
$$\begin{array}{llll}
\mathrm{Sentence} \to & \mathrm{Doc}'_j & & j \in 1, \ldots, m \\
\mathrm{Doc}'_j \to & _j & & j \in 1, \ldots, m \\
\mathrm{Doc}'_j \to & \mathrm{Doc}'_j & \mathrm{Doc}_j & j \in 1, \ldots, m \\
\mathrm{Doc}_j \to & \mathrm{Topic}_i & & i \in 1, \ldots, \ell \\
 & & & j \in 1, \ldots, m \\
\mathrm{Topic}_i \to & w & & i \in 1, \ldots, \ell \\
 & & & w \in \mathcal{V}
\end{array}$$



LDA topic models as PCFGs (4)

- $\text{Topic}_i \rightarrow \; w$ rules map *topics to words*

$$
\begin{array}{ll}
\text{Sentence} \rightarrow \; \text{Doc}'_j & j \in 1, \ldots, m \\
\text{Doc}'_j \rightarrow \; \_j & j \in 1, \ldots, m \\
\text{Doc}'_j \rightarrow \; \text{Doc}'_j \; \text{Doc}_j & j \in 1, \ldots, m \\
\text{Doc}_j \rightarrow \text{Topic}_i & i \in 1, \ldots, \ell \\
 & j \in 1, \ldots, m \\
\text{Topic}_i \rightarrow \; w & i \in 1, \ldots, \ell \\
 & w \in \mathcal{V}
\end{array}
$$



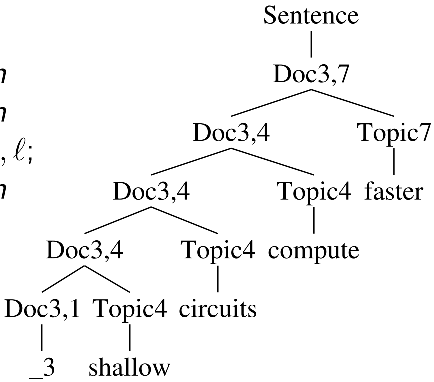
MACQUARIE UNIVERSITY

LDA topic models as PCFGs (5)

- *Not* suggesting blind use of PCFG inference for topic models
  - One iteration of LDA inference is *linear* in document length
  - One iteration of PCFG inference is *cubic* in document length
- Reduction of LDA topic models to PCFGs suggests ways of extending both kinds of models

“Sticky” topic models

$$
\begin{array}{lll}
\mathrm{Sentence} \rightarrow & \mathrm{Doc}_{j,i} & i \in 1, \ldots, \ell; \\
 & & j \in 1, \ldots, m \\
\mathrm{Doc}_{j,1} \rightarrow & \_j & j \in 1, \ldots, m \\
\mathrm{Doc}_{j,i} \rightarrow & \mathrm{Doc}_{j,i'} \ \mathrm{Topic}_i & i, i' \in 1, \ldots, \ell; \\
 & & j \in 1, \ldots, m \\
\mathrm{Topic}_i \rightarrow & w & i \in 1, \ldots, \ell; \\
 & & w \in \mathcal{V}
\end{array}
$$



- Prefer *adjacent words to have same topic*
- $\mathrm{Doc}_{j,i}$ means “document j, topic i”
- *Non-uniform Dirichlet prior* disprefers topic shift
  - $\alpha_{\mathrm{Doc}_{j,i} \rightarrow \mathrm{Doc}_{j,i} \ \mathrm{Topic}_i} \gg \alpha_{\mathrm{Doc}_{j,i} \rightarrow \mathrm{Doc}_{j,i'} \ \mathrm{Topic}_i}$ for $i' \neq i$

Outline

LDA topic models as PCFGs

Adaptor grammars

Finding topic-specific collocations

Learning the structure of proper nouns

Conclusion

MACQUARIE UNIVERSITY

From Multinomials to Dirichlet Processes

- Dirichlet Processes (DPs) are the *infinite-dimensional generalisation* of Dirichlet-Multinomials
- *Predictive distribution:* predict z_{n+1} given observations $\boldsymbol{z} = (z_1, \ldots, z_n)$
  - *Finite set* of outcomes $(1, \ldots, m)$:
    Dirichlet-multinomial with prior $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$

    $$\mathrm{P}(Z_{n+1} = k \mid \boldsymbol{z}) \quad \propto \quad n_k(\boldsymbol{z}) + \alpha_k$$

    where $n_k(\boldsymbol{z})$ is the number of times k appears in $\boldsymbol{z} = (z_1, \ldots, z_n)$
  - *Infinite set* of outcomes Ω:
    Dirichlet process $\mathrm{DP}(\alpha, \mathrm{P}_0)$ with *base distribution* $\mathrm{P}_0(Z)$ and *concentration parameter* α

    $$\mathrm{P}(Z_{n+1} = z' \mid \boldsymbol{z}) \quad \propto \quad n_{z'}(\boldsymbol{z}) + \alpha \mathrm{P}_0(z')$$

Dirichlet Processes as Adaptors

- DPs generalise Dirichlet-multinomials

$$\mathrm{P}(Z_{n+1} = z' \mid \boldsymbol{z}) \quad \propto \quad n_{z'}(\boldsymbol{z}) + \alpha \mathrm{P}_0(z')$$

- DPs follow a *"rich get richer"* law
  - frequent outcomes are increasingly likely to be predicted
- The DP is *stochastic:*
  in general, every sample $\boldsymbol{z} = (z_1, z_2, \ldots)$ is different
  $\Rightarrow$ DPs map a *base distribution* P_0 to a *distribution over distributions* $\mathrm{DP}(\alpha, \mathrm{P}_0)$
- *Pitman-Yor Processes* (PYPs) generalise Dirichlet Processes
- An *adaptor* is a function that *maps a base distribution* P_0 *to a distribution over distributions* with the same support as P_0
  - Dirichlet Processes and Pitman-Yor Processes are adaptors

Adaptor grammars as generalised PCFGs

- An *adaptor grammar* is a PCFG with a set $A \subseteq N$ of *adapted nonterminals*, and *adaptors* C_X for each $X \in A$
- *Dirichlet Process Adaptor Grammar:*
  - If $X \in W$ then $G_X(X) = 1$ (all mass on singelton tree X)
  - If $X \in N \setminus A$ is *not adapted* then X expands as in PCFG, i.e.,:

$$G_X \quad = \sum_{X \to Y_1 \ldots Y_m \in R_X} \theta_{X \to Y_1 \ldots Y_m} \mathrm{TD}_X(G_{Y_1}, \ldots, G_{Y_m})$$

  - If $X \in A$ is *adapted*, then PCFG distribution is adapted:

$$\begin{aligned} G_X &\sim \mathrm{DP}(\alpha, H_X) \\ H_X &= \sum_{X \to Y_1 \ldots Y_m \in R_X} \theta_{X \to Y_1 \ldots Y_m} \mathrm{TD}_X(G_{Y_1}, \ldots, G_{Y_m}) \end{aligned}$$

- Other kinds of adaptor grammars use different adaptors
  - *Pitman-Yor adaptor grammars* use Pitman-Yor Processes as adaptors

Predictive distribution of DP adaptor grammars

- Predictive distribution: predict next tree t_{n+1} given previously generated trees $\boldsymbol{t} = (t_1, \ldots, t_n)$
- Predictive model "caches" adapted subtrees:
  - An *unadapted nonterminal* B expands using $B \to \beta$ with probability $\theta_{B\to\beta}$
  - Each adapted nonterminal B is associated with a DP that caches previously generated subtrees in $\mathcal{T}_B$
  - An *adapted nonterminal* B expands:
    - to a subtree $t' \in \mathcal{T}_B$ probability proportional to the number of times t' was previously generated
    - using $B \to \beta$ with probability proportional to $\alpha\,\theta_{B\to\beta}$

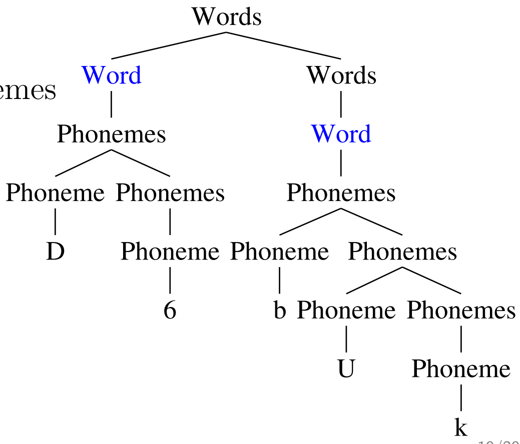
MACQUARIE UNIVERSITY

Adaptor grammars for word segmentation

- Input: phoneme sequences with *sentence boundaries* (Brent)
- Task: identify *word boundaries*, and hence words

y △ u ▲ w △ a △ n △ t ▲ t △ u ▲ s △ i ▲ D △ 6 ▲ b △ U △

Words → Word
Words → Word Words
<u>Word</u> → Phonemes
Phonemes → Phoneme Phonemes
Phonemes → Phoneme
Phoneme → a | . . . | z



- Adapted nonterminals (e.g., <u>Word</u>) highlighted and underlined

MACQUARIE UNIVERSITY

Outline

LDA topic models as PCFGs

Adaptor grammars

Finding topic-specific collocations

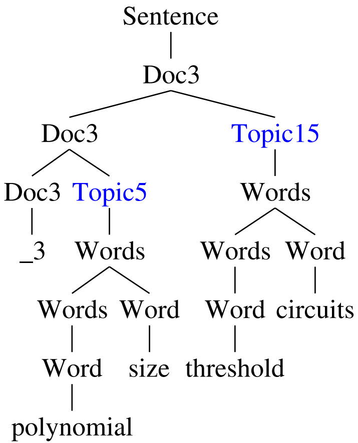
Learning the structure of proper nouns

Conclusion

Topic model with collocations

- Combines *PCFG topic model* and *segmentation adaptor grammar*

$$
\begin{array}{ll}
\text{Sentence} \rightarrow \text{Doc}_j & j \in 1, \ldots, m \\
\text{Doc}_j \rightarrow \_j & j \in 1, \ldots, m \\
\text{Doc}_j \rightarrow \text{Doc}_j \ \text{Topic}_i & i \in 1, \ldots, \ell; \\
 & j \in 1, \ldots, m \\
\underline{\text{Topic}_i} \rightarrow \text{Words} & i \in 1, \ldots, \ell \\
\text{Words} \rightarrow \text{Word} & \\
\text{Words} \rightarrow \text{Words Word} & \\
\text{Word} \rightarrow w & w \in \mathcal{V}
\end{array}
$$



Finding topical collocations in NIPS abstracts

- Run topical collocation adaptor grammar on NIPS corpus
- Run with $\ell = 20$ topics (i.e., 20 distinct Topic_i nonterminals)
- Corpus is segmented by punctuation
  - terminal strings are fairly short

  $\Rightarrow$ inference is fairly efficient
- Used Pitman-Yor adaptors
  - sampled Pitman-Yor *a* and *b* parameters
  - flat and "vague Gamma" priors on Pitman-Yor *a* and *b* parameters

See: Griffiths et al (2007), Johnson and Goldwater (2009)

MACQUARIE UNIVERSITY

Sample output on NIPS corpus, 20 topics

- Multiword subtrees learned by adaptor grammar:

T_0 → gradient descent	T_1 → associative memory
T_0 → cost function	T_1 → standard deviation
T_0 → fixed point	T_1 → randomly chosen
T_0 → learning rates	T_1 → hamming distance
T_3 → membrane potential	T_10 → ocular dominance
T_3 → action potentials	T_10 → visual field
T_3 → visual system	T_10 → nervous system
T_3 → primary visual cortex	T_10 → action potential

- Sample skeletal parses:

_3 (T_5 polynomial size) (T_15 threshold circuits)
_4 (T_11 studied) (T_19 pattern recognition algorithms)
_4 (T_2 feedforward neural network) (T_1 implements)
_5 (T_11 single) (T_10 ocular dominance stripe) (T_12 low)
(T_3 ocularity) (T_12 drift rate)

MACQUARIE UNIVERSITY

Outline

LDA topic models as PCFGs

Adaptor grammars

Finding topic-specific collocations

Learning the structure of proper nouns

Conclusion

MACQUARIE UNIVERSITY

Learning the structure of proper nouns

- Grammars offer *structural* and *positional sensitivity* not captured in topic models: can we use this somehow?
- The Penn WSJ assigns flat structures to names and other base NPs
- Identifying structure within names can be useful
  - *Bill Clinton* and *Hillary Clinton* are unlikely to corefer because *Bill* and *Hillary* are both first names
  - *Secretary Clinton* and *Hillary Clinton* can corefer because *Secretary* is an honorific
- There are many different types of names (e.g., company names, person names)
- Some components of a name can be filled by multi-word sequences
  - In *Jean-Claude van Damme*, *van Damme* is the surname

An adaptor grammar for names

$$\begin{array}{ll}
\mathrm{NP} \to (\mathrm{A0})\ (\mathrm{A1})\ \ldots\ (\mathrm{A6}) & \mathrm{NP} \to (\mathrm{B0})\ (\mathrm{B1})\ \ldots\ (\mathrm{B6}) \\
\underline{\mathrm{A0}} \to \mathrm{Word}^+ & \underline{\mathrm{B0}} \to \mathrm{Word}^+ \\
\ldots & \ldots \\
\underline{\mathrm{A6}} \to \mathrm{Word}^+ & \underline{\mathrm{B6}} \to \mathrm{Word}^+ \\
\mathrm{NP} \to \mathrm{Unordered}^+ & \underline{\mathrm{Unordered}} \to \mathrm{Word}^+
\end{array}$$

- *Sample parses:*

  (A0 barrett) (A3 smith)
  (A0 albert) (A2 j.) (A3 smith) (A4 jr.)
  (A0 robert) (A2 b.) (A3 van dover)
  (B0 aim) (B1 prime rate) (B2 plus) (B5 fund) (B6 inc.)
  (B0 balfour) (B1 maclaine) (B5 international) (B6 ltd.)
  (B0 american express) (B1 information services) (B6 co)
  (U abc) (U sports)
  (U sports illustrated)
  (U sports unlimited)

See: Elsner, Charniak and Johnson (2009)

MACQUARIE UNIVERSITY

Outline

LDA topic models as PCFGs

Adaptor grammars

Finding topic-specific collocations

Learning the structure of proper nouns

Conclusion

MACQUARIE UNIVERSITY

Conclusion

- LDA topic models can be expressed as Bayesian PCFGs
  - makes it easier to combine grammars and topic models
  - may help us to design new topic models that incorporate configurational sensitivity that is easy to express with grammars
- Adaptor grammars are a non-parametric extension of PCFGs which associate probabilities with entire subtrees
- Adaptor grammars can be used to express generalised topic models
  - learning topical collocations
  - learning the structure of names

MACQUARIE UNIVERSITY

Interested in Bayesian Inference and Language?

We're recruiting *PhD students* and *post-docs*.

Contact *Mark.Johnson@mq.edu.au* for more information.