条件付確率場とベイズ階層言語モデルの統合による半教師あり形態素解析

持橋大地*　鈴木潤　藤野昭典
NTTコミュニケーション科学基礎研究所
daichi@cslab.kecl.ntt.co.jp

言語処理学会2011
2011-3-10(Fri), 豊橋技術科学大学

最初に

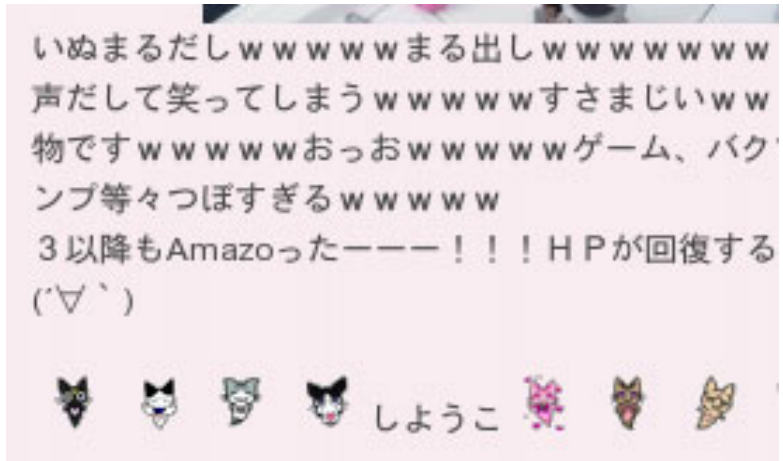
- 公式のPDFは完成度が充分でないため、
  ご興味のある方は下の完成版をお取り下さい:
  http://chasen.org/~daiti-m/paper/nlp2011semiseg.pdf
  - または, タイトルで検索で簡単に見つかります

“不自然”言語の解析

- 従来の新聞記事コーパスでは扱えない言語データが増えている (口語, 新語, 新表現, ...)

Twitter

Blog



– 全部人手で辞書登録...？(Brain damaged!)

話し言葉の解析

CSJ話し言葉コーパスの一部

何て言うんでしょうねその当時はそう思われていたっていうことを全部ドラマにしちゃうっていうところそういうところとかが面白くて凄く見てたんですけど最初の方は凄くまともで緋色の研究とか四人の署名とかそういうのから始まってたんですけどそれはもうとにかく原作に沿ってこういうこれこれホームズってこれっていうのが見たくて私はずっと見てます見てましたで凄くこれは何かＮＨＫが放送されてる時から...

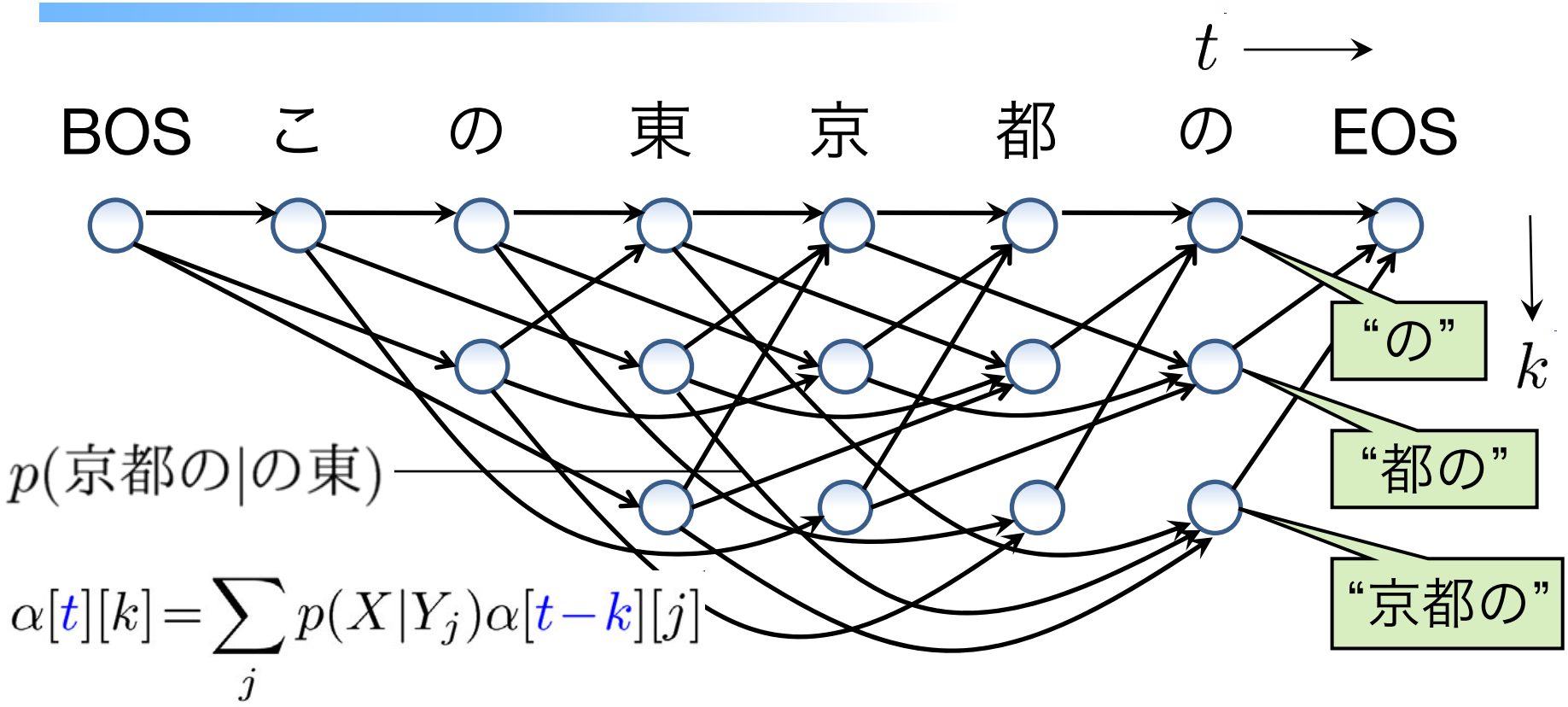
- 多数の口語表現 (“そいで”,“ってさぁ”,“ちゃって”...)
  - 無数のバリエーションが存在
- 自然な話し言葉の音声認識や音声科学のために、長期的にきわめて重要

教師なし形態素解析 (持橋+, ACL2009)

- 生の文字列だけから, 階層ベイズで「単語」を学習
  - モデル: NPYLM (Nested Pitman-Yor LM)

```
  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 。
```

NPYLM as a Semi-Markov model



- Semi-Markov HMM (Murphy 02, Ostendorf 96)の教師なし学習+MCMC法
- 状態遷移確率(nグラム)を超精密にスムージング

教師なし→半教師あり学習

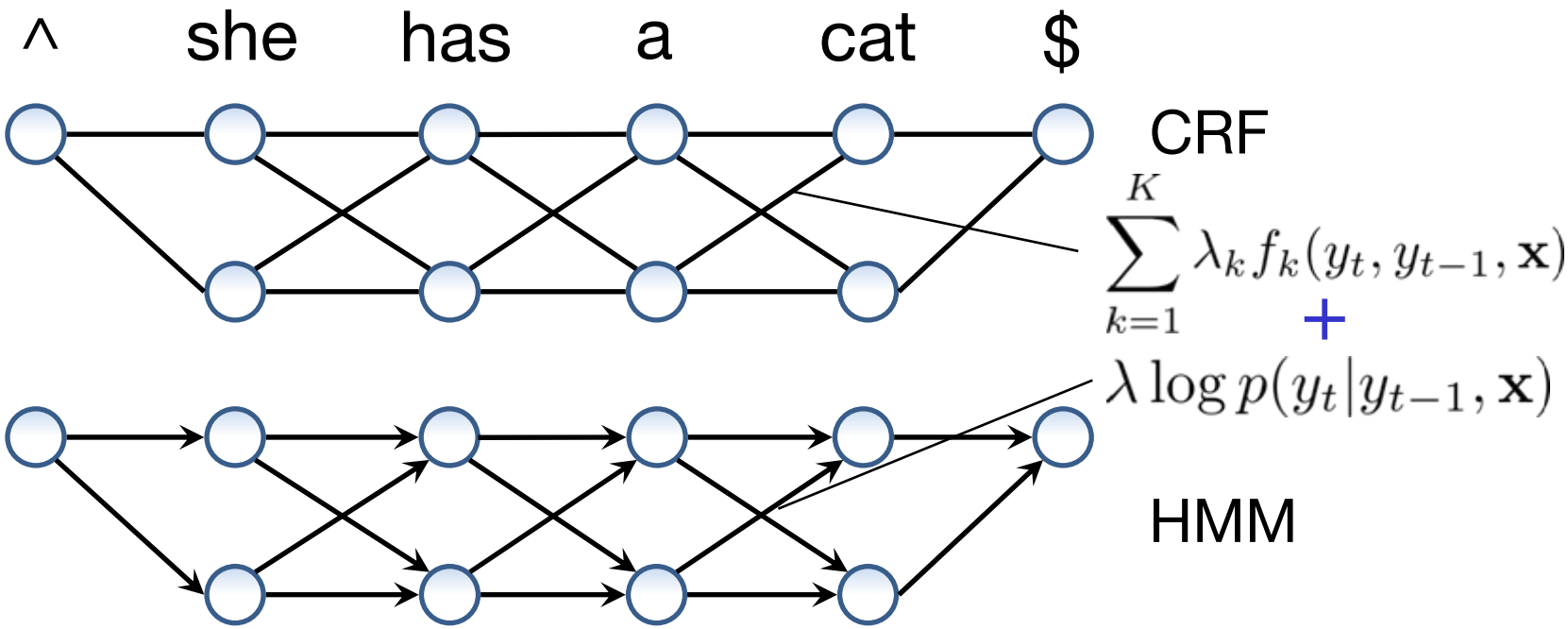
- 教師なし学習は低頻度語, 人間の基準に弱い
  - “阪急桂駅”, “首都グロズヌイ”
  - “歌 う”→“歌う”, “静か な”→“静かな”
- 生成と識別モデルの結合確率モデル: JESS-CM

識別モデル　生成モデル　重み λ も学習

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

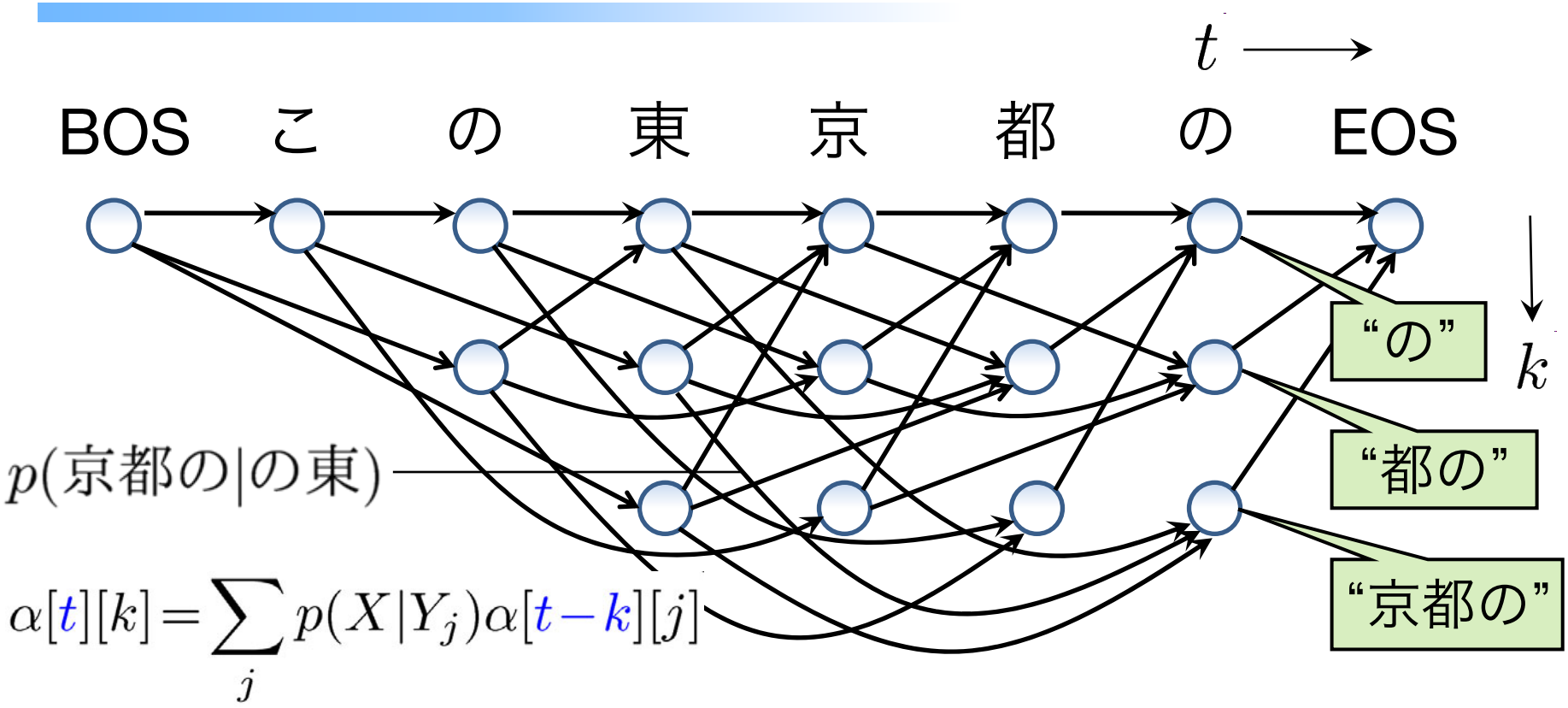
  - joint probability model embedding style semi-supervised conditional model (鈴木+ ACL08/09)
  - 現在世界最高性能の半教師あり学習
    - CRF/HMM, CRF/Naive Bayesなど

JESS-CM on CRF/HMM (鈴木+, ACL2008)



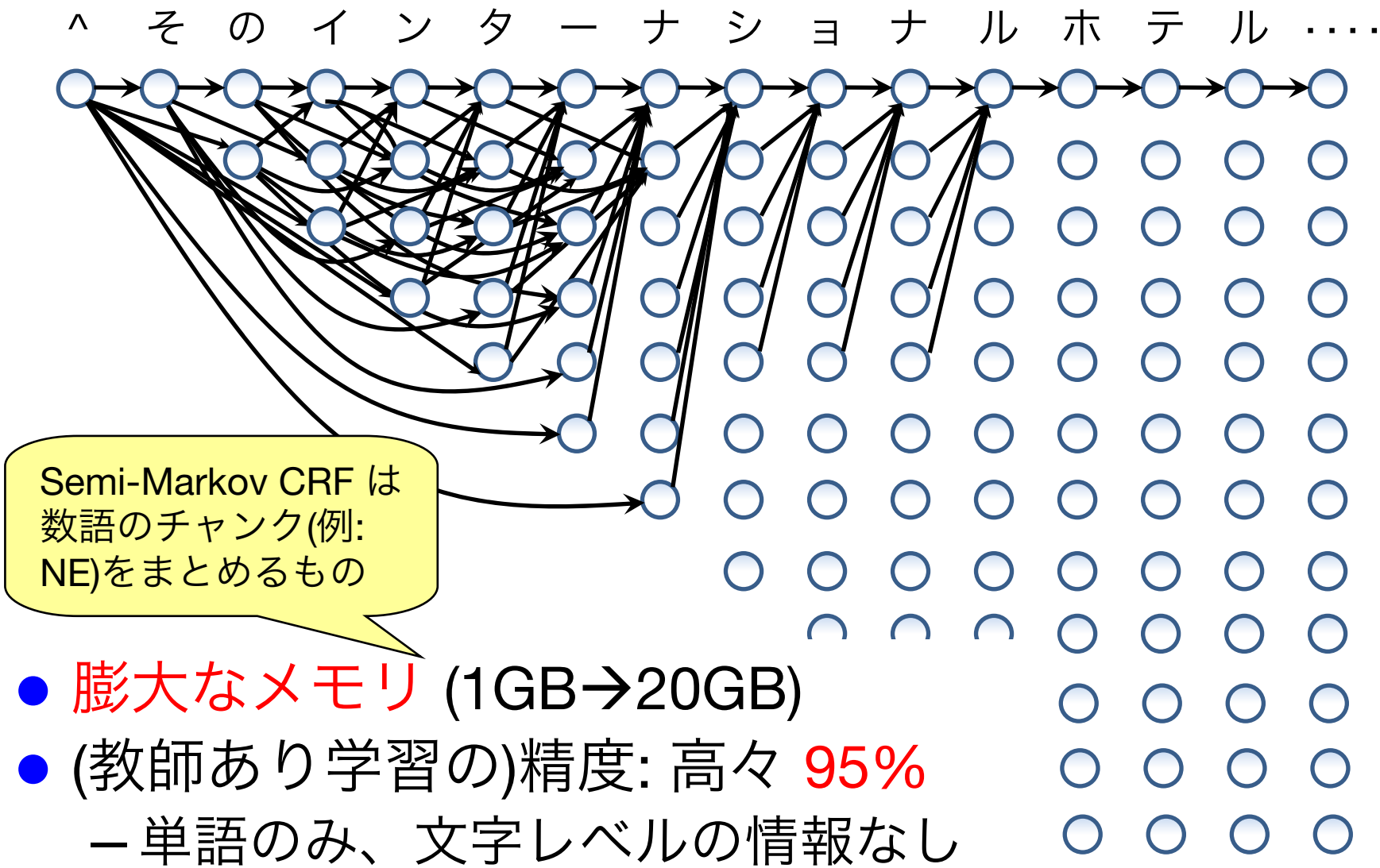
- 同じグラフィカルモデル上で, パスのコストを重みつきで足し合わせる
  →CRFとHMMを交互に学習(「教え合う」)
- NPYLMの場合は? (Semi-Markov)

NPYLM as a Semi-Markov model



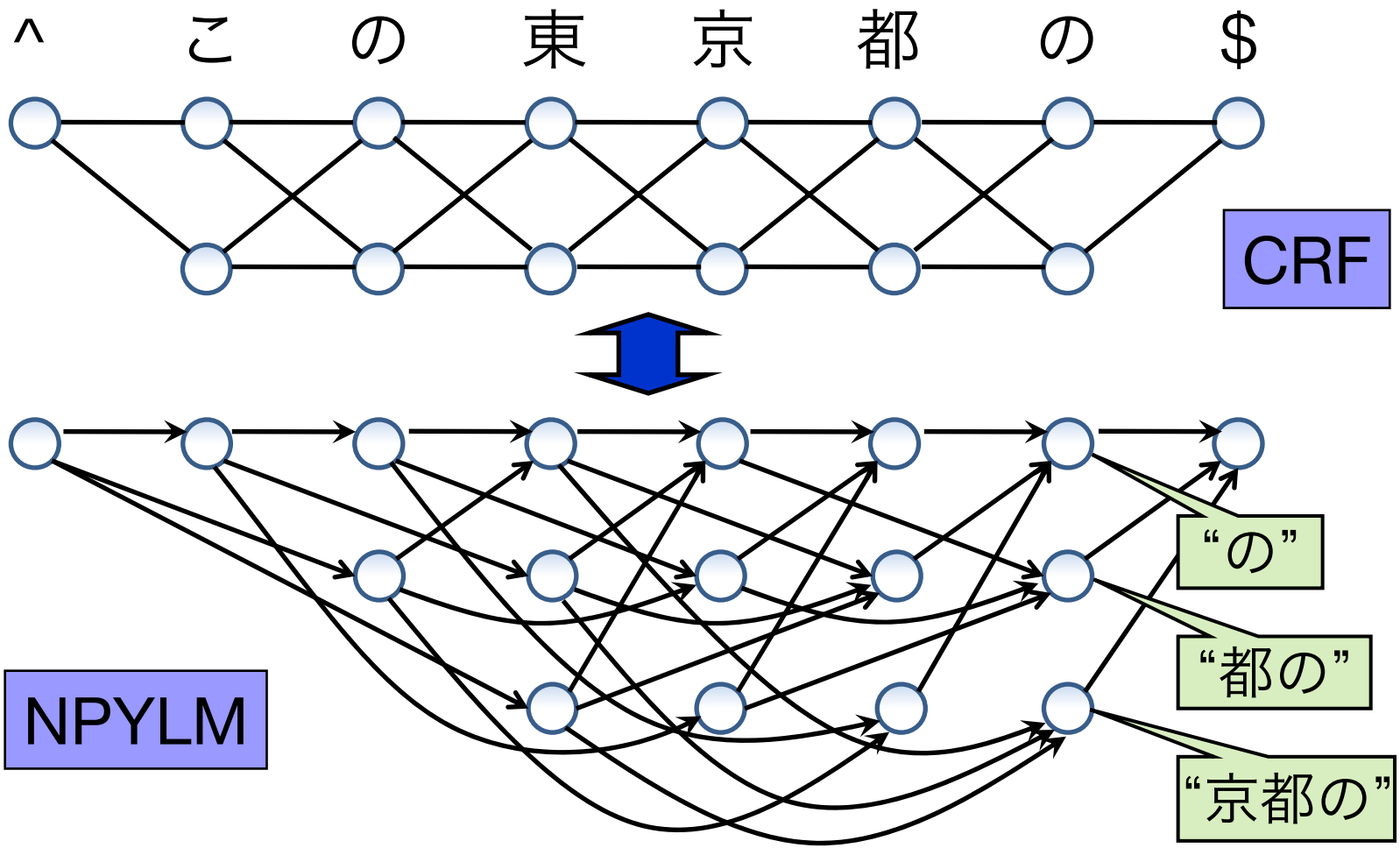
- Semi-Markov HMM (Murphy 02, Ostendorf 96)の教師なし学習+MCMC法
- 状態遷移確率(nグラム)を超精密にスムージング

Semi-Markov CRF (NIPS 2004)を適用?



- 膨大なメモリ (1GB→20GB)
- (教師あり学習の)精度: 高々 95%
  - 単語のみ、文字レベルの情報なし

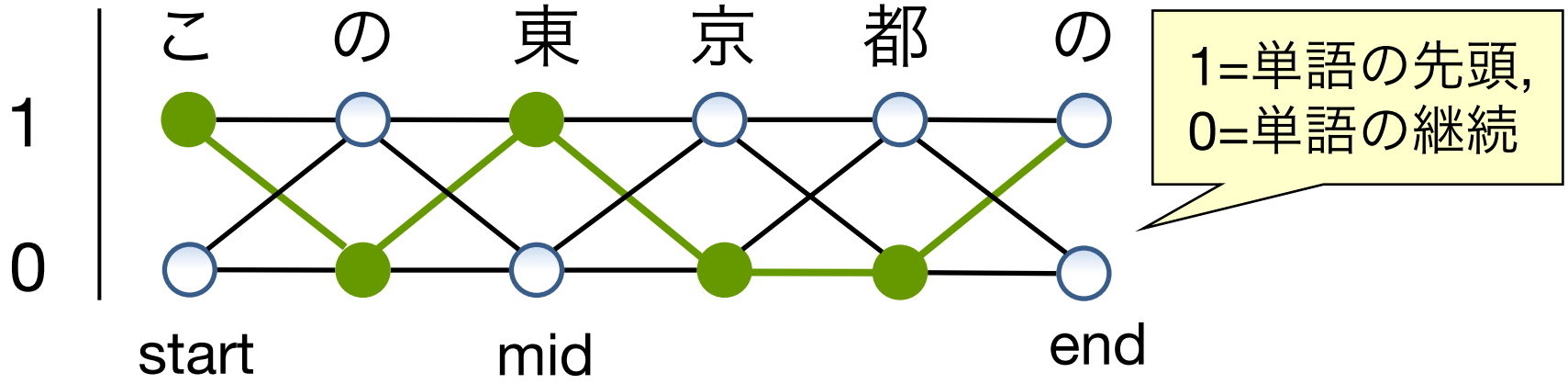
Markov CRF←→Semi-Markov LMの学習



- どうやって2つの違うモデルを統合するか?

CRF→NPYLM

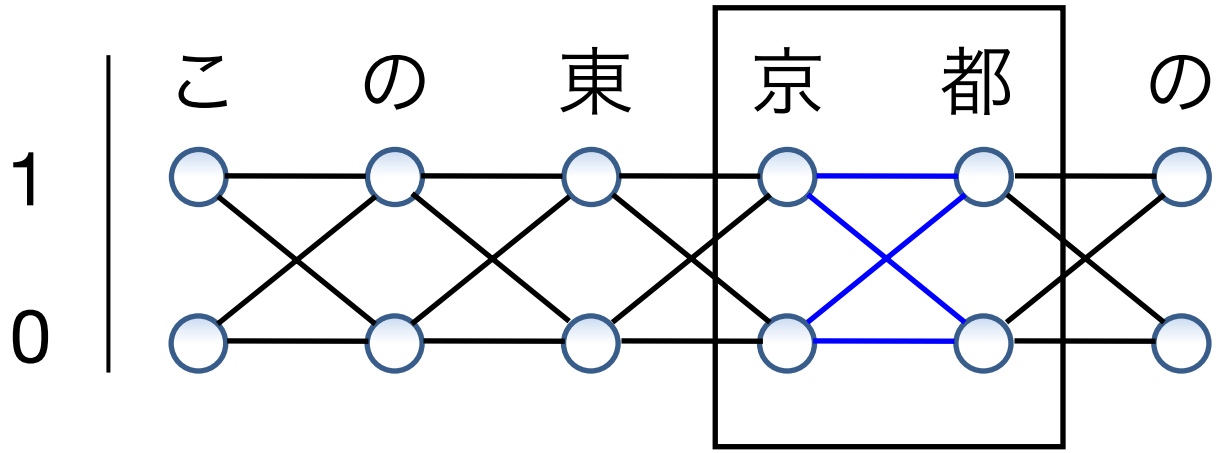
- Andrew+(EMNLP 2006)で既出、易しい
  - CRF→semi-Markov CRFの変換
  - *p*(“この”→“東京都”)



  - 上のパスに沿って素性の重みを足し合わせる
  - γ(start, mid, end)
    := γ(start, mid)+ γ(mid, end)

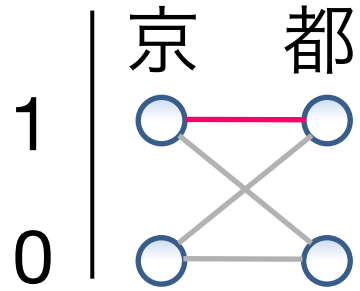
NPYLM→CRF (1)

- 難しい!!

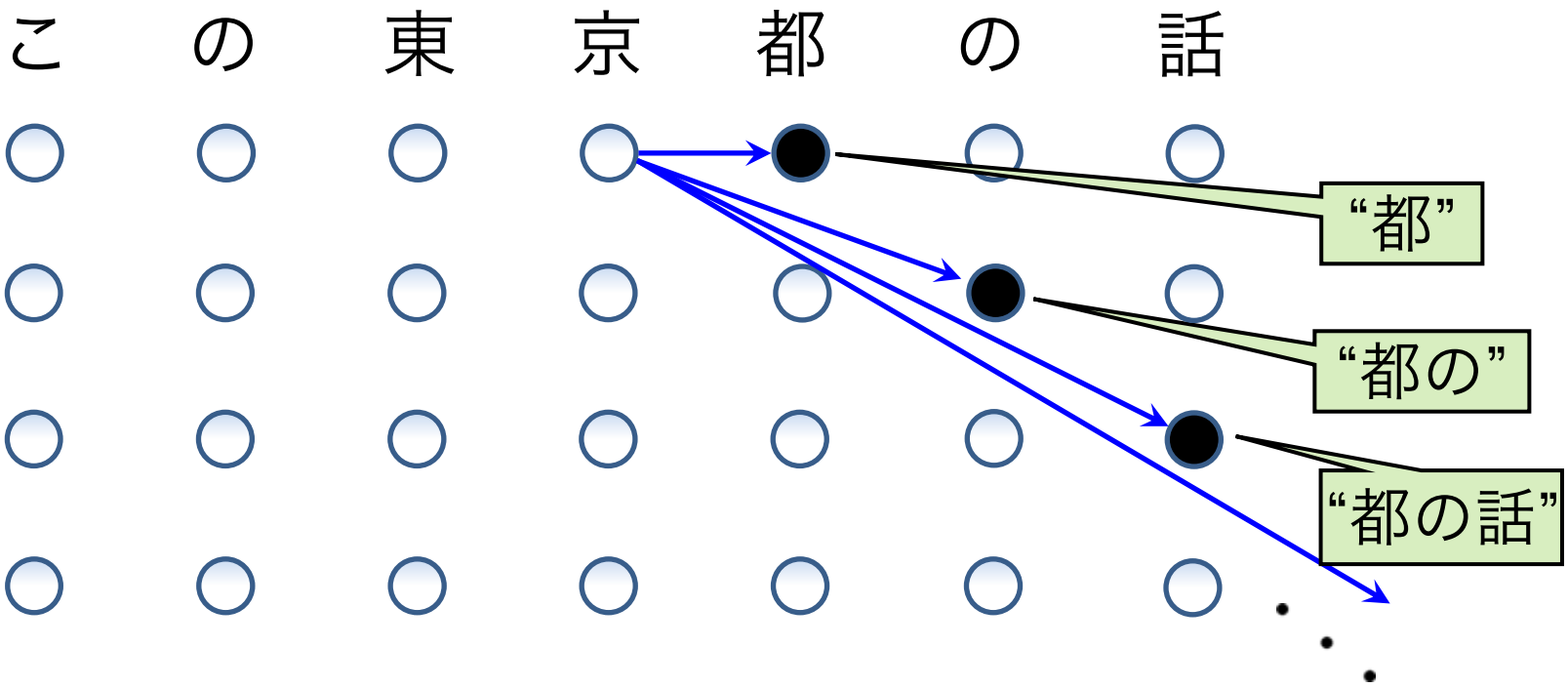


  - 4通りのパス: 0→0, 0→1, 1→0, 1→1
  - モデルがもしMarkov(=HMM)だった時の重みは, 文 $\mathbf{x}$が与えられれば、Semi-Markovの確率を複雑に足し合わせることで計算可能！

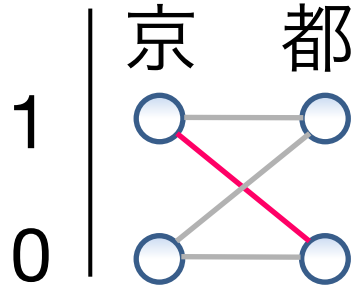
NPYLM→CRF (2)



- Case 1→1 :
  1→1＝“京→都”, “京→都の”, “京→都の話”, …

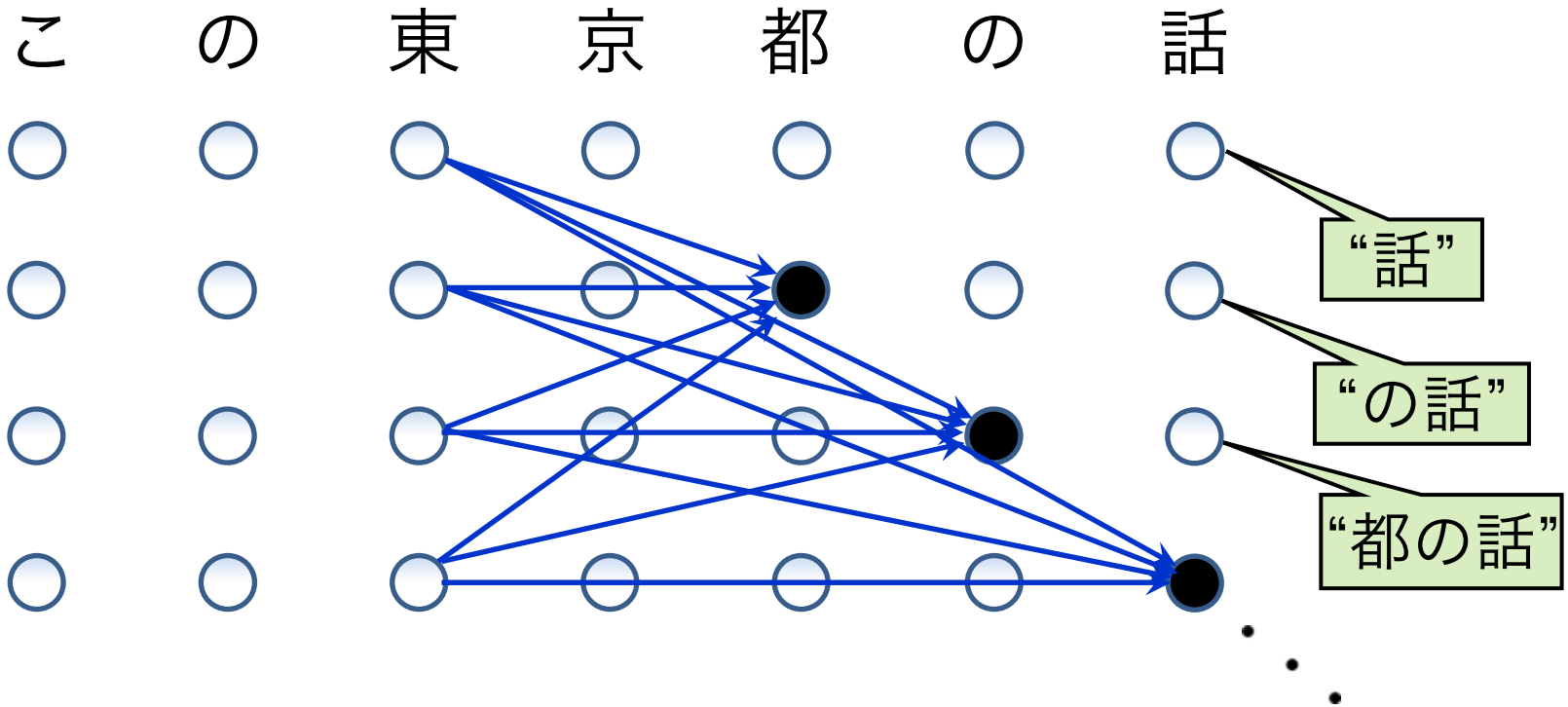


NPYLM→CRF (3)



- 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

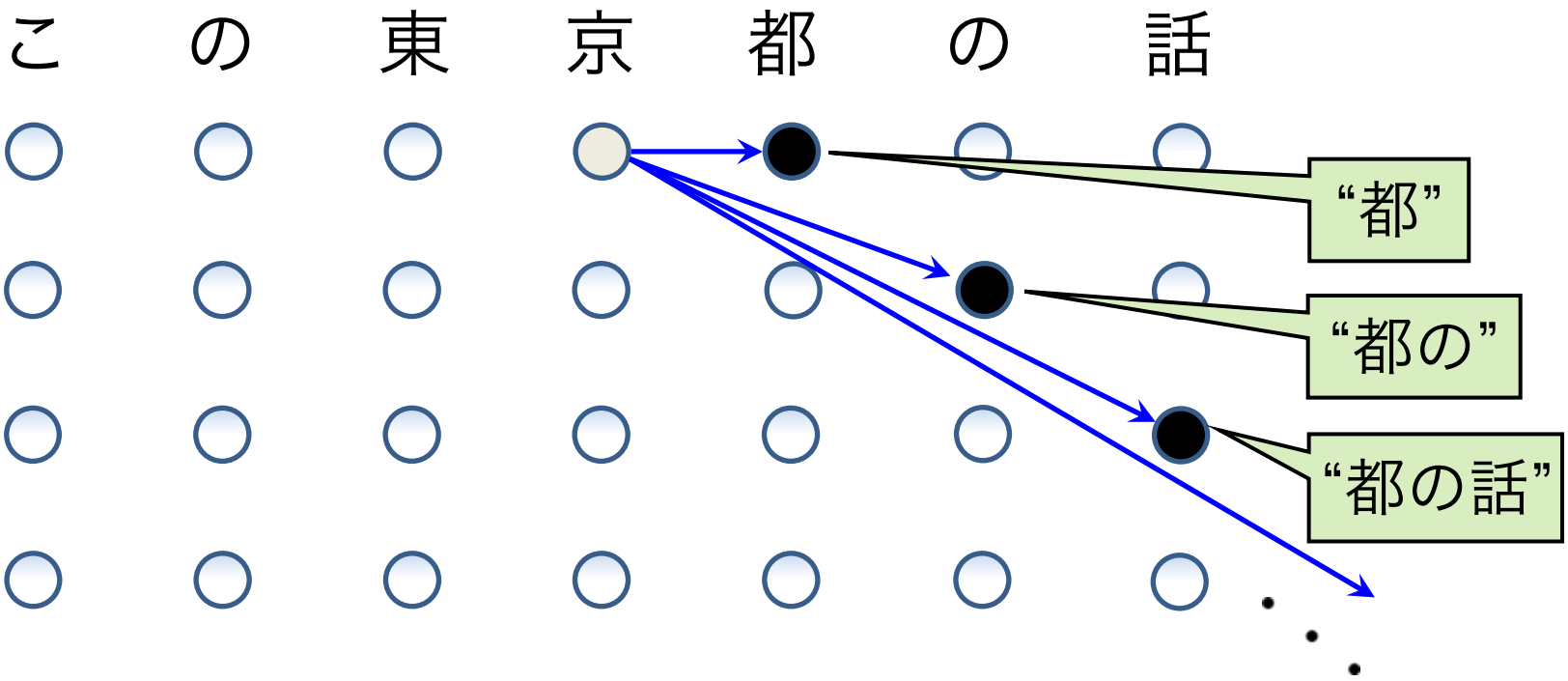
- “0→0”のポテンシャルを計算するC++コード

```
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;
}
```

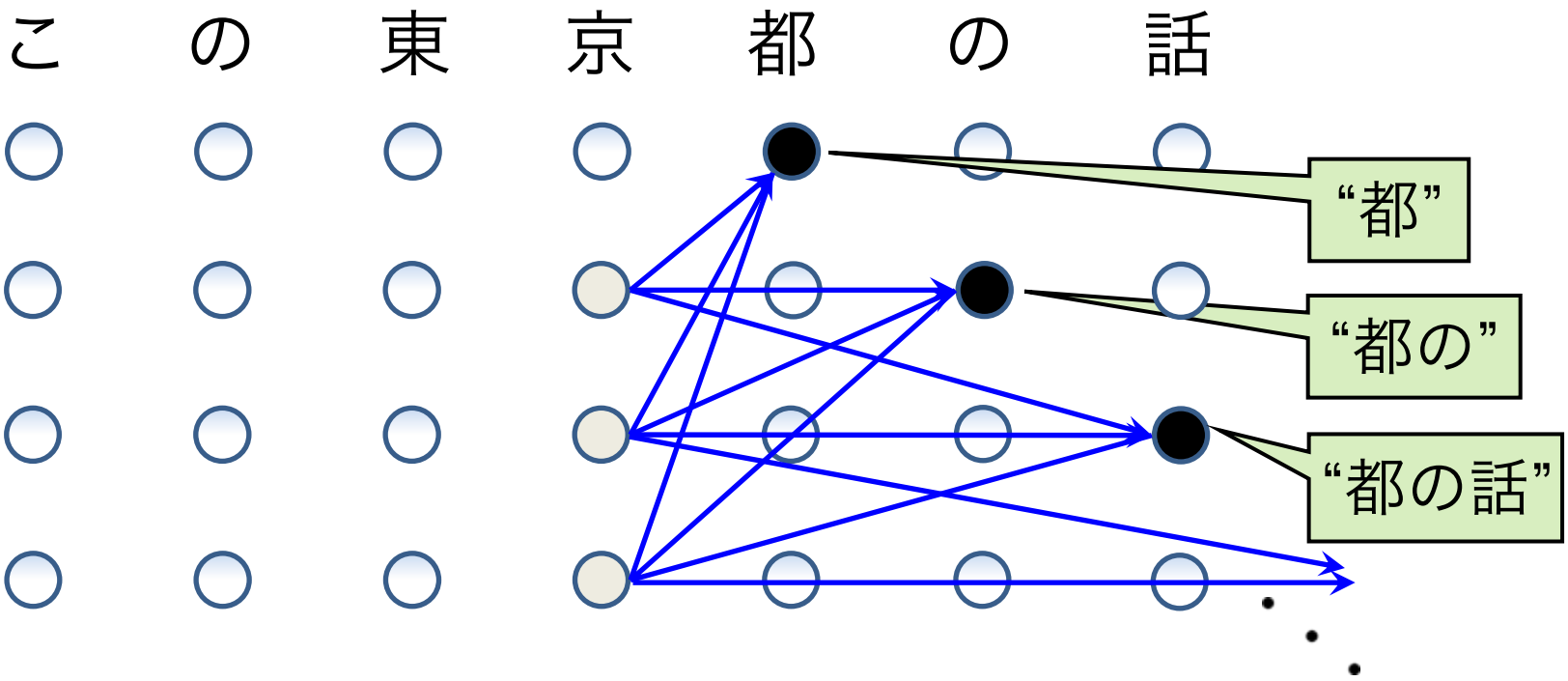
What are we doing? (1)

- グラフィカルモデル上で、1→1に対応する確率の和をとる：



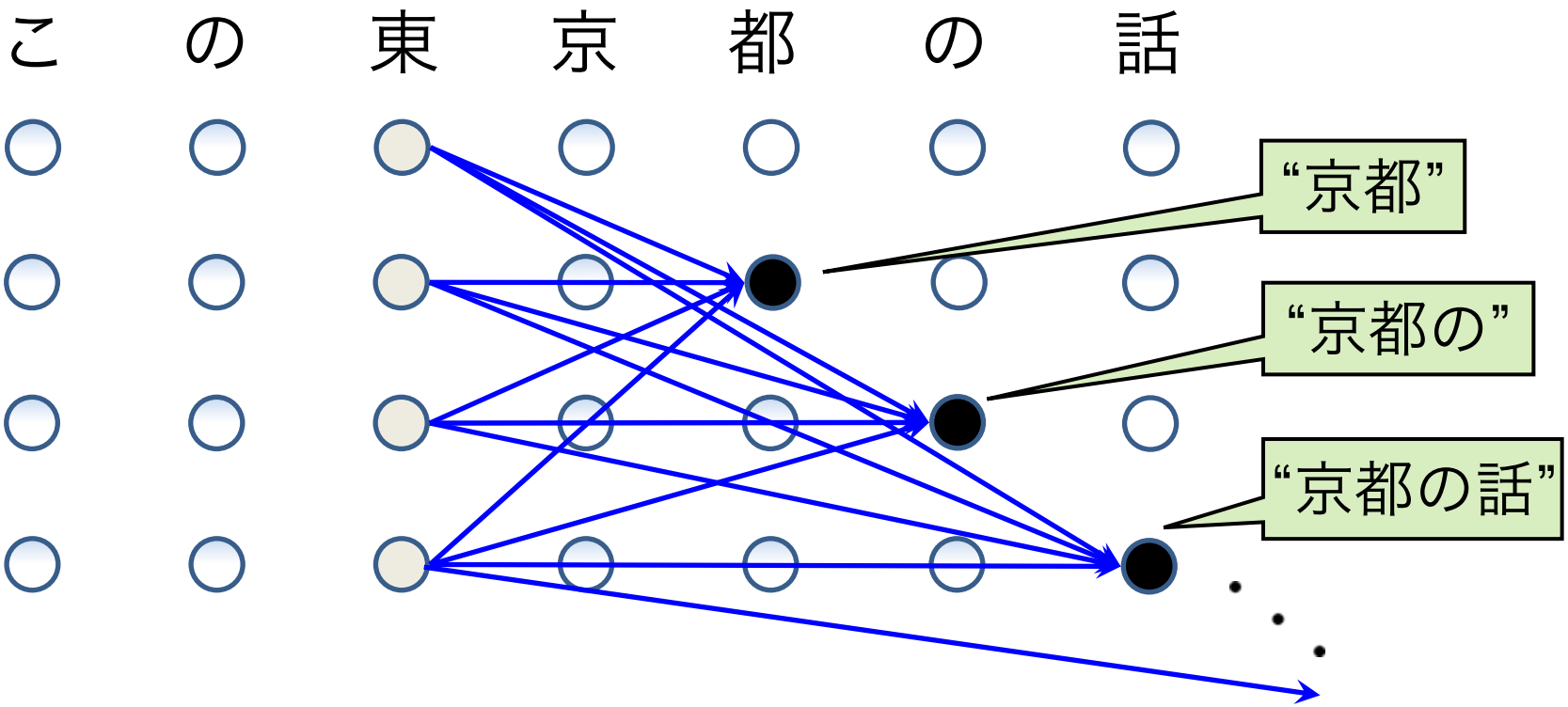
What are we doing? (1)

- グラフィカルモデル上で、0→1に対応する確率の和をとる：



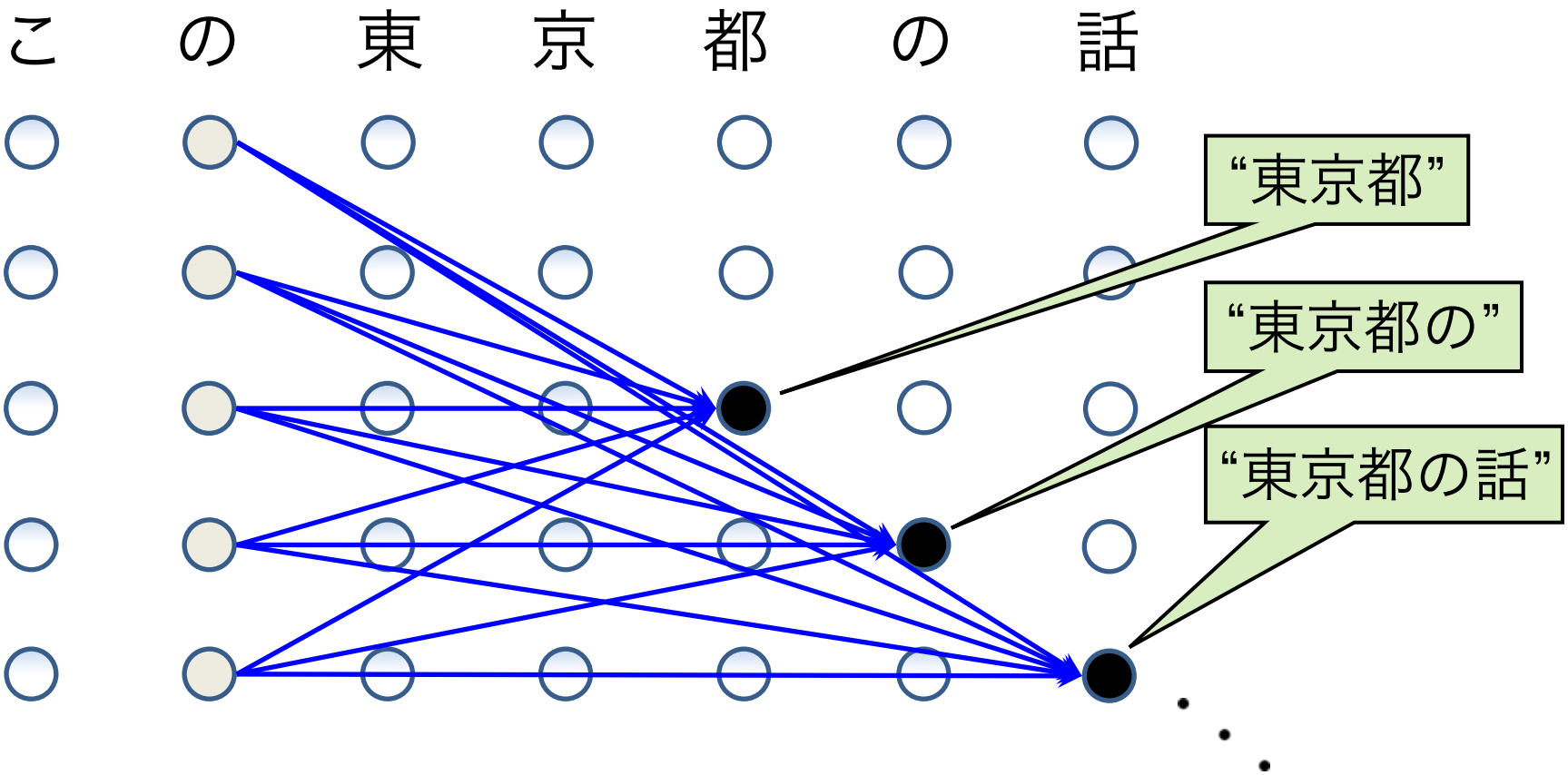
What are we doing? (1)

- グラフィカルモデル上で、1→0に対応する確率の和をとる：



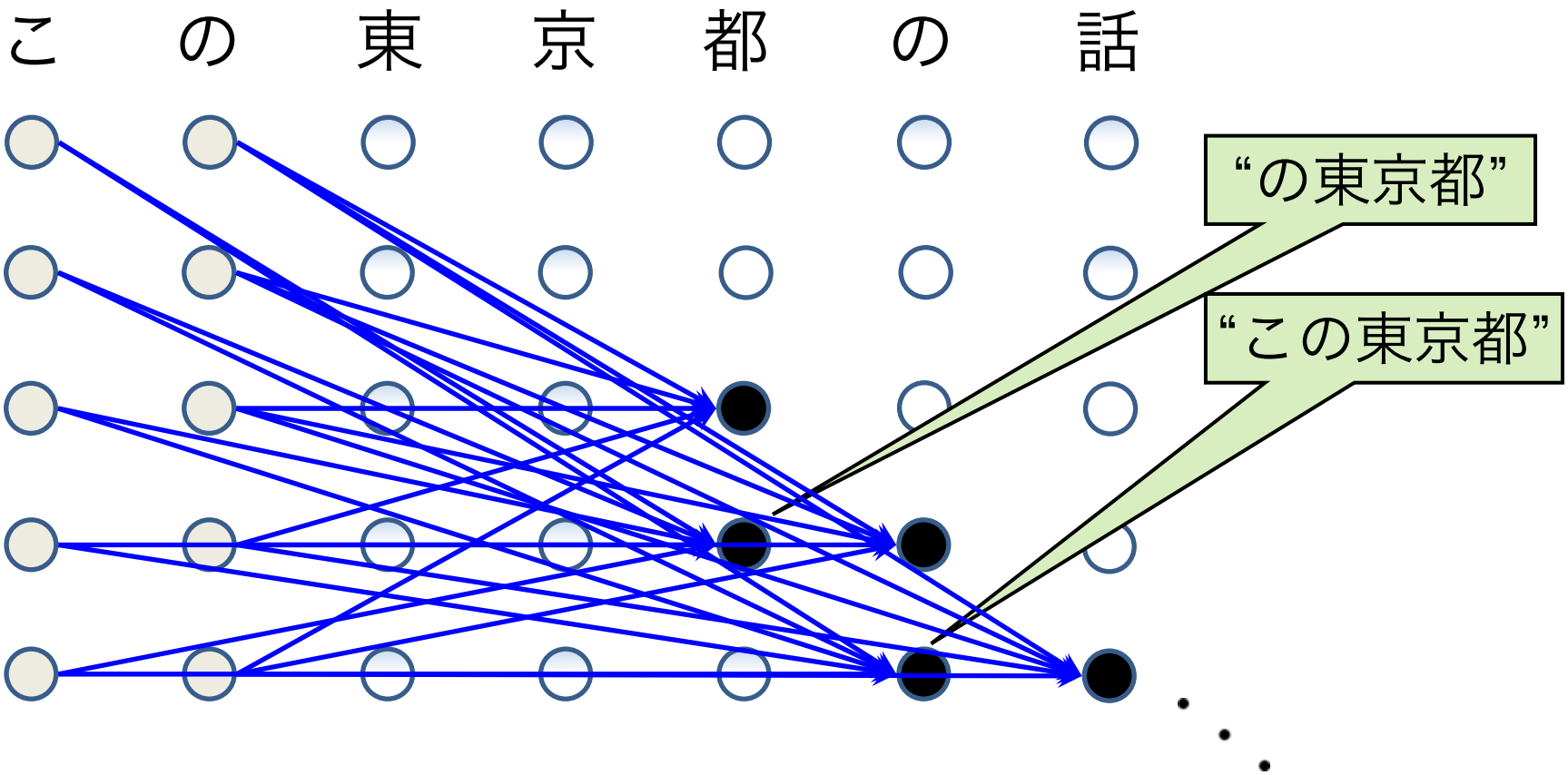
What are we doing? (1)

- グラフィカルモデル上で、0→0に対応する確率の和をとる：



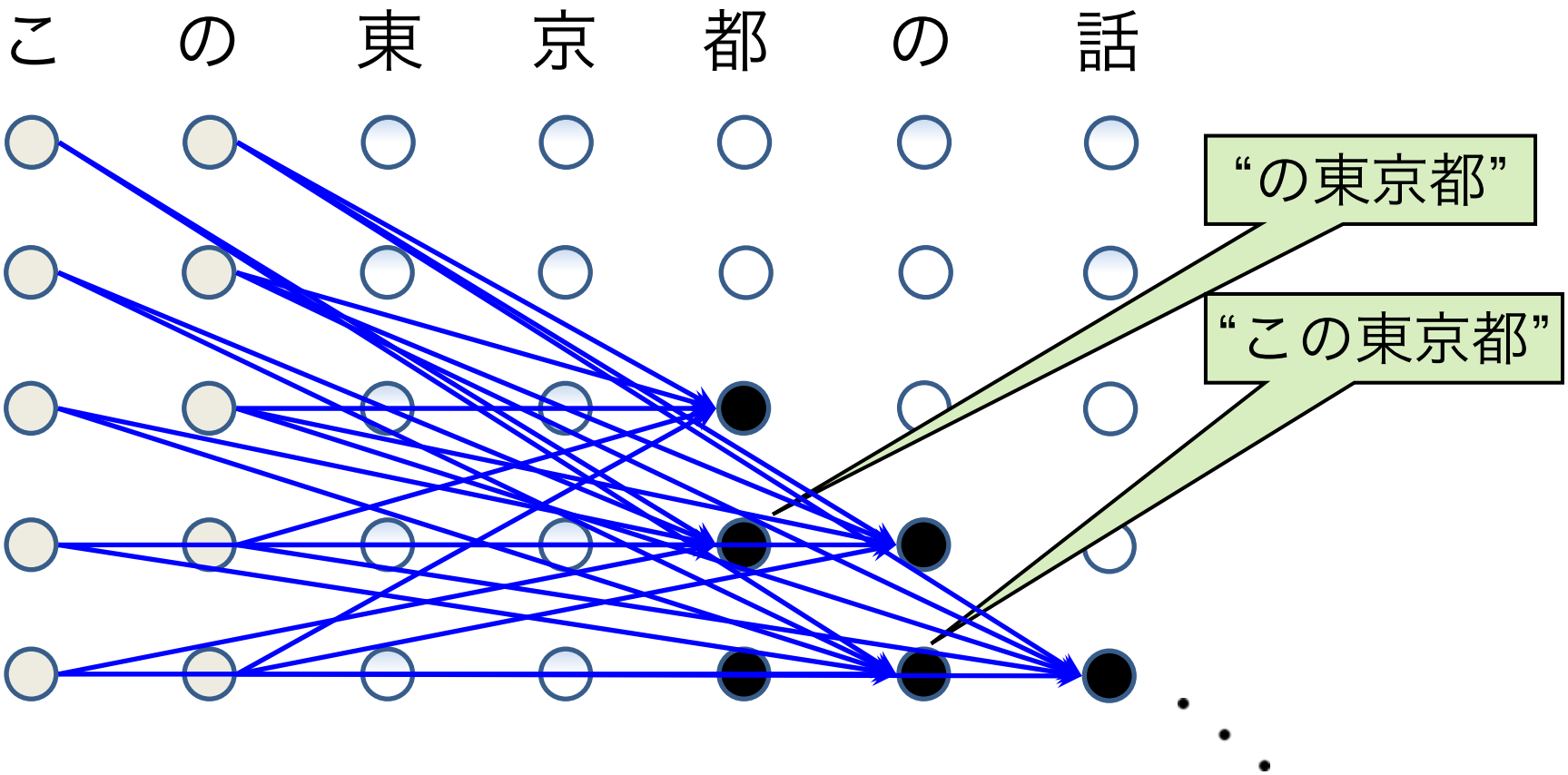
What are we doing? (1)

- グラフィカルモデル上で、0→0に対応する確率の和をとる：



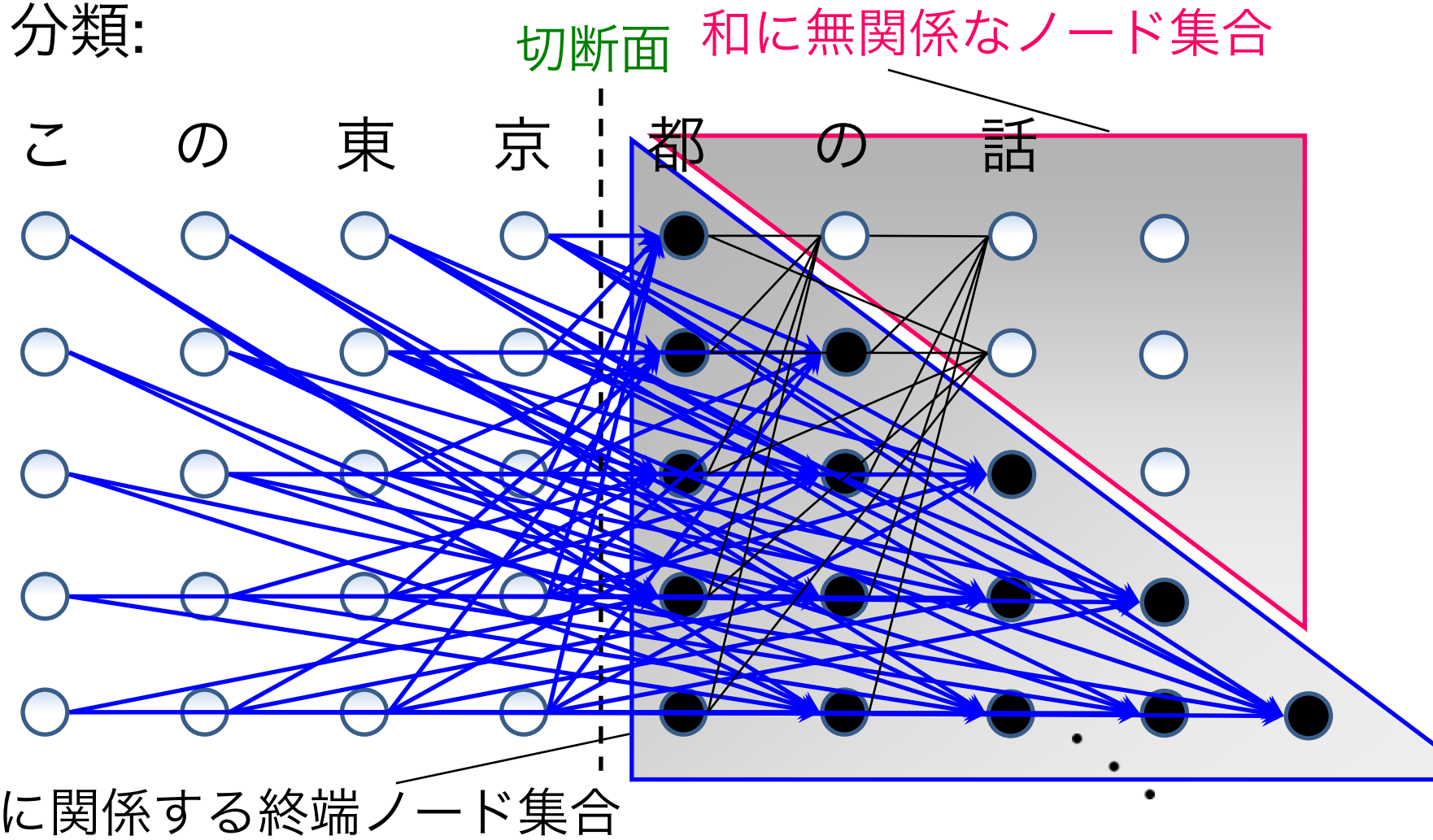
What are we doing? (1)

- グラフィカルモデル上で、0→0に対応する確率の和をとる：



What are we doing? (1)

- DAG上で, 切断面を横断するパスを4種類に分類:

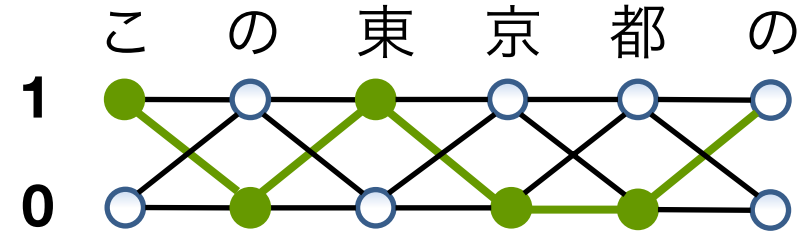


What are we doing? (2)

$$p(x) = \sum_Y p(x, y)$$

- 数学的には, この計算は確率の周辺化と等価
  - 定義に従って,

$$p(c_t^{u-1} | c_s^{t-1}) \propto \gamma(s, t, u)$$



$$= p(z_s = 1, z_{s+1} = 0, \cdots, z_t = 1, z_{t+1} = 0, \cdots, z_u = 1)$$

  - 周辺化により, 求める同時確率が得られる

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

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

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

Experiments (still ongoing)

- 新浪微博 (Sina Microblog)
  - 中国語圏のTwitter, 95000000人のユーザ
- 「しょこたんブログ」
  - “しょこたん語”で有名な、崩れた日本語blog
- CSJ日本語話し言葉コーパス (省略)
- SIGHAN Bakeoff 2005
  - 公開データセット、新聞記事

Tremendous!

「しょこたんブログ」の分割

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 ^)

- 教師あり: 京大コーパス 37,400文
  教師なし: しょこたんブログ 40,000文

しょこたんブログの「単語」

- 頻度2以上の単語をランダムに抽出

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	スターーーーートゥハッツッ	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

新浪微博 (Sina microblog)

中国／香港／台湾での
超有名サイト(Twitter)

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 ~~~ (ノ ﹏ ╰) 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 。。。。。

- 教師あり: MSR 87000 文 (普通話)
  教師なし: Sina API, 98700 文

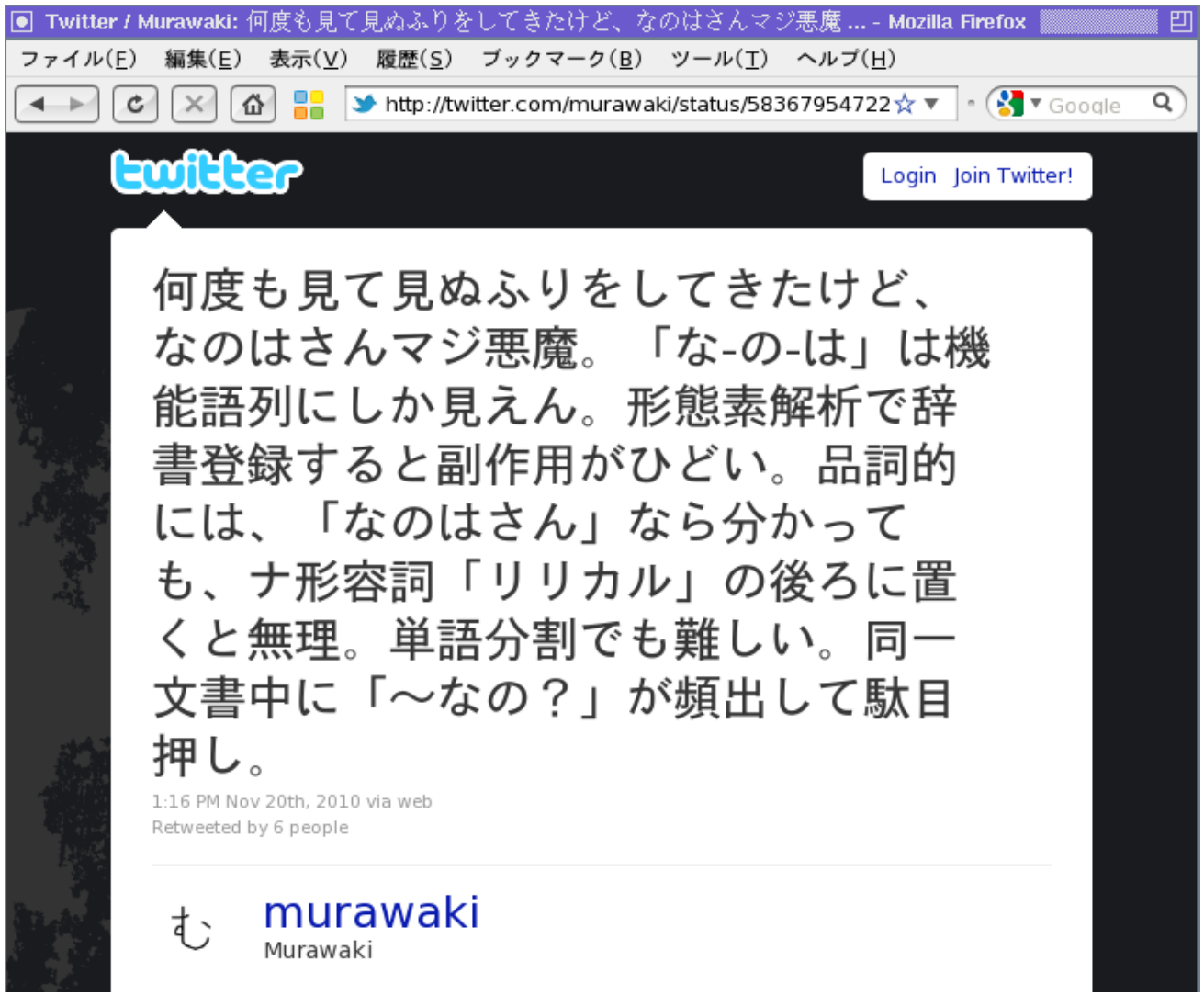
SIGHAN Bakeoff 2005

- 中国語単語分割の標準データセット
  - 新聞記事のため, 半教師あり学習の目的とはやや異なるが...
- データ: MSR Asia 87k+ Chinese Gigaword 200k

Model	CRF	NPYCRF	+辞書
Token F値	97.4	**97.5**	**97.5**
OOV 再現率	83.5	**84.1**	82.1
IV 再現率	98.5	**98.6**	**98.8**

  - 改善しているが, 大きな改善にはもっとデータ量が必要 (現在2004年の約半年分)
  - ベースラインの97.4%は教師ありでは世界最高

解析が非常に困難な例 [注: 以下お祭りです]



実際にやってみました

- データ: 2ch 魔法少女なのは総合スレ874-883
  - 10スレッド分(10000レス), 26474文
  - NPYCRF/K=12, 教師データ: 京大コーパス
- 結果: 難しい！

2004 年秋 に 放送 さ れた 「 魔法 少女 リリカルな の は 」 と 、
2005 年秋 に 放送 さ れた 「 魔法 少女 リリカルな の は A ‘s 」 、
2007 年春 に 放送 さ れた 「 魔法 少女 リリカルな の は StrikerS 」
に ついて 語り ましょう 。
な の は は 基本 的に 個人 戦 だ から ほとんど 描写 さ れ ない が
な の は さん は 仕事 が 恋人な ワーカーホリック です

- 不可能なのか? → No！

教師なし学習の結果

- 京大コーパス生文+なのはスレ26474文の
  教師なし学習

  2004 年秋に放送 された 「 魔法少女リリカルなのは 」 と 、
  2005 年秋に放送 された 「 魔法少女リリカルなのは A ‘ s 」 、
  2007 年春に放送 された 「 魔法少女リリカルなのは StrikerS 」
  について語りましょう。
  脚本 漫画 全て 都築作の な の は 出 す気だったりして
  なのは SS で 雄 犬 が ずっと 犬形態 な の は なぜ ？
  なんで なのは さん って 俺の嫁 なの ？
  防衛 政策 に 不可欠 な の は 国民的 合意 で あり 、 可能な 限り...

- 一部完全ではないが、重要な部分は本来、
  完全に自動的に学習できる！

山形弁の解析

http://www.nhk.or.jp/namara03-blog/

- NHK「今夜はなまらナイト」の視聴者ブログをクロール, 重複を除いて6,000文

教師なし:

次回 の 放送 、 あっかもすんね すない がもすんね ん だ べ けど 、
いずれ は 全国 放送 に … (笑)
おばん か だ っす 。 おらほ の 家 で は 、 生卵 や とろろ を 入れて
食べ ます 。 ほんて 、 んまえ ぞ ～ 。
えっちゃん さ ～ ん 、 是非 やまがだ さ き て の ～ 。

半教師あり:

次回 の 放送 、 あっかも すんねす ない が も すん ねんだべ けど 、
いずれ は 全国 放送 に … (笑)
おばん か だっす 。 おら ほの 家 で は 、 生卵 や とろろ を 入れて
食べ ます 。 ほんて 、 んまえ ぞ ～ 。
えっちゃん さ～ん 、 是非やまがだ さきて の ～ 。

何が問題?

- 半教師あり学習の一般的なモデル(JESS-CM):

識別モデル　　生成モデル

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

  - 2つのモデルの補間重みが常に固定！

- 言語モデルの高精度な補間と同様な問題
  - Jelinek-Mercer スムージング ＝線形補間

$$p(w|v) = \lambda_1 p(w) + (1-\lambda_1) p(w|v)$$

重みは常に一定

  - 階層Bayes (Dirichlet) スムージング (MacKay 1994)

$$p(w|v) = \frac{f(v)}{f(v)+\alpha_0}\hat{p}(w|v) + \frac{\alpha_0}{f(v)+\alpha_0}\bar{\alpha}_w$$

HMM的な仕組みの導入が必要

重みが動的に変化!

まとめ

- CRFと教師なし形態素解析を組み合わせた、半教師あり形態素解析
  - JESS-CM法の枠組で「教え合う」
  - Semi-Markov(単語)<->Markov(文字) の情報相互変換
- 人手の基準を守りつつ、未知語を完全に自動的に認識
- モデル補間重みの文脈依存化が今後の課題
  [半教師あり学習の一般的な問題]