階層ベイズ言語モデルと条件付確率場の統合による半教師あり形態素解析

2011年7月14日　統計数理研究所
オープンハウス

持橋 大地　モデリング研究系　准教授
鈴木潤、藤野昭典　NTTコミュニケーション科学基礎研究所 (共同研究)

形態素解析 ＝文字列を単語に分けること

統数研はいかがですか ➡ 統数研 は いかが です か
世界杯赛场地已准备停当 ➡ 世界杯 赛 场地 已 准备 停当

特にアジア圏では, 自然言語処理に不可欠

現在の問題点 “正しい”テキストしか解析できない

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…

CSJ日本語話し言葉コーパス
(国立国語研究所)

对@mo小颖 说:星期五吃饭~~！6点啊 [乐乐]
快下班了。。。。。可是我的时间才刚刚开始！
注册口碑卡~\(≧▽≦)/~啦啦啦
亲爱的，现在的我很成熟，我知道你对我意味着什么，我要好好珍惜你，我不要在错过一个值得我爱一生的女孩。
快艺考了。中西上戏，最差也要山艺锕。
哈哈哈……天真不嘛～可耐不嘛！雕塑课……明显是摆拍的-.-

新浪微博
(中国語版Twitter)

教師なし形態素解析　何か方法はないのか?

- 生の文字列だけから、完全に自動的に「単語」を学習
- 階層-階層Pitman-Yor過程(NPYLM) + Blocked MCMC

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 。

- 限界: 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”

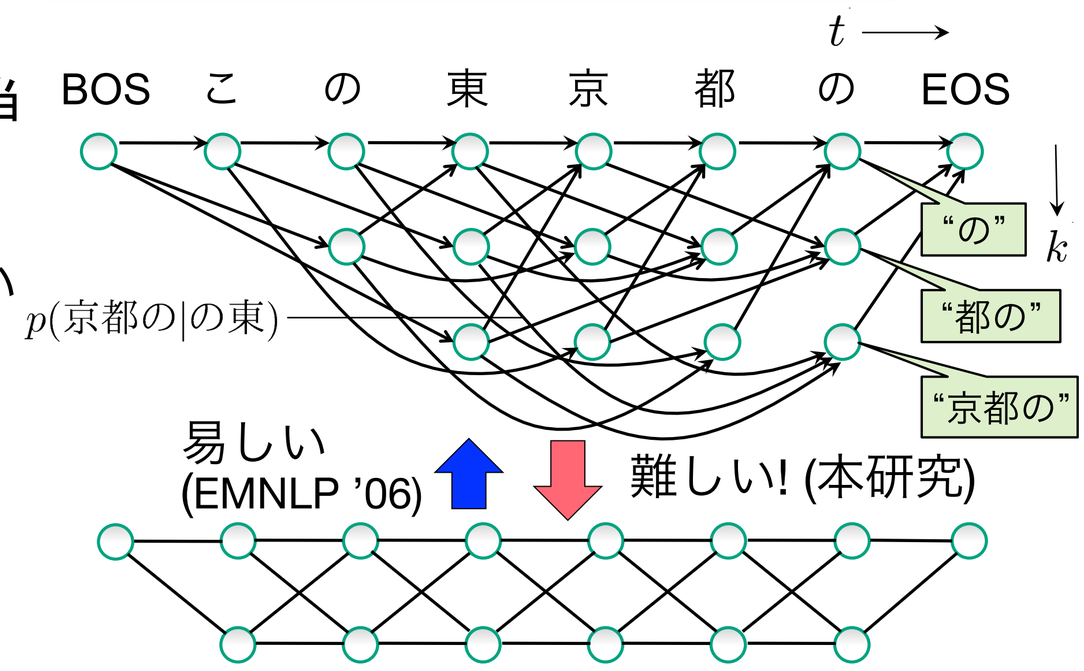
半教師あり学習: JESS-CM

識別モデル　生成モデル　重みλも学習

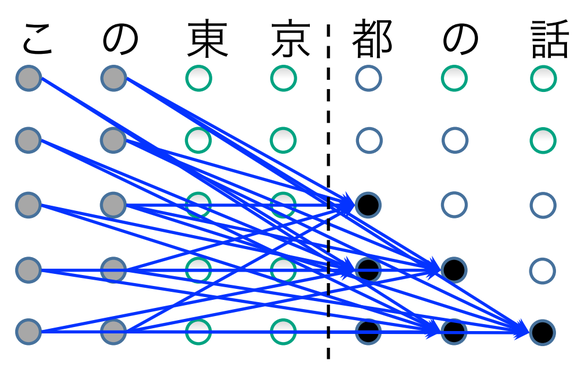
$$p(\mathbf{y}|\mathbf{x};\Lambda,\Theta) \propto p_{\mathrm{DISC}}(\mathbf{y}|\mathbf{x};\Lambda)\,p_{\mathrm{GEN}}(\mathbf{y},\mathbf{x};\Theta)^{\lambda}$$

- joint probability model embedding style semi-supervised conditional model (Suzuki+ ACL2008, ACL2009)
- 現在世界最高性能の半教師あり学習

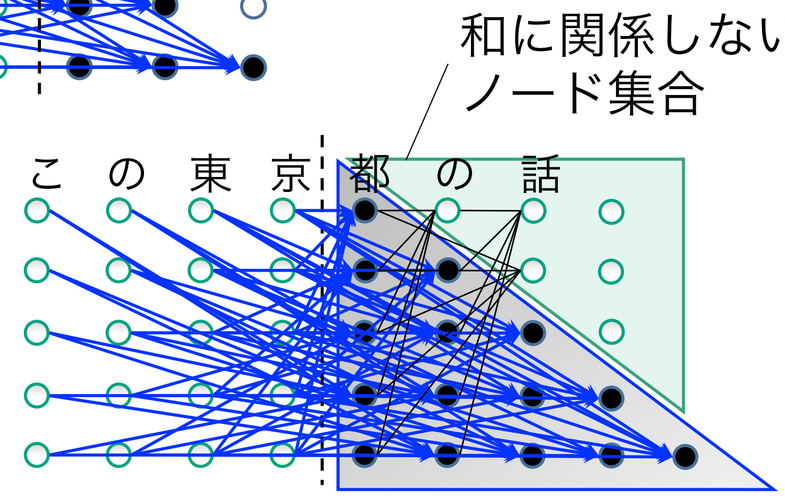
NPYLM=semi-Markov モデル



semi-Markov→Markov変換



例: Markovモデルの状態0→0に対応するパス集合



実験

「しょこたんブログ」の解析例

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(゜ω゜)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

新浪微博の解析結果

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
好饿 啊....走 ！ 妈妈 带 你 出去 吃饭 去~.....(((((ヾ(o=^･ェ･)o ┌┅┐ 喵~o(=∩ω∩=)m
學 校 學生 超 愛牠 的 ！ ！ ！[哈哈]

SIGHAN Bakeoff 2005 データセット MSRでの結果

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**

論文

持橋,鈴木,藤野:「ベイズ階層言語モデルと条件付確率場の統合による半教師あり形態素解析」,NLP2011, 2011.3.