Unsupervised and Semi-supervised learning of Nonparametric Bayesian word segmentation

Daichi Mochihashi
NTT Communication Science Laboratories, Japan
daichi@cslab.kecl.ntt.co.jp

NOE Statistical Machine Learning Seminar
2011-1-19 (Wed)
ISM

Word segmentations in NLP

- Segment string into “words”

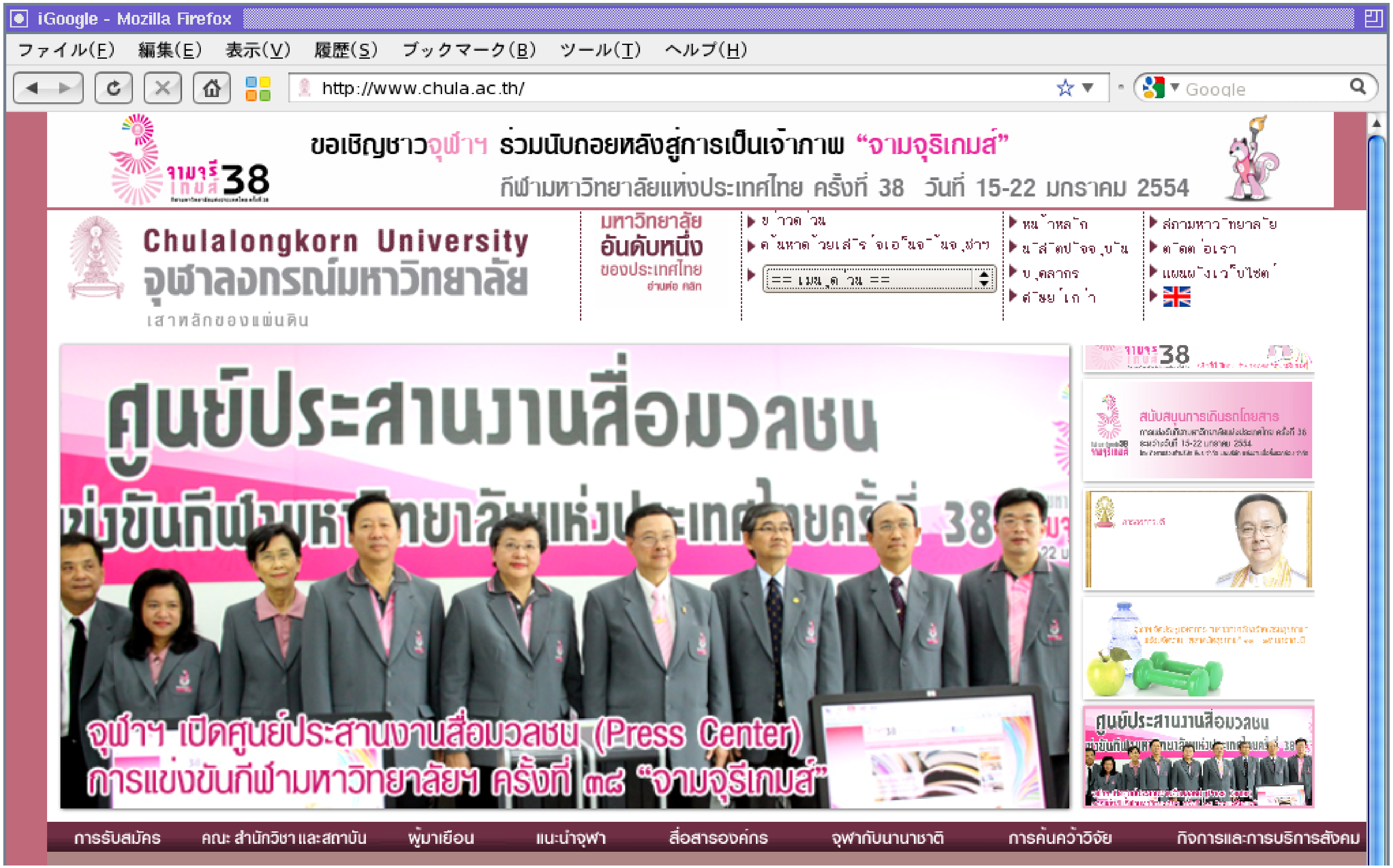
```
% echo “やあこんにちは, 統数研はどうですか。“
 | mecab -O wakati
やあ こんにちは , 統数研 は どう です か 。
(やあこん にち は , 統 数研 はどう で すか 。) ×
```

- Very important problem for unsegmented languages like Japanese, Chinese, Thai, Hebrew, ...

Chinese

Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://www.tsinghua.edu.cn/qhdwzy/board_xxqk2.jsp?board=12&bid2=1201&pagen
Google
清華大學
Tsinghua University
学校概况 | 院系设置 | 管理机构 | 科学研究 | 教师队伍 | 人才培养 | 招生就业 | 虚拟校园
学校沿革
历任校长
统计资料
校长致辞
喜迎百年华诞 再铸新的辉煌
——2011年新年献辞
●校长 顾秉林
一世纪沧桑砥砺，一百年春华秋实。此时此刻，2010年的余晖散去，2011年的曙光在前！此时此刻，清华正在送走她的第一个百年，迈向新百年的征程！在辞旧迎新、钟声回荡之际，请允许我代表学校，向清华全体同学、教职员工、离退休人员和海内外广大校友，向长期关心支持清华发展的各界人士，致以崇高的敬意和新年的祝福！
一百年来，清华大学的发展始终与国家民族的命运休戚与共，形成了优良的精神传统和鲜明的办学特色。一代代清华人"自强不息、厚德载物"，涌现出众多学术大师、兴业英才和治国栋梁，为中国社会进步和世界文明发展作出了重要贡献。特别是近年来，在国家的大力支持下，学校致力于世界一流大学建设，积极探索中国特色的"大学之道"，各项事业不断取得新的进展，正在跻身世界一流大学的行列。
大学之道，育人为本。一年来，以"清华新百年人才培养的使命与战略"为主题的第23次教育工作讨论会顺利举行，全校师生就推动办学优势转化、培养拔尖创新人才进一步取得共识。"清华学堂人才培养计划"以及多项教育教学改革措施相继实施。招生工作大力推进多元评价、兼顾拔尖与公平，生源质量进一步提高。就业毕业生中，超过80%选择到国家重要行业和领域建功立业。外国留学生规模不断扩大，结构进一步优化，外国研究生在学规模居全国高校首位。
大学之道，学术为魂。一年来，我校积极面向国际学术研究前沿和国家重大战略需求开展高

Thai



iGoogle - Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://www.chula.ac.th/
Google
ขอเชิญชาวจุฬาฯ ร่วมนับถอยหลังสู่การเป็นเจ้าภาพ “จามจุรีเกมส์”
กีฬามหาวิทยาลัยแห่งประเทศไทย ครั้งที่ 38 วันที่ 15-22 มกราคม 2554
Chulalongkorn University
จุฬาลงกรณ์มหาวิทยาลัย
เสาหลักของแผ่นดิน
มหาวิทยาลัย
อันดับหนึ่ง
ของประเทศไทย
ศูนย์ประสานงานสื่อมวลชน
จุฬาฯ เปิดศูนย์ประสานงานสื่อมวลชน (Press Center)
การแข่งขันกีฬามหาวิทยาลัยฯ ครั้งที่ ๓๘ “จามจุรีเกมส์”
สนับสนุนการเดินรถโดยสาร
ศูนย์ประสานงานสื่อมวลชน
การรับสมัคร
คณะ สำนักวิชา และสถาบัน
ผู้มาเยือน
แนะนำจุฬา
สื่อสารองค์กร
จุฬากับนานาชาติ
การค้นคว้าวิจัย
กิจการและการบริการสังคม

Persian

(Isfahan university of technology, Iran)

Word segmentations in NLP

- Crucial first step for unsegmented languages like Japanese, Chinese, Thai, Hebrew, …
  - Very important and fundamental problem
  - Especially: Chinese (1,200,000,000 speakers)
  - SIGHAN word segmentation Bakeoff: 2003-2010
    - Many intensive research!

Word segmentation (2)

- Learning methods so far: *Supervised*

```
# S-ID:950117245-006 KNP:99/12/27
* 0 5D
一方 いっぽう * 接続詞 * * *
、 、 * 特殊 読点 * *
* 1 5D
震度 しんど * 名詞 普通名詞 * *
は は * 助詞 副助詞 * *
```

Standard “Kyoto corpus”: 38400 **hand-segmented** newspaper sentences

- “Correct” segmentations with huge human effort
  - CRF, SVM, Maxent, .. classifiers to learn
- “Correct” segmentations for speech? Twitter? Unknown language..?
  - |女御|更衣|あ|また|さ|ぶら|ひ|た|ま|ひける|中|に|、|…

MeCab analysis of *“The Tale of Genji”*, AD1000

Overview

- Unsupervised learning of word segmentation
  - HPYLM^2 prior + Blocked MCMC
  - “Words” for even unknown language!
- Semi-supervised learning of word segmentation
  - Systematic integration with CRF
  - Markov<->semi-Markov conversion (general)
  - MCMC-EM like algorithm for inference

Unsupervised Word Segmentation

- Basic idea: Maximize the probability of the segmentation $\mathbf{w}$ of a string s :

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}}\ p(\mathbf{w}|s)\ .$$

  - Ex. p(he sings a song) > p(hes ingsao ng)
  - *No dictionary, no "correct" supervised data*
  - Find the "most natural segmentation" of a string
- Note: Exponential number of candidates
  - A sentence of 50 characters:
    2^{50}= 1,125,899,906,842,624 different segmentations

Probability of a sentence: *n*-grams

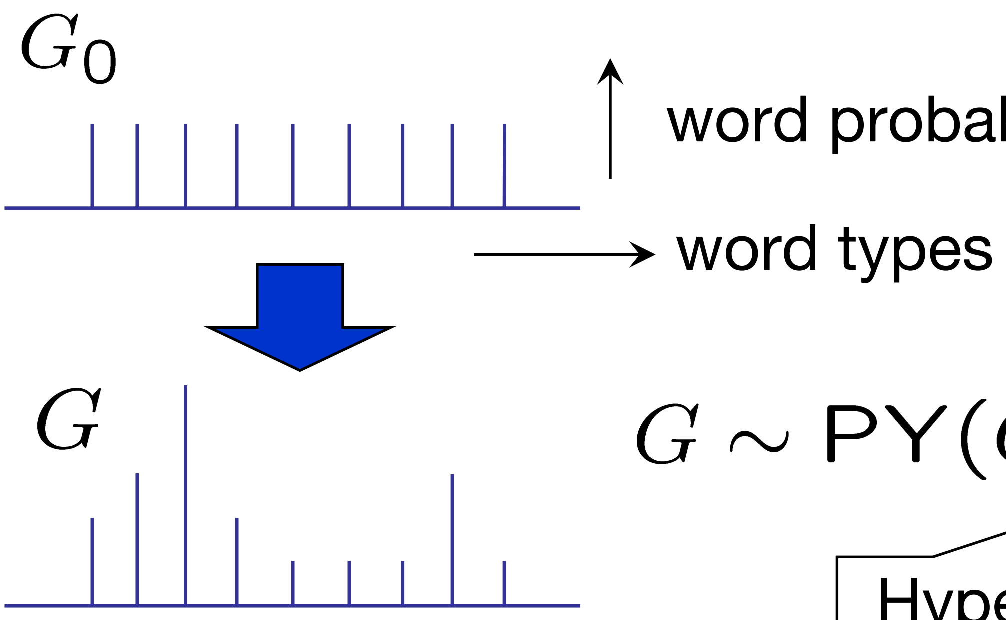
p(she likes music)
$= p(\text{she}|\wedge)\ p(\text{likes}|\text{she})\ p(\text{music}|\text{likes})\ p(\$|\text{music})$

- Markov model on words: very common in NLP
  - First introduced by Shannon (1948)
- PCFG is ok too.. but simple model will suffice
  - PCFG on Twitter yellings??
- Probability tables are mostly 0
  - Hierarchical smoothing is necessary
  - Every substring could be a “word”

➡ Bayesian treatment: HPYLM (yw’s talk)

Pitman-Yor n-gram model

- The Pitman-Yor (=Poisson-Dirichlet) process:
  - Draw distribution from distribution
  - Extension of Dirichlet process (w/ frequency discount)

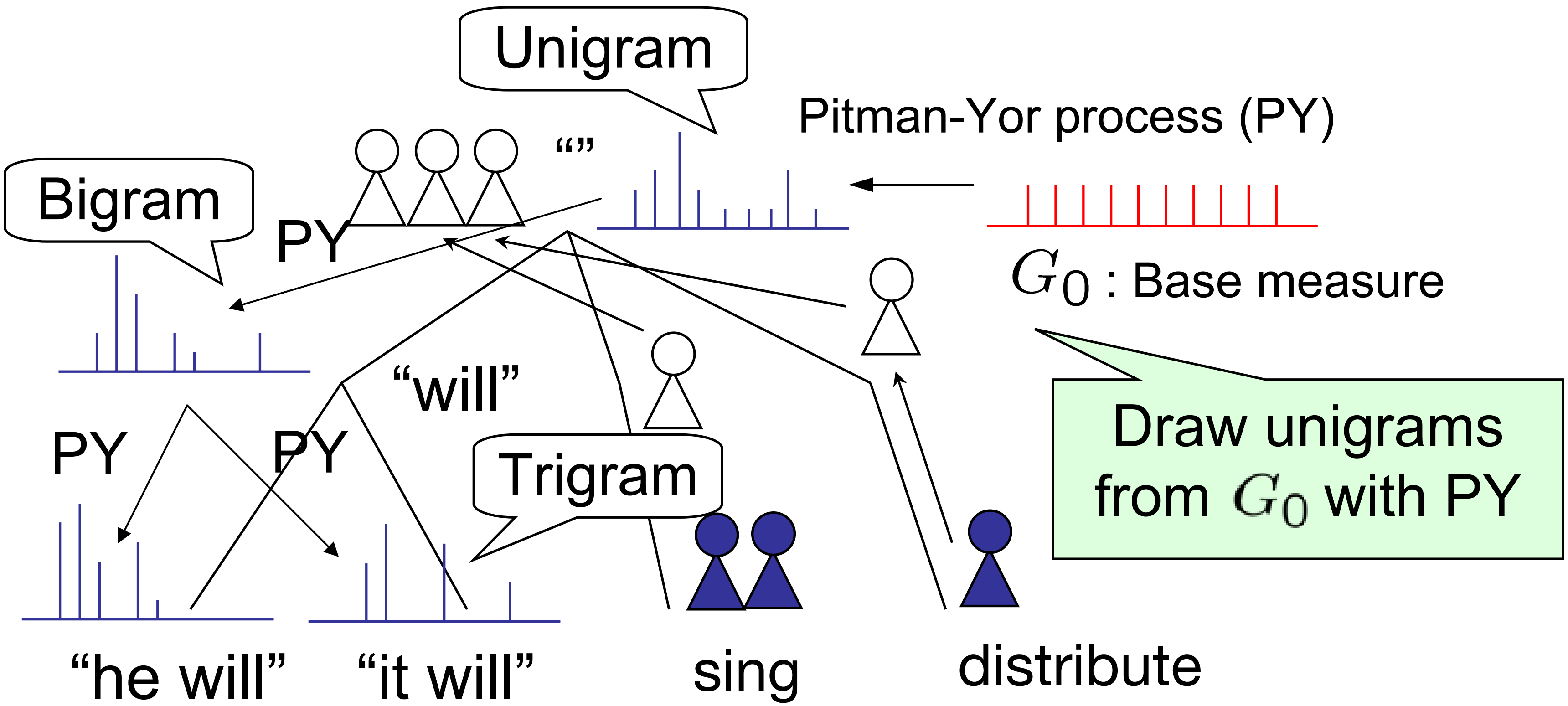


Called "base measure"

$$G \sim \mathrm{PY}(G_0, \boxed{\theta, d})$$

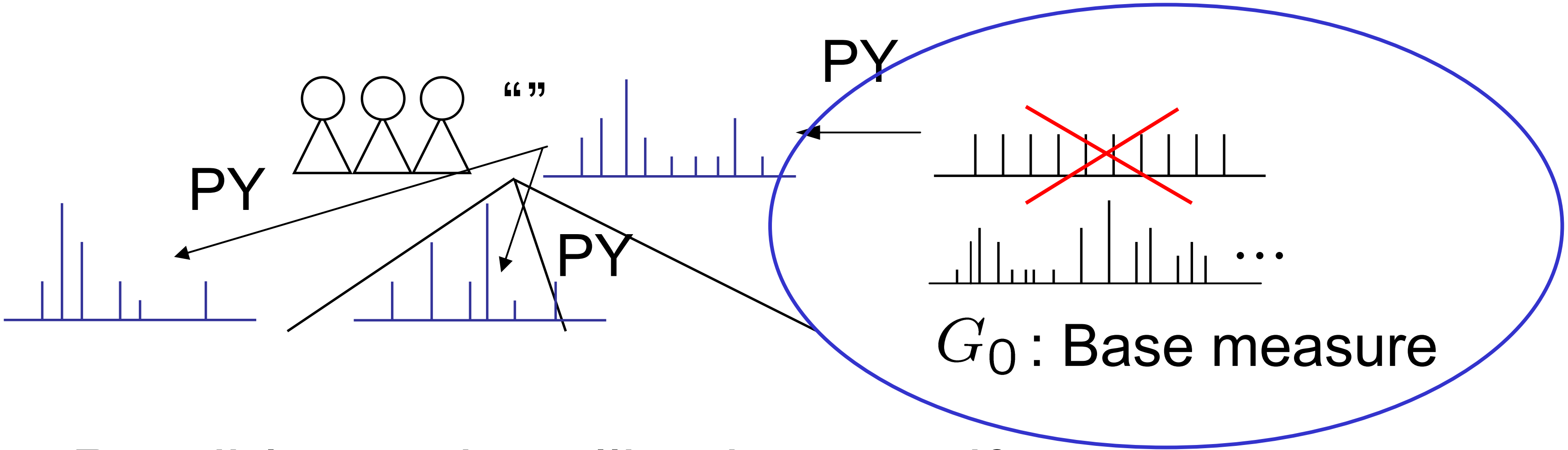
Hyperparameters (can be learned)

Hierarchical Pitman-Yor n-gram



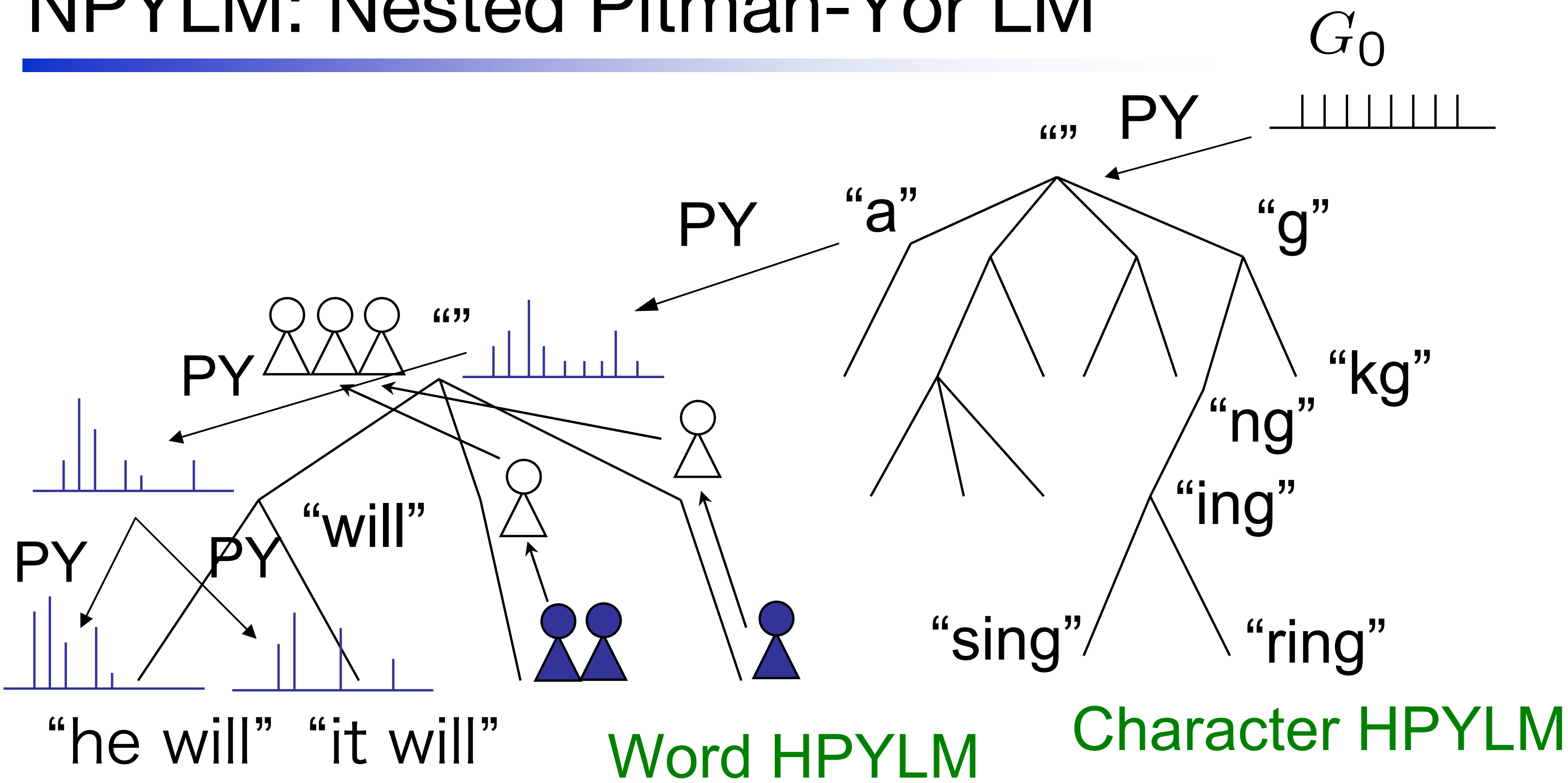
- Kneser-Ney smoothing is an approximation of hierarchical Pitman-Yor process (Teh, ACL 2006)
  - HPYLM = “Bayesian Kneser-Ney n-gram”

Problem: Word spelling



- Possible word spelling is not uniform
  - Likely: “will”, “language”, “hierarchically”, …
  - Unlikely: “illbe”, “nguag”, “ierarchi”, …
- Replace the base measure using character information
  →Character HPYLM!

NPYLM: Nested Pitman-Yor LM



- Character n-gram embedded in the base measure of word n-gram
  - i.e. hierarchical Markov model
  - Poisson word length correction (see the paper)

Inference and Learning

- Simply maximize the probability of strings
  - i.e. minimize the perplexity per character of LM
- X : Set of strings $s_1, s_2, \cdots, s_N$
  Z : Set of hidden word segmentation indicators
  $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N$

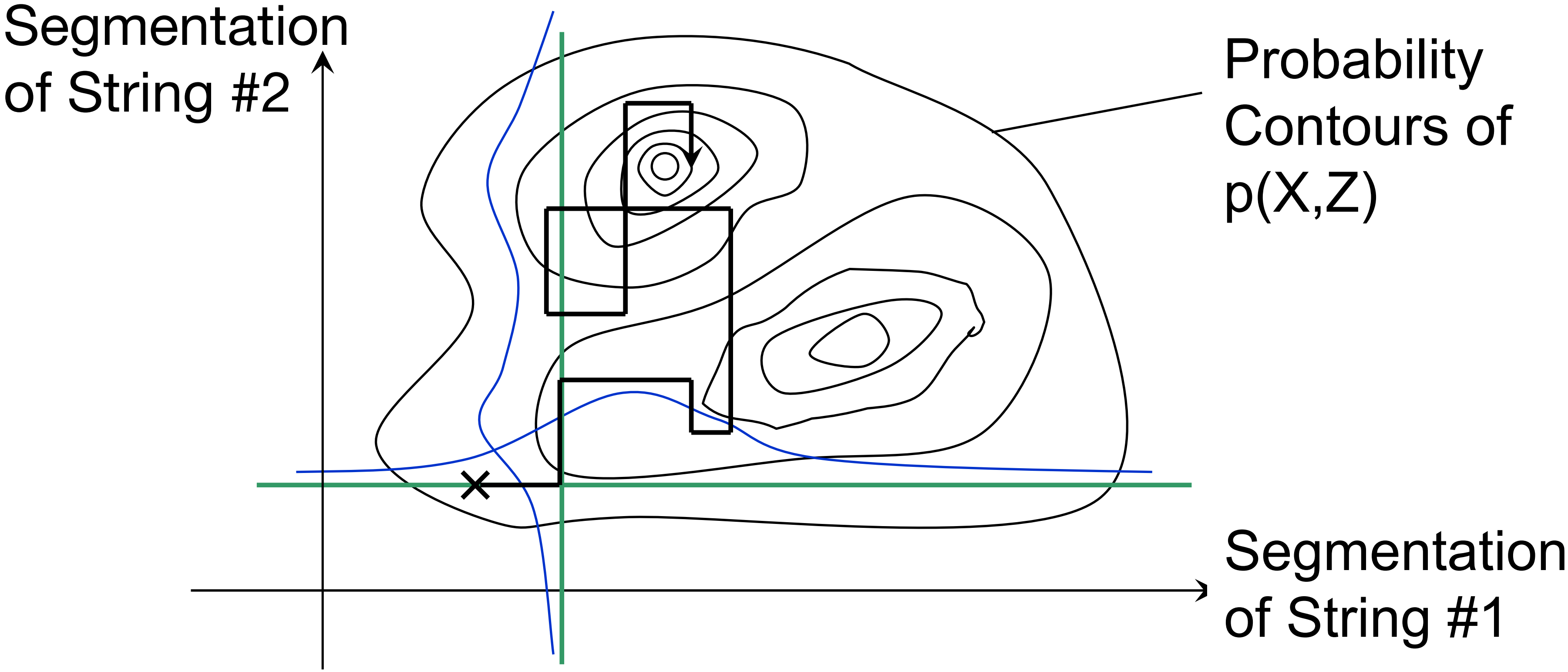
$$p(X) = \prod_n p(s_n)$$

$$p(s_n) = \sum_{\mathbf{z}_n} p(s_n, \mathbf{z}_n)$$

Hidden word segmentation of string s_n

  - Notice: *Exponential possibilities* of segmentations!

Blocked Gibbs Sampling



- Sample word segmentation block-wise for each sentence (string)
  - High correlations within a sentence

Blocked Gibbs Sampling (2)

- Iteratively improve word segmentations: words(s) of s

0. For $s = s_1 \cdots s_N$ do
   parse_trivial(s, Θ).

Whole string is a single "word"

1. For j = 1..M do
   For s = *randperm*($s_1 \cdots s_N$) do
     Remove words(s) from NPYLM Θ
     Sample words(s) $\sim p(w|s, \Theta)$
     Add words(s) to NPYLM Θ
   done
   Sample all hyperparameters of Θ
  done

Gibbs Sampling and word segmentation

1 神戸では異人館　街の 二十棟 が破損した 。

2 神戸 では 異人館 街の 二十棟 が破損した 。

10 神戸 では 異人館 街の 二十棟 が破損した 。

50 神戸 で は異人 館 街 の 二 十 棟 が 破損 し た 。

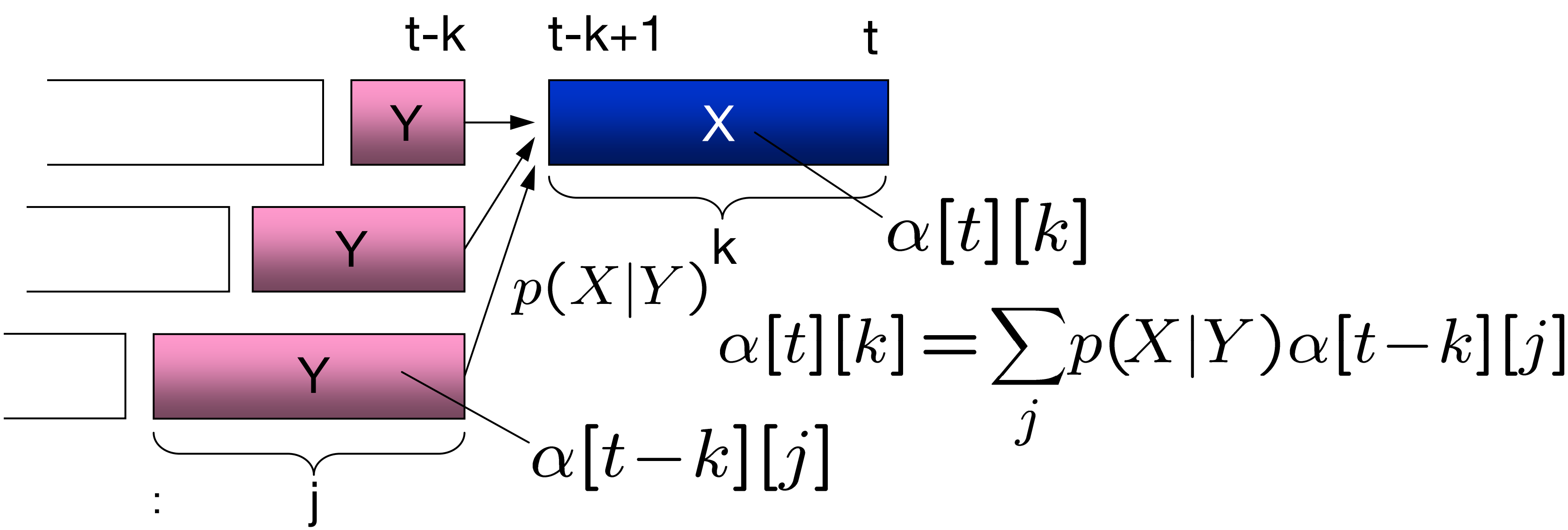
100 神戸 で は 異 人館 街 の 二 十 棟 が 破損 し た 。

200 神戸 で は 異人館 街 の 二 十 棟 が 破損 し た 。

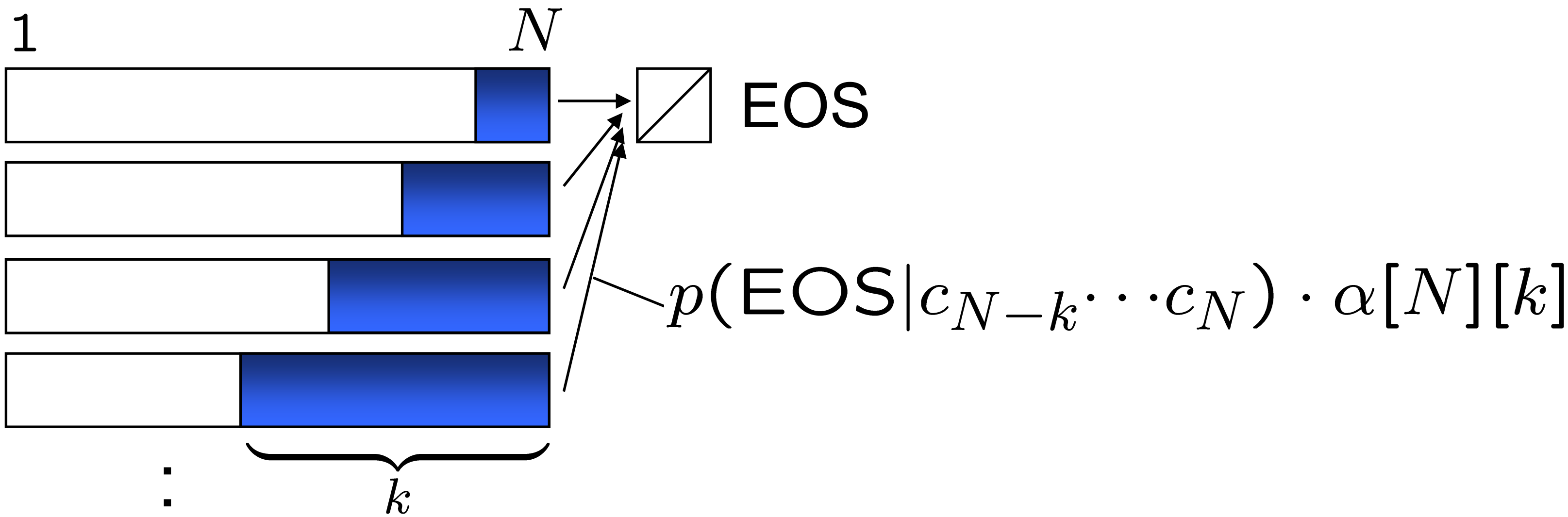
- Iteratively resample word segmentations and update language models accordingly.

Sampling through Dynamic Programming

- Forward filtering, Backward sampling (Scott 2002)
- $\alpha[t][k]$: inside probability of substring $c_1 c_2 \cdots c_t$ with the last k characters constituting a word
  - Recursively marginalize segments before the last k

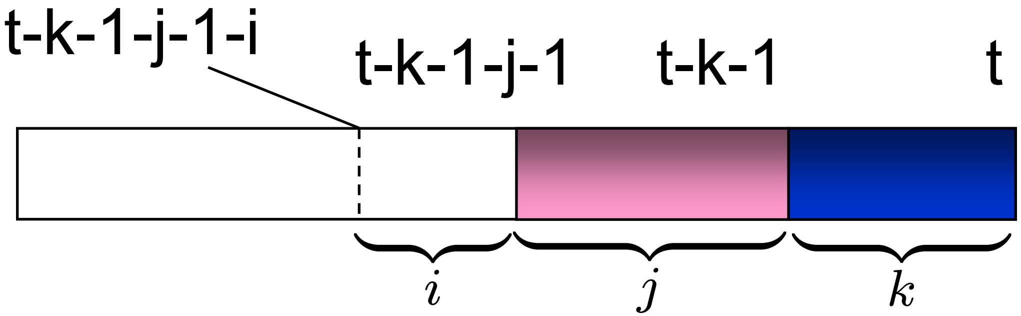


Sampling through Dynamic Programming (2)



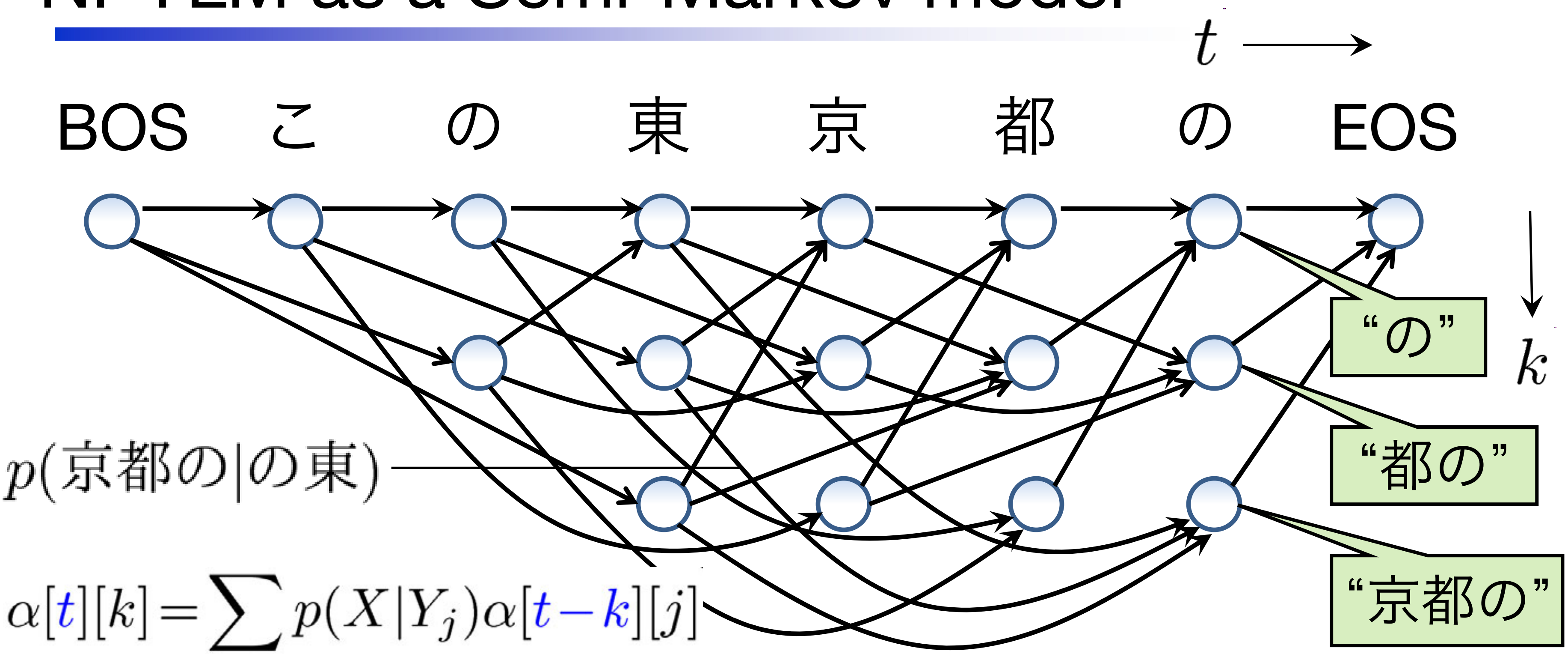
- $\alpha[N][k]$ = probability of the entire string $c_1\cdots c_N$ with the last k characters constituting a word
  - Sample k with probability to end with EOS
- Now the final word is $c_{N-k}\cdots c_N$: use $\alpha[N-k-1][k']$ to determine the previous word, and repeat

The Case of Trigrams



- In case of trigrams: use $\alpha[t][k][j]$ as an inside probability
  - $\alpha[t][k][j]$ = probability of substring with the final k chars and the further j chars before it being words
  - Recurse using $\alpha[t-k-1][j][i]$ $(i = 0 \cdots L)$
- >Trigrams? Practically not so necessary, but use Particle MCMC (Doucet+ 2009 to appear) if you wish

NPYLM as a Semi-Markov model



$$\alpha[t][k] = \sum_{j} p(X|Y_j)\,\alpha[t-k][j]$$

- Unsupervised learning of Semi-Markov HMM (Ostendorf 96, Murphy 02)
- State transition = word transition with an intensive smoothing w/ NPYLM + MCMC

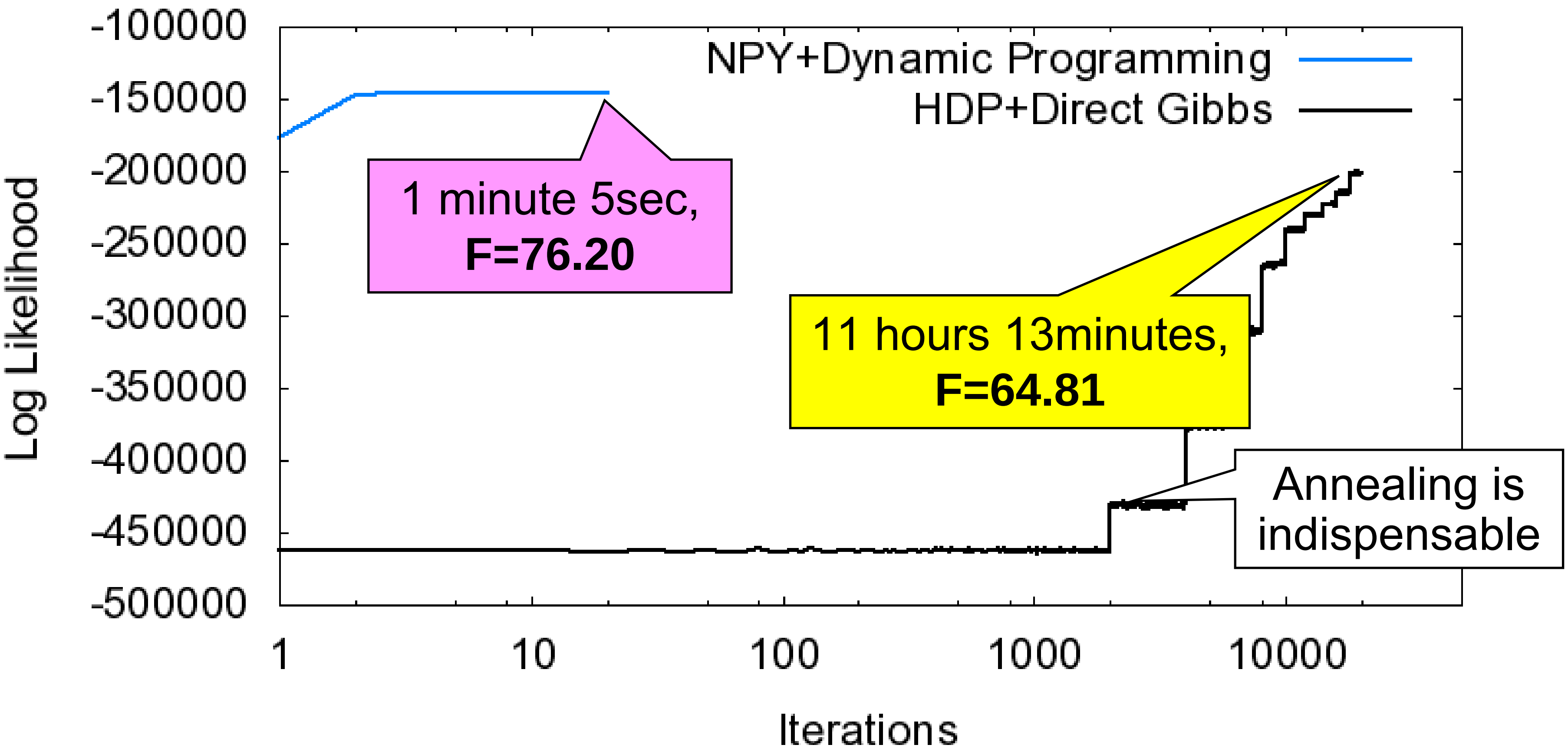
English Phonetic Transcripts

- Comparison with HDP bigram (w/o character model) in Goldwater+ (ACL 2006)
- CHILDES English phonetic transcripts
  - Recover “WAtsDIs”→”WAts DIs” (What’s this)
  - Johnson+(2009), Liang(2009) use the same data

Model	P	R	F	LP	LR	LF
NPY(3)	74.8	75.2	75.0	47.8	**59.7**	53.1
NPY(2)	74.8	**76.7**	**75.7**	57.3	56.6	57.0
HDP(2)	**75.2**	69.6	72.3	**63.5**	55.2	**59.1**

  - Very small data: 9,790 sentences, 9.8 chars/sentence

Convergence & Computational time



- NPYLM is very efficient & accurate! (600x faster here)

Chinese and Japanese

Perplexity per character

Model	MSR	CITYU	Kyoto
NPY(2)	0.802 (51.9)	**0.824** (**126.5**)	0.621 (23.1)
NPY(3)	**0.807** (**48.8**)	0.817 (128.3)	**0.666** (**20.6**)
NPY(+)	0.804 (**38.8**)	0.823 (**126.0**)	**0.682** (**19.1**)
ZK08	0.667 (—)	0.692 (—)	—

- MSR&CITYU: SIGHAN Bakeoff 2005, Chinese
- Kyoto: Kyoto Corpus, Japanese
- ZK08: Best result in Zhao&Kit (IJCNLP 2008)

Note: Japanese subjective quality is much higher (proper nouns combined, suffixes segmented, etc..)

Arabic

- Arabic Gigawords 40,000 sentences (AFP news)

الفلسطينيبسببتظاهرةلانصارحركةالمقاومةالاسلاميةحماس.

و اذاتحققذلكفانكيسلوفسكيكونقدحازثلاث ... فيابرزثلاثة

صحية

+قائد

الايقل

وقالتدانييلتومسونالتيكتبتالسيناريو.وقداستغرقاعدادهخمسةاعوام."تاريخي

Google translate:
"Filstinebsbptazahrplansarhrkpalmquaompalaslamiphamas."

NPYLM

الفلسطيني بسبب تظاهرة ل انصار حركة المقاومة الاسلامية حماس .

و اذا تحقق ذلك ف ان كيسلوفسكي يكون قد حاز ثلاث ... فيابرز ثلاثة

الصحية

سطينية ... + " .

الايقل

وقد استغرق اعداد ه خمسةاعوام . و قال ت دان ييل تومسون التي " تاريخي

Google translate:
"Palestinian supporters of the event because of the Islamic Resistance Movement, Hamas."

English ("Alice in Wonderland")

first,shedreamedoflittlealiceherself,andonceagainthetinyhandswereclaspedupo
nherknee,andthebrighteagereyeswerelookingupintohersshecouldheartheveryto
nesofhervoice,andseethatqueerlittletossofherheadtokeepbackthewanderinghai
rthatwouldalwaysgetintohereyesandstillasshelistened,orseemedtolisten,thewho
leplacearoundherbecamealivethestrangecreaturesofherlittlesister'sdream.thelo
Nggrassrustledatherfeetasthewhiterabbithurriedbythefrightenedmousesplashed
Hiswaythroughtheneighbouringpoolshecouldheartherattleoftheteacupsasthemar
chhareandhisfriendssharedtheirneverendingmeal,andtheshrillvoiceofthequeen…

first, she dream ed of little alice herself ,and once again the tiny hand s were clasped upon her knee ,and the bright eager eyes were looking up into hers -- shecould hearthe very tone s of her voice , and see that queer little toss of herhead to keep back the wandering hair that would always get into hereyes -- and still as she listened , or seemed to listen , thewhole place a round her became alive the strange creatures of her little sister 'sdream. thelong grass rustled ather feet as thewhitera bbit hurried by -- the frightened mouse splashed his way through the neighbour ing pool -- shecould hearthe rattle ofthe tea cups as the marchhare and his friends shared their never -endingme a l ,and the …

Conclusion of the first part

- Completely unsupervised word segmentation of arbitrary language strings
  - Combining word and character information via hierarchical Bayes
  - Very efficient using forward-backward+MCMC
- Directly optimizes Kneser-Ney language model
  - N-gram construction without any "word" information
  - Sentence probability calculation with all possible word segmentations marginalized out
    - Easily obtained from dynamic programming

Problems of Unsupervised segmentation?

- Optimize *n*-gram language models
  → Must be optimized for different tasks
  - For machine translation: Nguyen+ (COLING 2010)
  - For speech recognition : Neubig+ (Interspeech 2010)
- Not always fit for human standards
  - Ex. Inflection, proper nouns, human preference
    - “咲か”-“咲き”-“咲く”, “盧前大統領”, “そ の”
  - Remedy:
    1. Make the generative model more complex
    2. *Semi-supervised learning*
       - human standards are usually closed and general

Semi-supervised learning : JESS-CM

better than Druck+ (ICML2010)

- JESS-CM (Suzuki+ ACL-HLT 2008): "joint probability model embedding style semi-supervised conditional model"
  - *highest performance semi-supervised learning on POS tagging, NE chunking, dependency parsing*
- Model:

Discriminative model

Generative model

$$p(\mathbf{y}|\mathbf{x};\Lambda,\Theta) \propto p_{\mathrm{DISC}}(\mathbf{y}|\mathbf{x};\Lambda)p_{\mathrm{GEN}}(\mathbf{y},\mathbf{x};\Theta)^{\lambda}$$

  - Product model: discriminative and generative
  - However no naïve product
    - "model weight" λ is included in Λ
    - Θ is influenced by Λ through learning (i.e. recursive)

JESS-CM : inference

Discriminative model

Generative model

$$p(\mathbf{y}|\mathbf{x};\Lambda,\Theta) \propto p_{\mathrm{DISC}}(\mathbf{y}|\mathbf{x};\Lambda)p_{\mathrm{GEN}}(\mathbf{y},\mathbf{x};\Theta)^{\lambda}$$

- If the discriminative model is log-linear (like CRF):

$$p_{\mathrm{DISC}}(\mathbf{y}|\mathbf{x}) \propto \exp\left(\sum_{k=1}^{K}\lambda_k f_k(\mathbf{y},\mathbf{x})\right)$$

- Then the model is again loglinear:

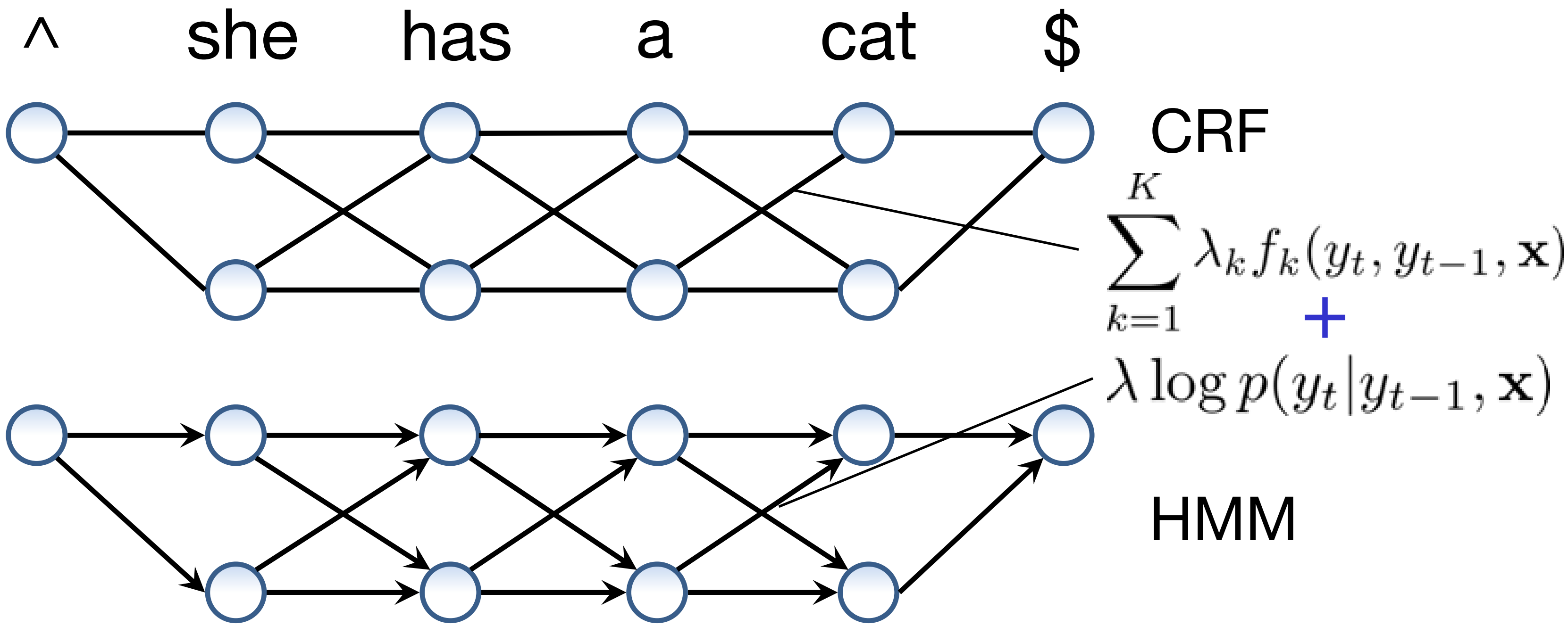
$$p(\mathbf{y}|\mathbf{x}) \propto \exp\left(\underbrace{\lambda \log p_{\mathrm{GEN}}(\mathbf{y},\mathbf{x})}_{\text{Additional feature!}} + \sum_{k=1}^{K}\lambda_k f_k(\mathbf{y},\mathbf{x})\right)$$

- Inference = maximize the objective

$$\log p(D|\Lambda,\Theta) = \log p(\mathbf{Y}_s|\mathbf{X}_s;\Lambda,\Theta) + \log p(\mathbf{X}_u;\Lambda,\Theta)$$

  – Fix Θ and optimize Λ on $\mathbf{Y}_s, \mathbf{X}_s$
  – Fix Λ and optimize Θ on $\mathbf{X}_u$

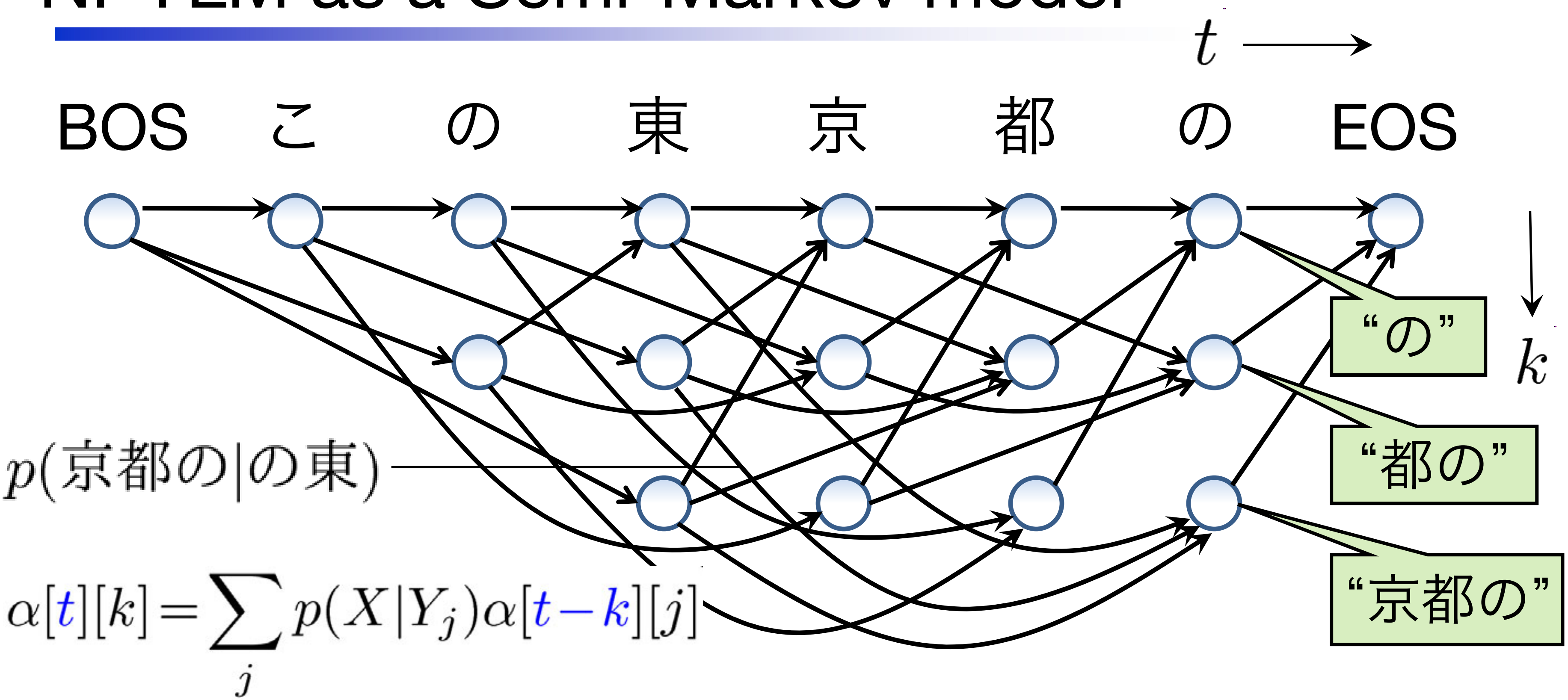
JESS-CM on CRF-HMM (Suzuki+ ACL08)



- Sum the corresponding weights of the path *on the same graphical model*

Interleave CRF and HMM optimization

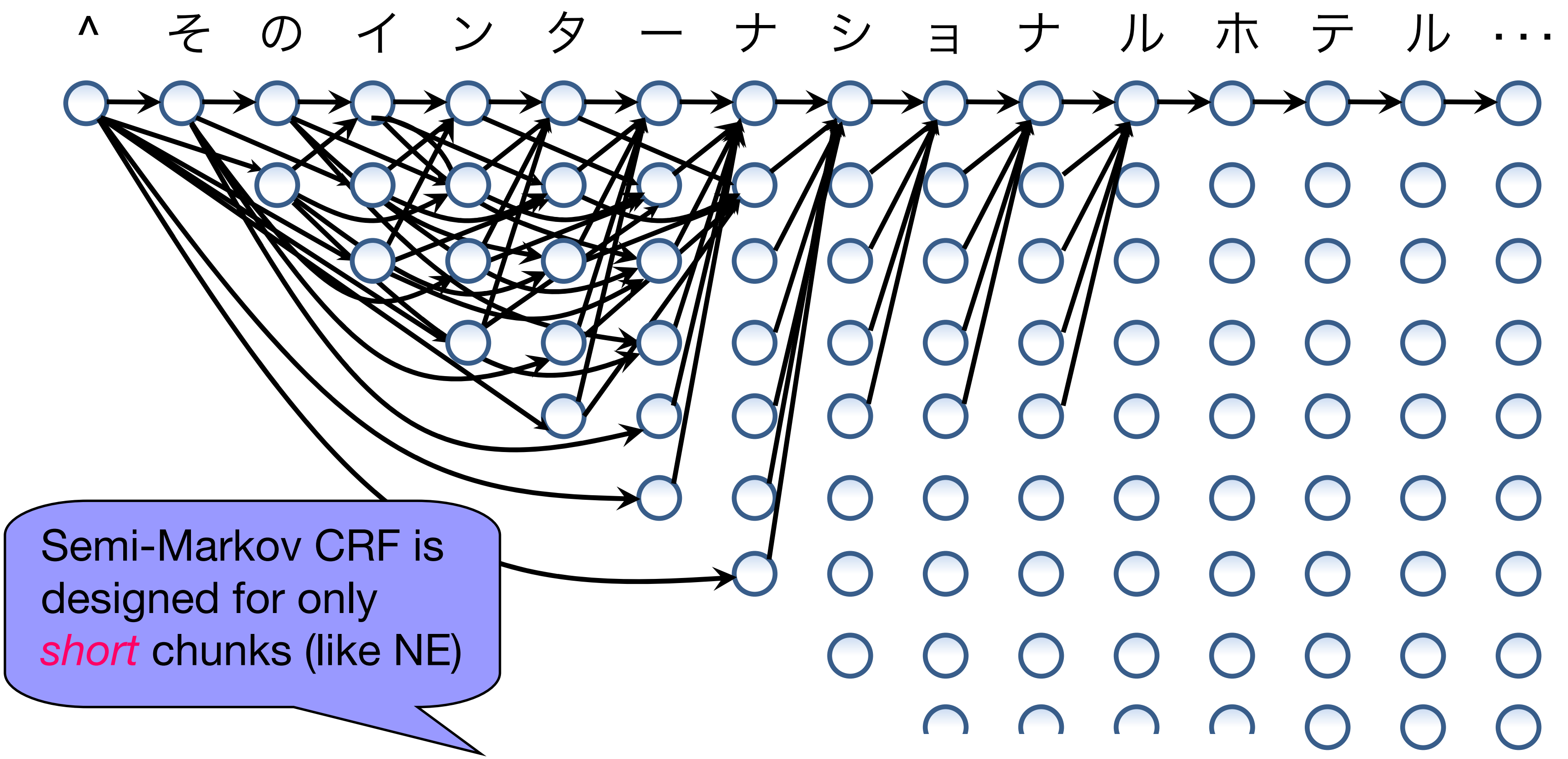
NPYLM as a Semi-Markov model



$$\alpha[t][k] = \sum_j p(X|Y_j)\alpha[t-k][j]$$

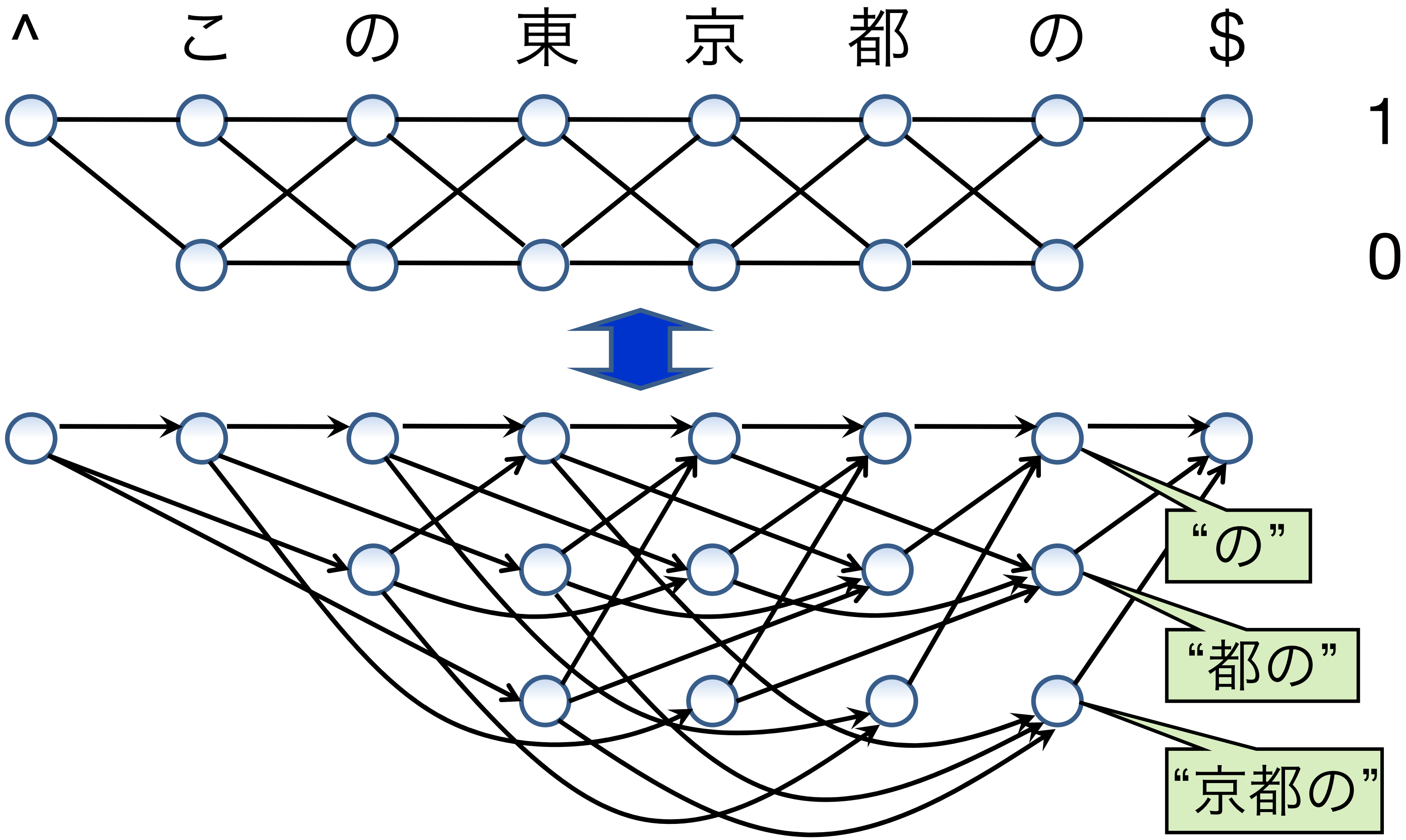
- NPYLM is not Markov, but semi-Markov
- State transition = word transition with an intensive smoothing w/ NPYLM + MCMC
  - CRF combination?

Semi-Markov CRF (NIPS 2004)?



- Enormous memory (1GB→20GB)
- (Supervised) precision: at most 95%
  - Only words, no character-level information

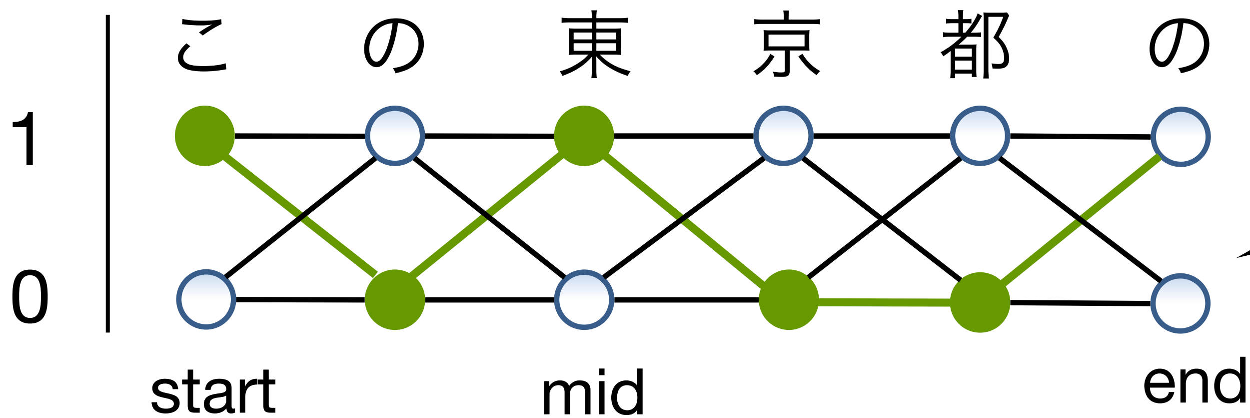
Learning Markov CRF←→Semi-Markov LM



- *How to combine two different models?*
  - CRF→NPYLM, NPYLM→CRF

CRF→NPYLM

- Easy, proposed in (Andrew+ EMNLP 2006)
  - CRF→semi-Markov CRF
  - p(“この”→“東京都”)

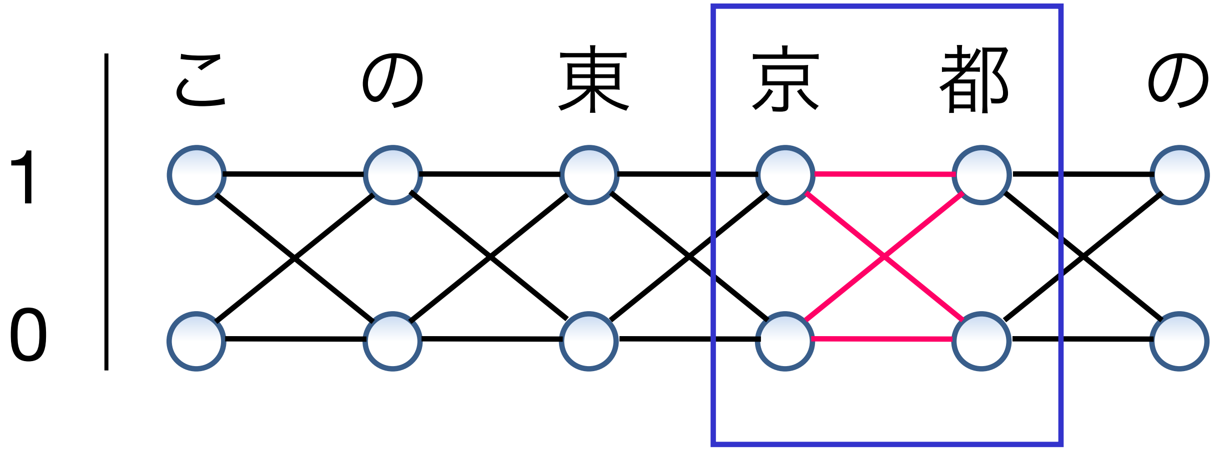


1=beginning of word,
0=continuation of word

  - Summing up the weight of features along the path
  - γ(start, mid, end)
    := γ(start, mid)+ γ(mid, end)

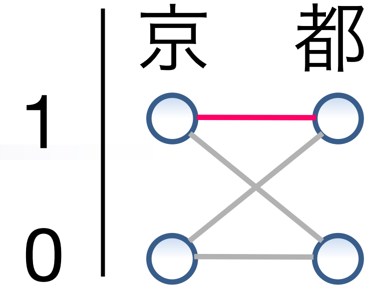
NPYLM→CRF (1)

- Nontrivial !



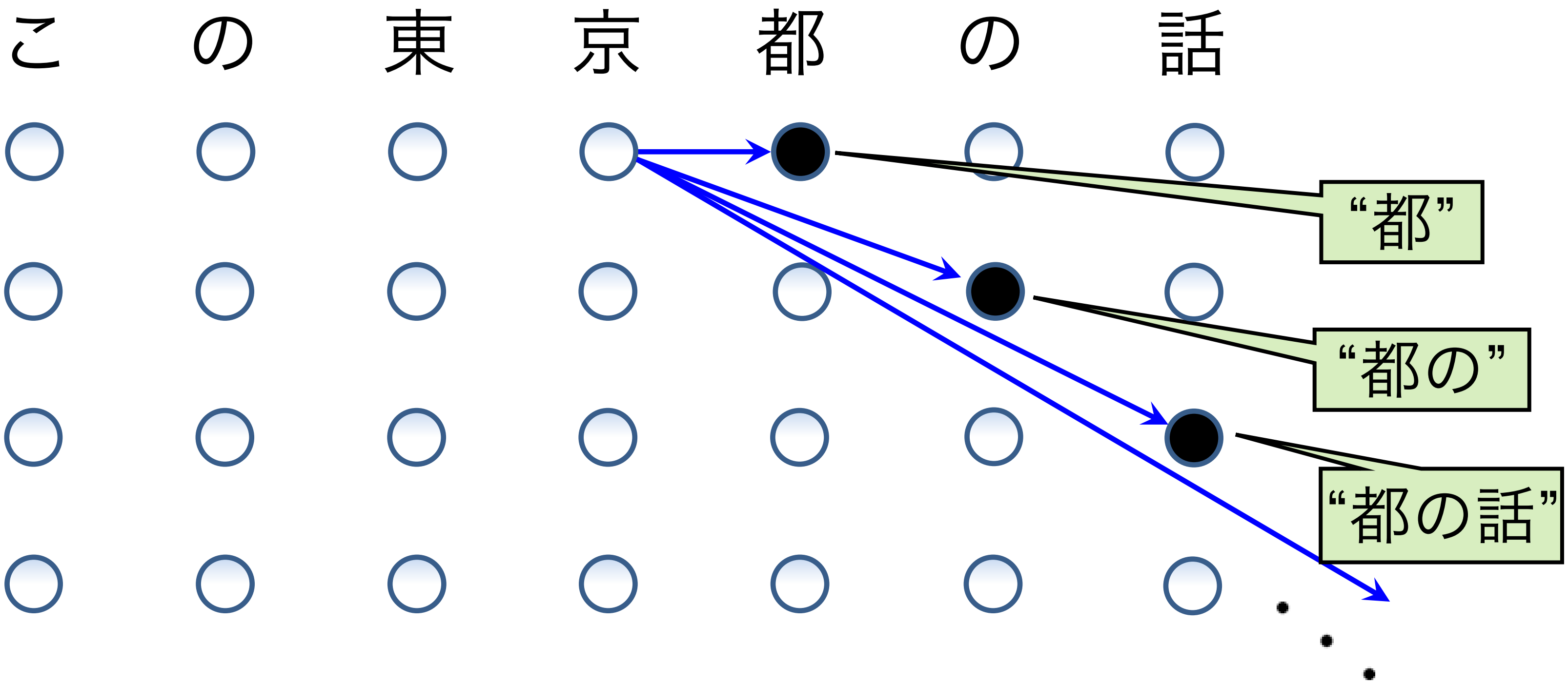
  - Four patterns: 0→0, 0→1, 1→0, 1→1
  - Pretend if the model is Markov HMM (not semi-Markov)
  - When sentence $\mathbf{x}$ is given, we can compute the corresponding potentials by intricately summing up NPYLM probabilities!

NPYLM→CRF (2)

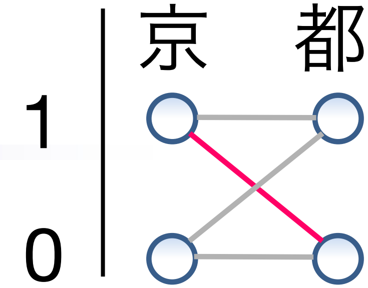


- Case 1→1 :

  1→1＝“京→都”, “京→都の”, “京→都の話”, …

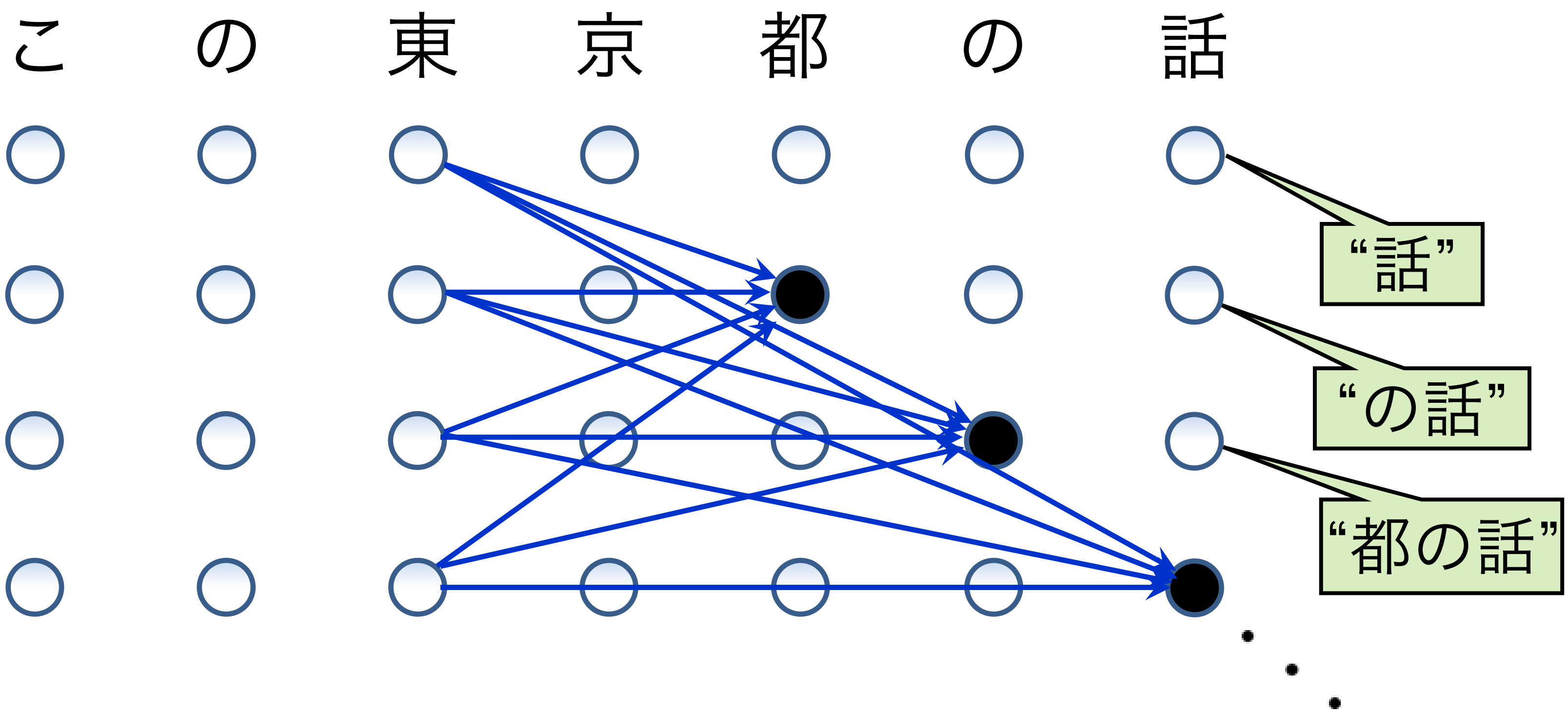


NPYLM→CRF (3)



- Case 1→0 :

1→0＝“東→京都”, “の東→京都”, “この東→京都”,
“東→京都の”, “の東→京都の”, “この東→京都の”,
“東→京都の話”, “の東→京都の話”, ・・・・



NPY→CRF: Code example

- C++ code for summing probabilities to compute "0→0" potential:

```
double
sentence::ccz (int t, HPYLM *lm)
{
	wstring w, h;
	int i, j, k, L = src.size();
	double z = 0;

	for (k = 0; k < MAX_LENGTH - 2; k++) {
		if (!(t + 1 + k < L)) break;
		for (j = 2 + k; j < index[t + 1 + k]; j++) {
			w = src.substr(t + 1 + k - j, j + 1);
			if (t + k - j < 0) {    /* (t + 1 + k - j) - 1 */
				h = EOS;
				z += lm->ngram_probability (w, h);
			} else {
				for (i = 0; i < index[t + k - j]; i++) {
					h = src.substr(t + k - j - i, i + 1);
					z += lm->ngram_probability (w, h);
				}
			}
		}
	}
	return z;
}
```

What are we doing? (1)

- *Mathematically*, this computation is a marginalization
  - By definition,

$$
\begin{aligned}
&p(c_t^{u-1} | c_s^{t-1}) \\
&= \gamma(s,t,u) \\
&= p(\underline{z_s = 1}, z_{s+1} = 0, \cdots, \underline{z_t = 1}, z_{t+1} = 0, \cdots, \underline{z_u = 1})
\end{aligned}
$$

  - Then we can marginalize:

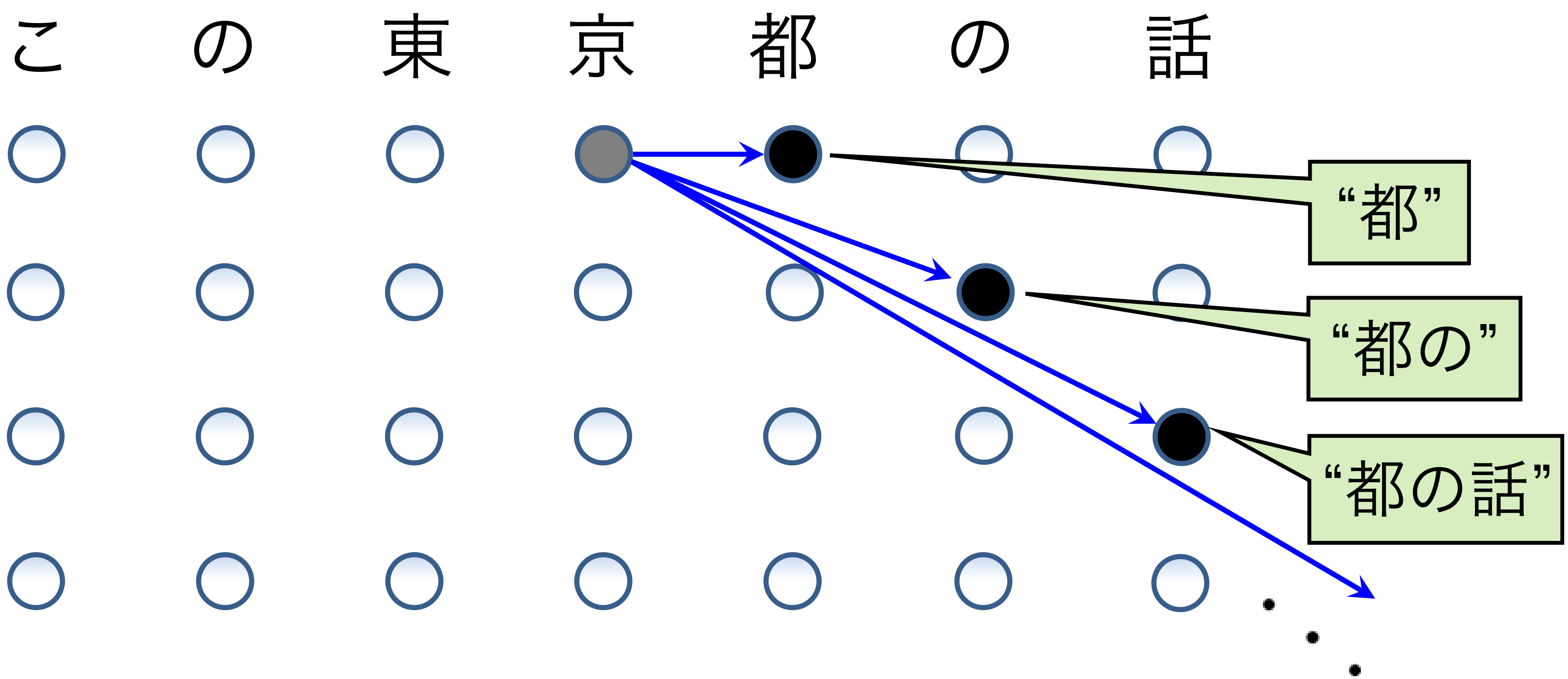
$$
p(z_t = 1, z_{t+1} = 1) = \sum_k p(\underline{z_t = 1}, \underline{z_{t+1} = 1}, \cdots, \underline{z_k = 1})
$$

$$
p(z_t = 1, z_{t+1} = 0) = \sum_l \sum_k p(\underline{z_t = 1}, z_{t+1} = 0, \cdots, \underline{z_k = 1}, \cdots, \underline{z_l = 1})
$$

$$
\begin{aligned}
p(z_t = 0, z_{t+1} = 0) = \sum_j \sum_l \sum_k p(&\underline{z_{t-1} = 1}, z_t = 0, z_{t+1} = 0, \cdots, \underline{z_l = 1}, \\
&\cdots, \underline{z_j = 1})
\end{aligned}
$$

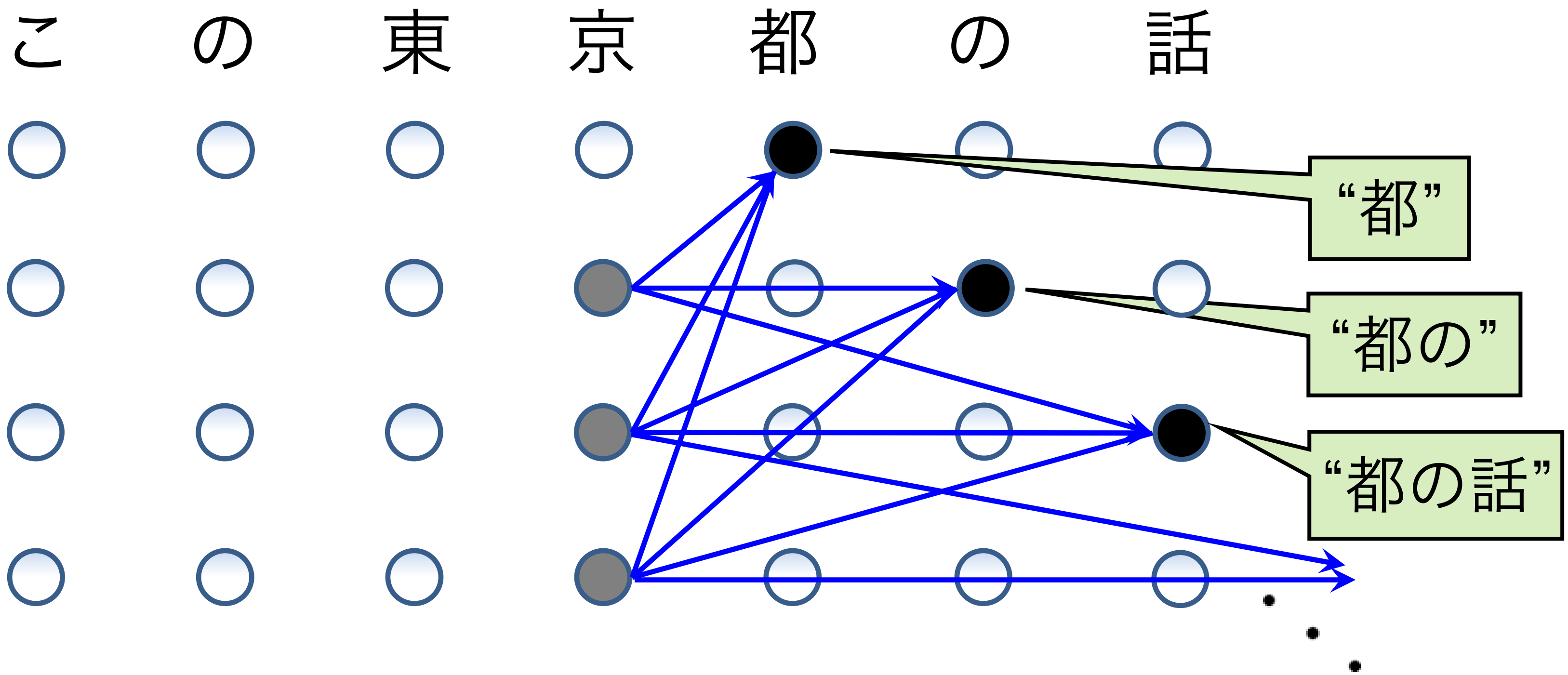
What are we doing? (2)

- *Graphically*, summing the collection of paths for the case of 1→1 :



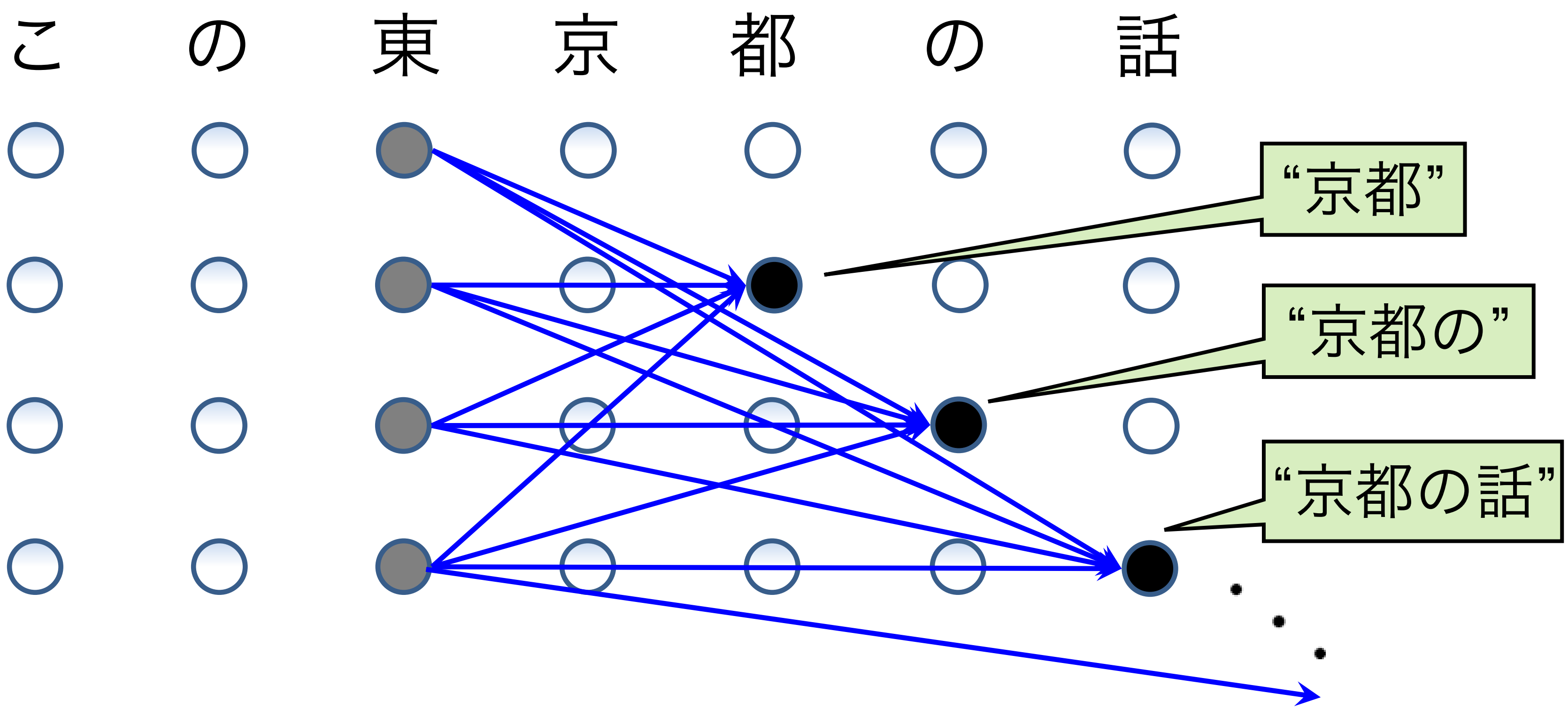
What are we doing? (2)

- *Graphically*, summing the collection of paths for the case of 0→1 :



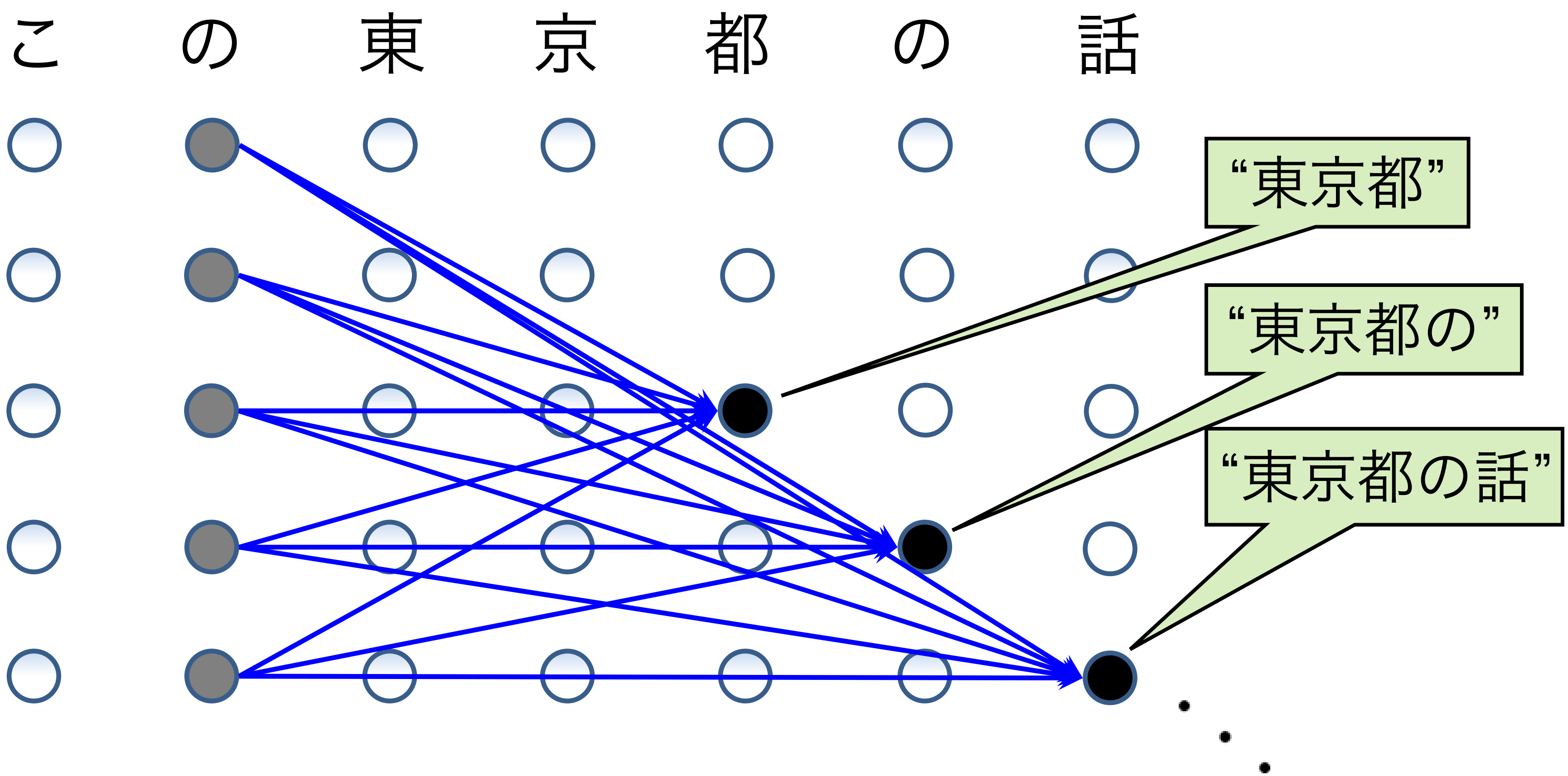
What are we doing? (2)

- *Graphically*, summing the collection of paths for the case of 1→0:



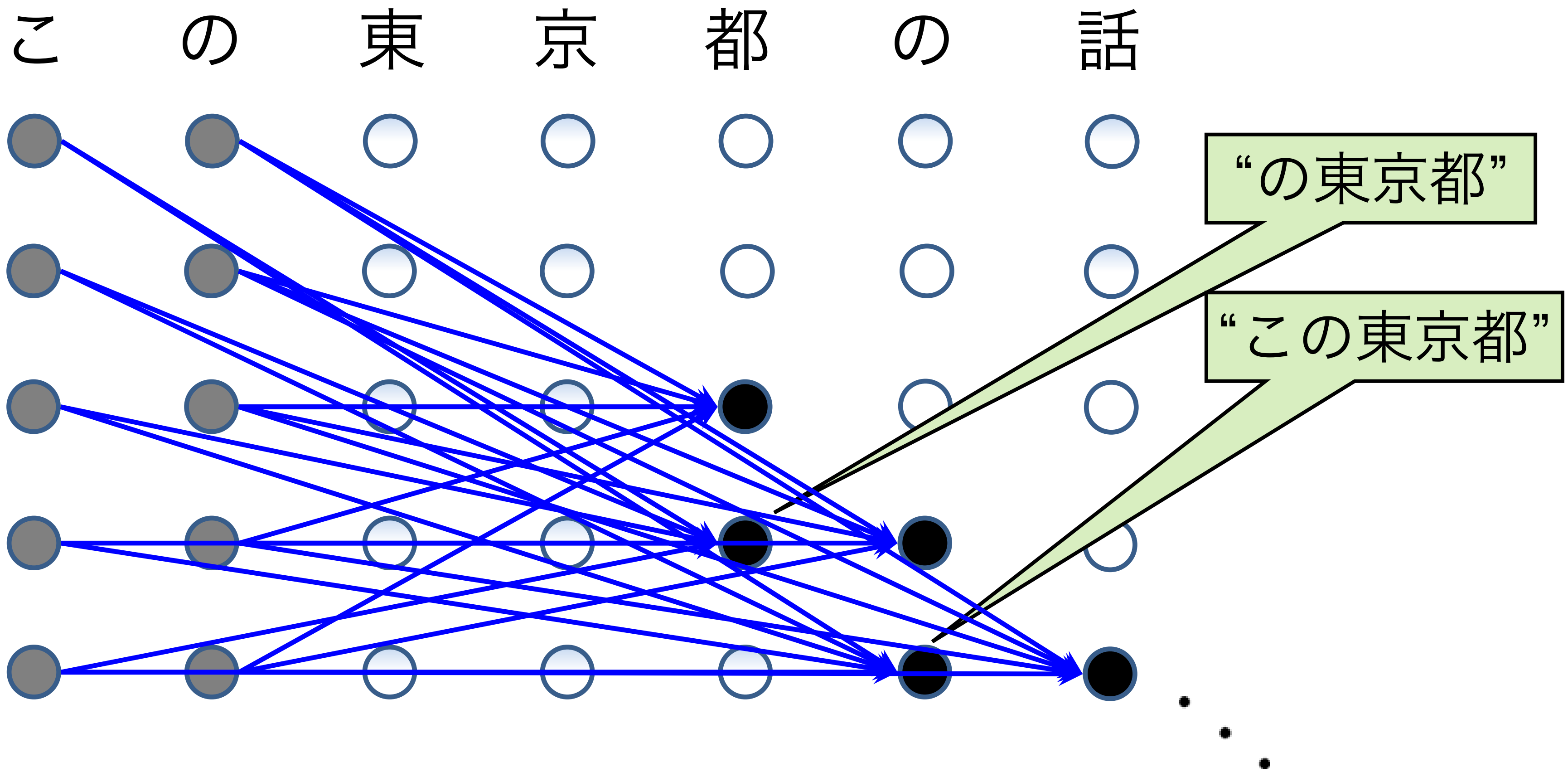
What are we doing? (2)

- *Graphically*, summing the collection of paths for the case of 0→0:



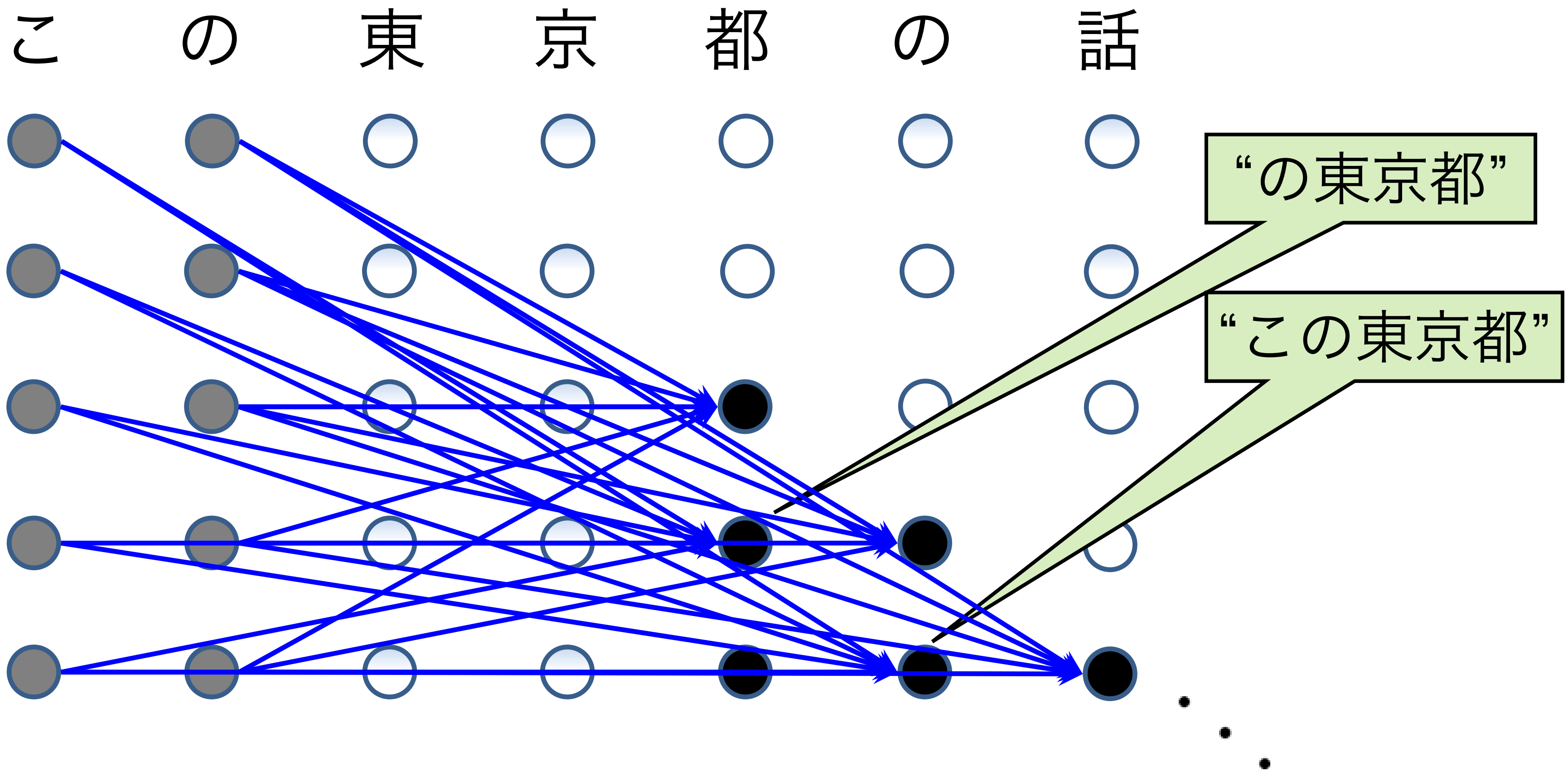
What are we doing? (2)

- *Graphically*, summing the collection of paths for the case of 0→0:



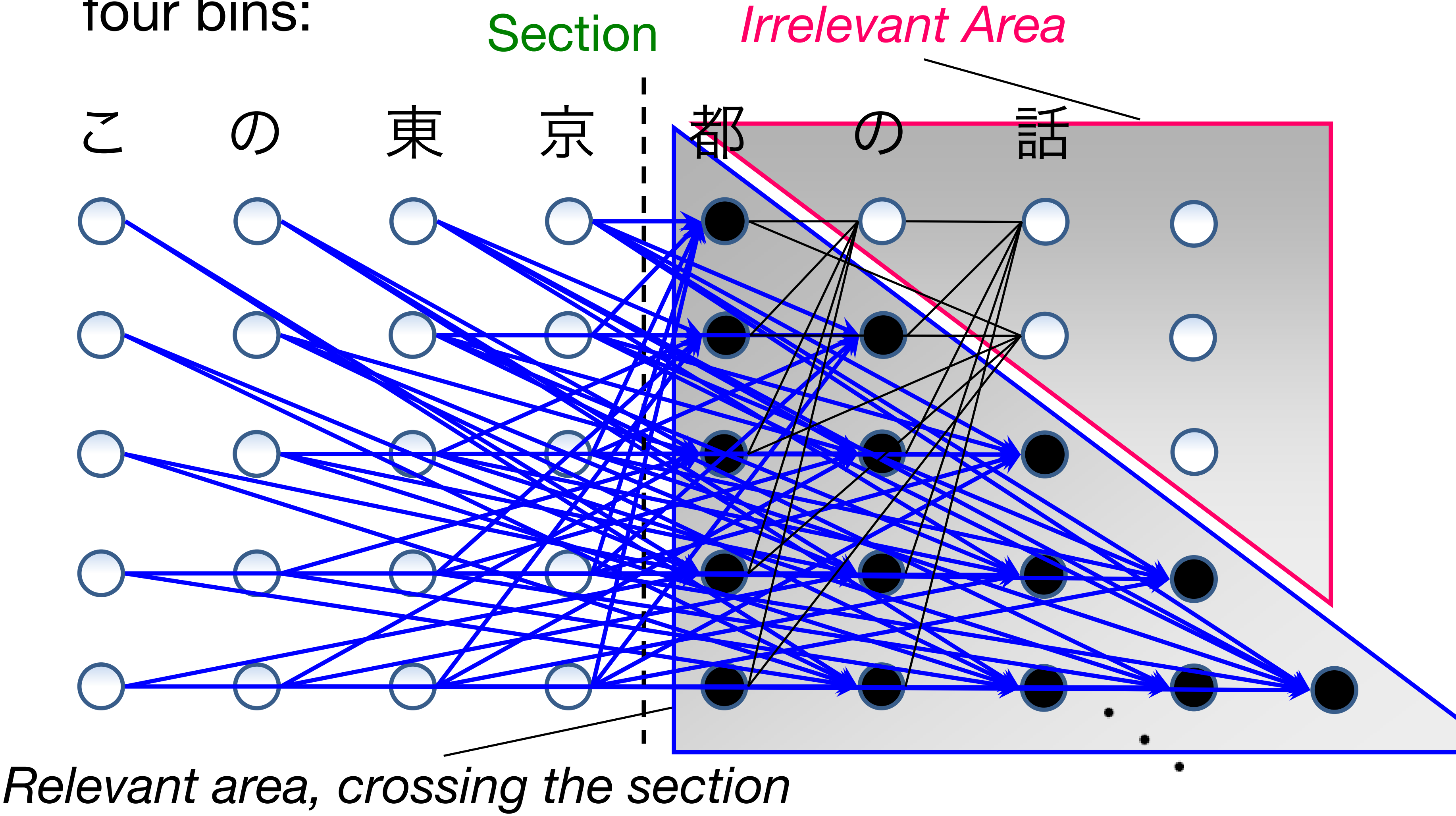
What are we doing? (2)

- *Graphically*, summing the collection of paths for the case of 0→0:



What are we doing? (2)

- *Graphically*, divide the paths at the section into four bins:



Experiments (still ongoing)

Tremendous!

- Sina Microblog (新浪微博)
  - Chinese equivalent of Twitter, 94,800,000 users
  - Full of non-standard Chinese sentences
- Japanese blog (Shokotan blog)
  - Famous for being full of jargons, new proper nouns, ..
- CSJ (Corpus of Spontaneous Japanese)
  - by National Institute of Japanese language and linguistics, next to ISM
- SIGHAN word segmentation Bakeoff 2005
  - Public dataset, intensively studied, newspaper

“Shokotan Blog” segmentations

中 三 の とき 後楽園 遊園地 に タイムレンジャー ショー を 見 に 行き まくってた ころ の こと そう だ お ね 、 セル画 は 修正 に 全て 塗り 直し と か ある だろう けど デジタル なら 一 発 で レイヤー 直せる から …
戸田 さん の 生 歌声 が 最高に 美しくて チャーム 状態 に なり まんた 。
そして ザ ・ ピーナッツ 役の 堀内 敬子 さん 瀬戸 カトリーヌ さん が 息 ピッタリ に 渡辺 プロ の そうそうたる 名曲 を 歌いあげ、 最高 の ハーモニー で とにかく すばらしかった です 。 生歌 で あの 美し さ … 。
四 つ と も 全部 この 川柳 ｗｗｗｗｗｗ お茶 ｗｗｗｗｗ イト カワユス ｗｗｗｗｗ イト カワユス ｗｗｗｗｗ (^ω ^) (^ω ^) (^ω ^)
深夜 まで お疲れさ マミタス (ﾟ ω ﾟ) ギャル 曽根 たん ！ 最近 よく 一緒に なる の と 楽屋 に 遊び に きて くれる ので いろいろ おしゃべり して タノシス ！ (^ω ^) 今日 も いろいろ 話し たお ね
イプサ の 化粧 水 が ケア 楽 チン だし 肌 に ギザ あう ！ これ ギガント 肌調子 よく なり まんた (^ω ^)

- Supervised : Kyoto corpus 37,400 setences
  Unsupervised: Shokotan blog 40,000 sentences

Shokotan blog: words

- Excerpt of random words with frequency >= 2

あるるるる	2	スワロフスキー	3	早いし	2
ますえ	2	わたる	11	信じろ	6
そびれちゃった	2	コマ送り	3	似てる	26
メリクリスマース	3	おおっお	7	居る	10
シクシク	3	にじむ	4	よる	85
チーム	45	簿	12	LaQua	7
ロック	11	ギギ	2	ただただ	7
キムタク	12	呼んで	29	ストロベリメロディ	21
うなぁ	2	席	31	スターーーートゥハッッッ	2
したろう	3	100	55	ひろがって	3
去った	4	グラビア	85	しろま	3
死兆星	4	田尻	3	カワユスピンク	2
スッキリ	6	より焼き	2	海馬	3
ドバァア	2	ヒャダルコ	3	除外	3
開催	47	永久	34	けえ	6
おく	17	ヤマト	2	なんとゆう	2

Sina microblog (新浪微博)

One of the most popular sites in China/HK/Taiwan

今天 一大早 就 被 电话 吵醒 了， 折磨 死 我 了， 昨天 太 晚 睡 了， 早上 被 这 电话 搞 的 晕 忽忽 ！
头疼， 发热 。。貌似 感冒 了， 晚上 睡觉 不能 裸 睡 了 。要 穿 睡衣 了 。
咿 ~ ？ 半个钟 前 发 的 围 脖 咋 不见 了 咧 ~~ 只是 感慨 了 一下 今天 的 归途 特顺 嘛 ~~~ (╯ ﹏ ╰) b
下雨 了， 不 知道 广州 那边 有没有 下雨， 明天 的 同学 聚会 我 去 不 了 了，[伤心] 大 哭
學 校 付 近 一 隻 很 可 愛 的 狗 狗， 做 了 點 特效 [心] [心] [心] 我 們 學 校 學生 超 愛牠 的 ！ ！ ！ [哈哈]
明儿 我 要 把 中山陵 搞定 ~~~~~ 玛丽 隔壁 的 ~~~ (¬ _ ¬)
好饿 啊 . . . . 走 ！ 妈妈 带 你 出去 吃饭 去 ~ . . . . . (((((ヾ (o = ^ · ェ ·) o ┏━┓ 喵 ~ o (= ∩ ω ∩ =) m
梦 。。。混乱 的 梦 。。清晰 的 梦 。。。。。

- Supervised : MSR 87000 sentences (Mandarin)
  Unsupervised : Sina API, 98700 sentences

Corpus of Spontaneous Japanese (CSJ)

- Experimental setting
  - Supervised: Kyoto corpus+CSJ 10% (21000 sents)
  - Unsupervised: CSJ 90% (190000 sents)
  - Training: CRF (only supervised) / NPYLM (sup+unsup)
- Results w.r.t. CSJ hand segmentations
  - P/R/F=0.928/0.893/0.91→0.931/0.899/0.915
  - Generally better, but ocasionally gets worse:

CRF> ええ だ から 何 て 言う ん でしょ う か 立っ てる と お腹に 力..
NPY> ええ だ から 何 て 言う ん でしょ う か 立っ てる と お腹 に 力..
CRF> 神奈川 寄り と か あっち の 方 行っちゃい ます よ ね 値段 ..
NPY> 神奈川 寄り と か あっち の 方 行っ ちゃい ます よ ね 値段 ..
CRF> 力強い 方向 性 ある タッチ
NPY> 力 強い 方向 性 ある タッチ

SIGHAN Bakeoff 2005

- Publicly available dataset of Chinese word segmentation, MSR portion
  - Mostly newspaper, *not so much suited for our task but standard*
- Data: MSR supervised (87000 sentences) + Chinese Gigaword (200000 sentences)
- Results:
  - Baseline: 97.4% F-measure with features of (Xu+ 2009)
  - MSR+Gigaword: 97.5%
  - MSR+Gigaword+dict: 97.5%

World best baseline! (closed)

97.3% on DPLVM

Difficult to beat.. Much more data to cover test data (5 days on Xeon W5590 to compute)

Conclusion

- Semi-supervised learning of CRF-NPYLM
  - Product models, each depends each
- Convert Semi-Markov ←→ Markov models
  - General observation, not only to NPYLM
- Good for blogs and Twitters
  - Maintaining accuracy on supervised data
  - Need: huge unsupervised data to improve on standard datasets
    - Parallelization or other implementation techniques

Future work

- Predicting POS (part-of-speech) simultaneously
  - “Class-based” n-gram models + Multiclass CRF
  - Class-based language models will also improve on unsupervised data
    - Ex. Predict “ギザ カワユス” → “ギザ スゴス”

Recognize “ギザ” as adjective

- However, in the current NLP…
  - No standard theory for building class-based n-grams.. only heuristic (like Google Mozc)
  - Dealing with “hierarchy” of part-of-speech
    - Ex. Noun→Proper noun→Name of a person
    - Theory of Markov model on tree-structured classes? (imagine hLDA)