ACL 2013 参加報告

NTTメディアインテリジェンス研究所
西川 仁

目次

1. ACL 概要
2. 分野動向
3. 論文紹介
   - Unsupervised Transcription of Historical Documents (Berg-Kirkpatrick et al.)
   - HEADY: News headline abstraction through event pattern clustering (Alfonseca et al.)
4. まとめ

ACL 2013 概要

- The association for computational linguistics (ACL) 主催の国際会議で，今回51回目
  - 計算言語学分野を代表する会議で，世界中から関連分野の研究者が一堂に会する
- 期間：2013年8月4日～9日
- 開催地：ソフィア（ブルガリア）
  - 概ね欧州，米州，亜州で交互に開催
  - 昨年は済州島（韓国），来年はボルチモア（米国）
- 採択率は25% (Long: 26% (174/662), Short: 25% 154/624))
- 投稿論文は1286件，参加者数は990人となり，過去最大の規模となった

研究動向：主要分野の採択件数

Machine Translation & Multilinguality	**64**
Semantics	36
Sentiment Analysis, Opinion Mining and Text Classification	27
Syntax and Parsing	**26**
NLP Applications	25
Summarization and Generation	17
Statistical and Machine Learning Methods in NLP	16
Text Mining and Information Extraction	16
Language Resources	14
Discourse, Coreference, and Pragmatics	13
NLP for the Web and Social Media	**10**
Others	64

- 傾向としては例年と同様，機械翻訳や 構文解析が多数
- 多様性を重視する方向性にあり，珍しい研究が割と採択されている
- ソーシャルメディアに関する研究が増加

研究動向：分野概観 1/3

- 機械翻訳
  - 主流は syntax を考慮した方法， Parsing と関連した話題が多い (Liu)
  - Rich な情報を使って翻訳する方向性も多数 (Goto et al.; Green et al)
  - Alignment では DNN を使ったものも (Yang et al.)
- 構文解析
  - DNN を利用した手法が出現 (Socher et al.)
  - 大多数は個別の問題をしっかり解いている印象

研究動向：分野概観 2/3

- NLP 応用
  - ソーシャルメディアの普及に伴い Sentiment Analysis が復権 (Si et al., 他多数)
  - ツイートに含まれる実体の同定 (Liu et al.) など，ツイッター 関連するものが多数
  - コード中のコメントを予測するという変わり種も (Movshovitz-Attias and Cohen)
- Semantics
  - 多言語での意味役割付与 (Andreas et al.; Kozhevnikov and Titov) など
  - English possessive の解析 (Tratz and Hovy) や量化子のスコープ同定 (Manshadi and Allen) など渋い話も

研究動向：分野概観 3/3

- 要約・生成
  - 要約についてはまだ抽出＋短縮が主流 (Almeida and Martins; Morita et al.; Dasgupta et al.,)
  - 一方，徐々に生成的手法が出現 (Alfonseca et al ; Cheung and Penn)
- その他
  - 古い文書の Transcription （ OCR の精度向上） (Berg-Kirkpatrick et al.) は興味深い（例えば google books に1900年代以前の古い文献を追加するために使える）
  - その他，暗号解読 (Ravi) を SMT の精度向上に使うもの（教師なし統計的機械翻訳とみなせる）など

ご紹介する文献

1. Unsupervised Transcription of Historical Documents (Berg-Kirkpatrick et al.)
2. HEADY: News headline abstraction through event pattern clustering (Alfonseca et al.)

Unsupervised Transcription of Historical Documents

- T. Berg-Kirkpatrick, G. Durrett and D. Klein (U. Cal. at Berkeley)
- 昔のテキストを OCR したい
- 既存の OCR ツールでは太刀打ちできない
- 生成モデルを考えて Unsupervised に解く

Berg-Kirkpatrick et al., Unsupervised Transcription of Historical Documents, Proc. of ACL, pp. 207—217, 2013. Figure 1 より引用

Unsupervised Transcription of Historical Documents

- 大きくわけて3つの問題

Berg-Kirkpatrick et al., Unsupervised Transcription of Historical Documents, Proc. of ACL, pp. 207—217, 2013. Figure 1 より引用

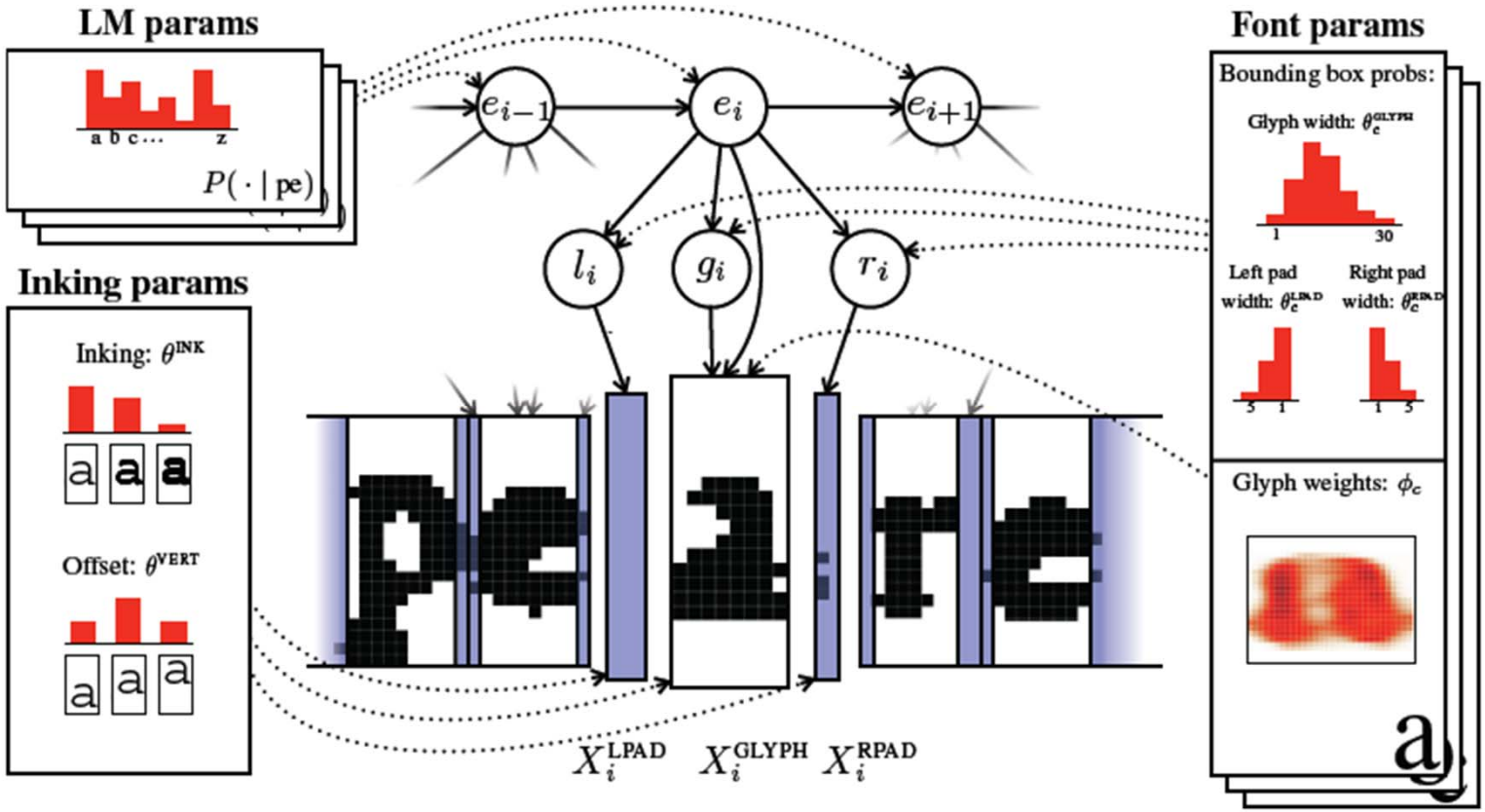
謎のフォント（今と昔でフォントが違う）

字の高さが違い過ぎる（活版印刷なので大変）

インクがにじみ過ぎ

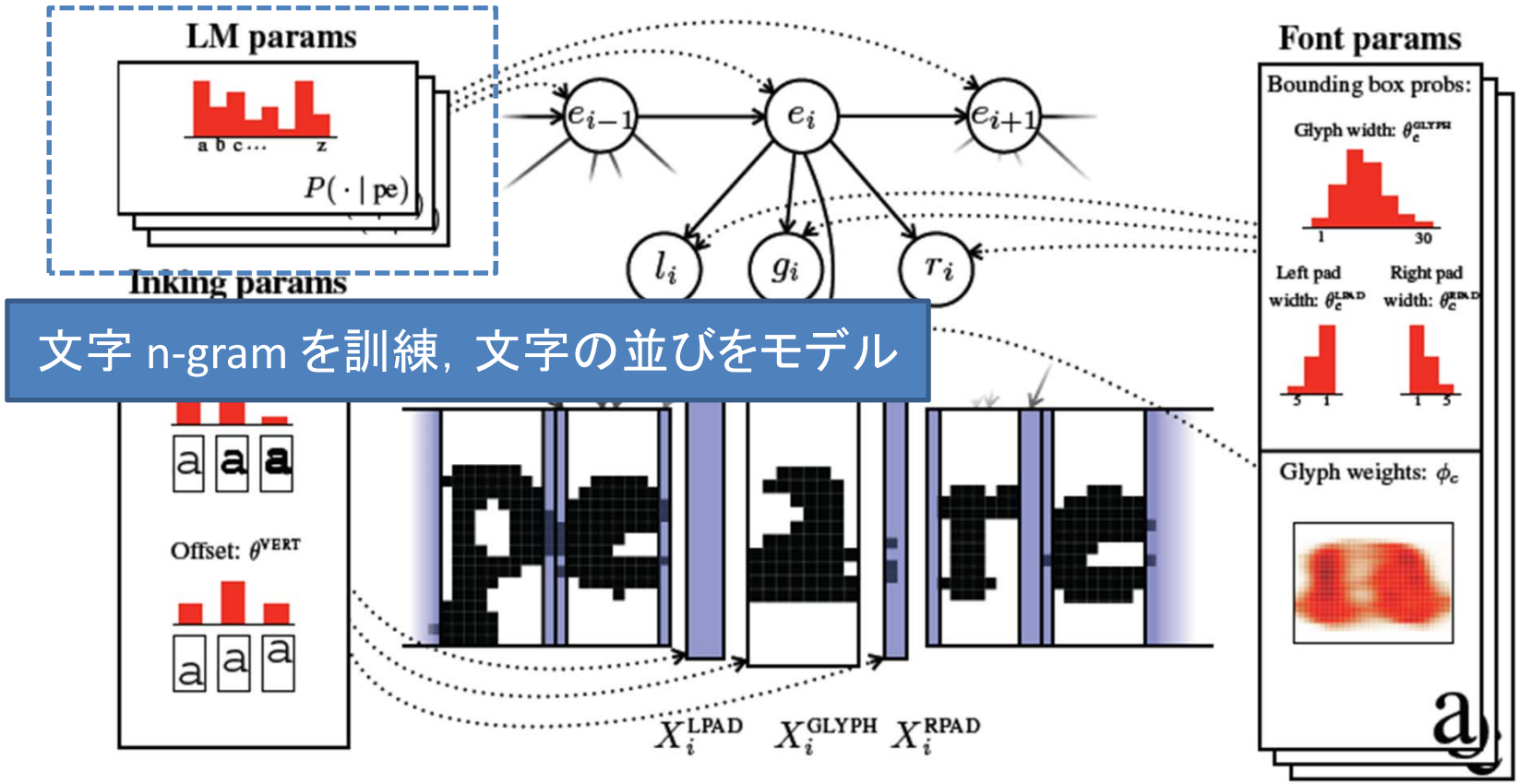
Berg-Kirkpatrick et al.,
Unsupervised Transcription
of Historical Documents,
Proc. of ACL, pp. 207—217,
2013.
Figure 3 より引用

The big picture



Berg-Kirkpatrick et al.,
Unsupervised Transcription
of Historical Documents,
Proc. of ACL, pp. 207—217,
2013.
Figure 3 より引用

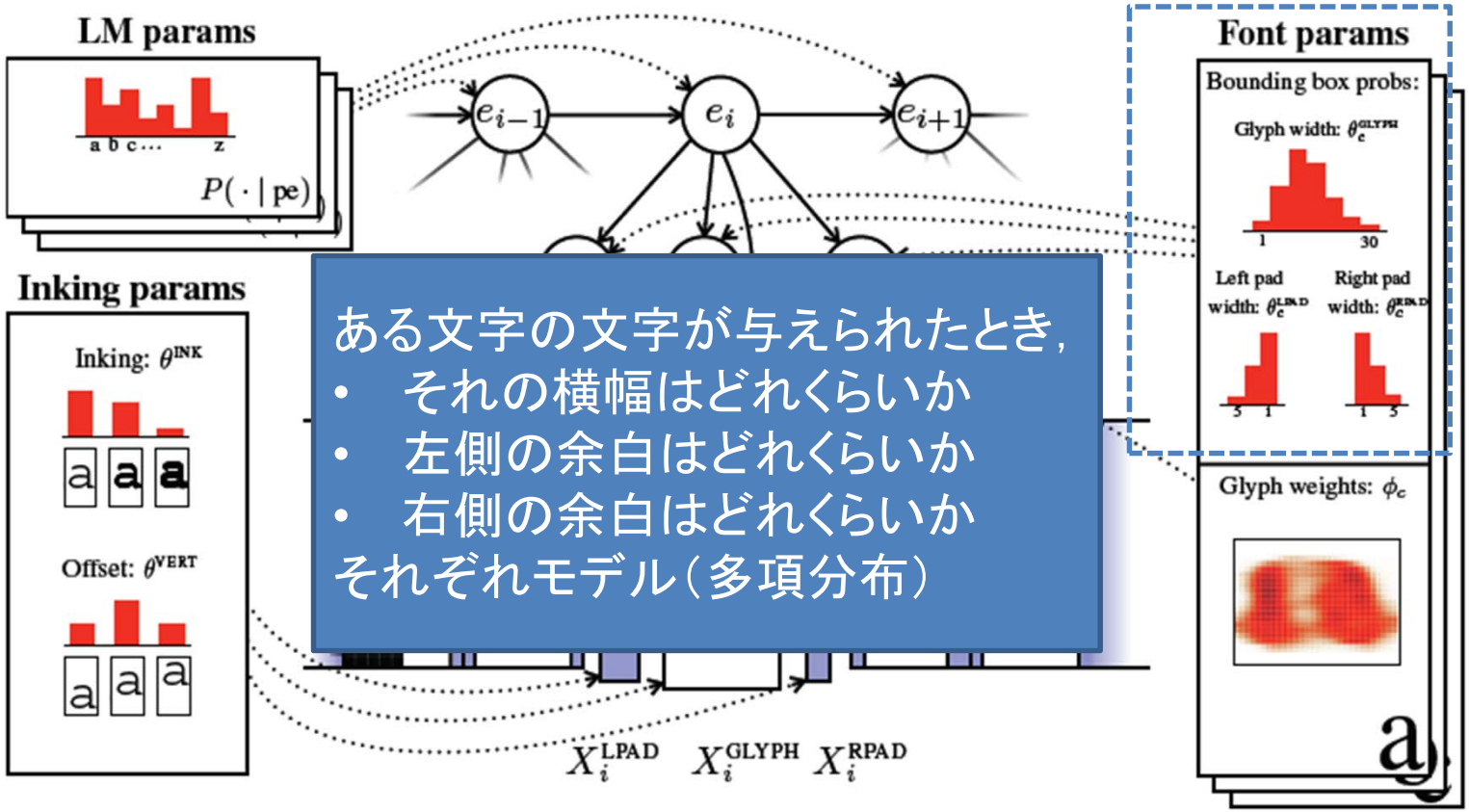
The big picture



文字 n-gram を訓練，文字の並びをモデル

Berg-Kirkpatrick et al.,
Unsupervised Transcription
of Historical Documents,
Proc. of ACL, pp. 207—217,
2013.
Figure 3 より引用

The big picture

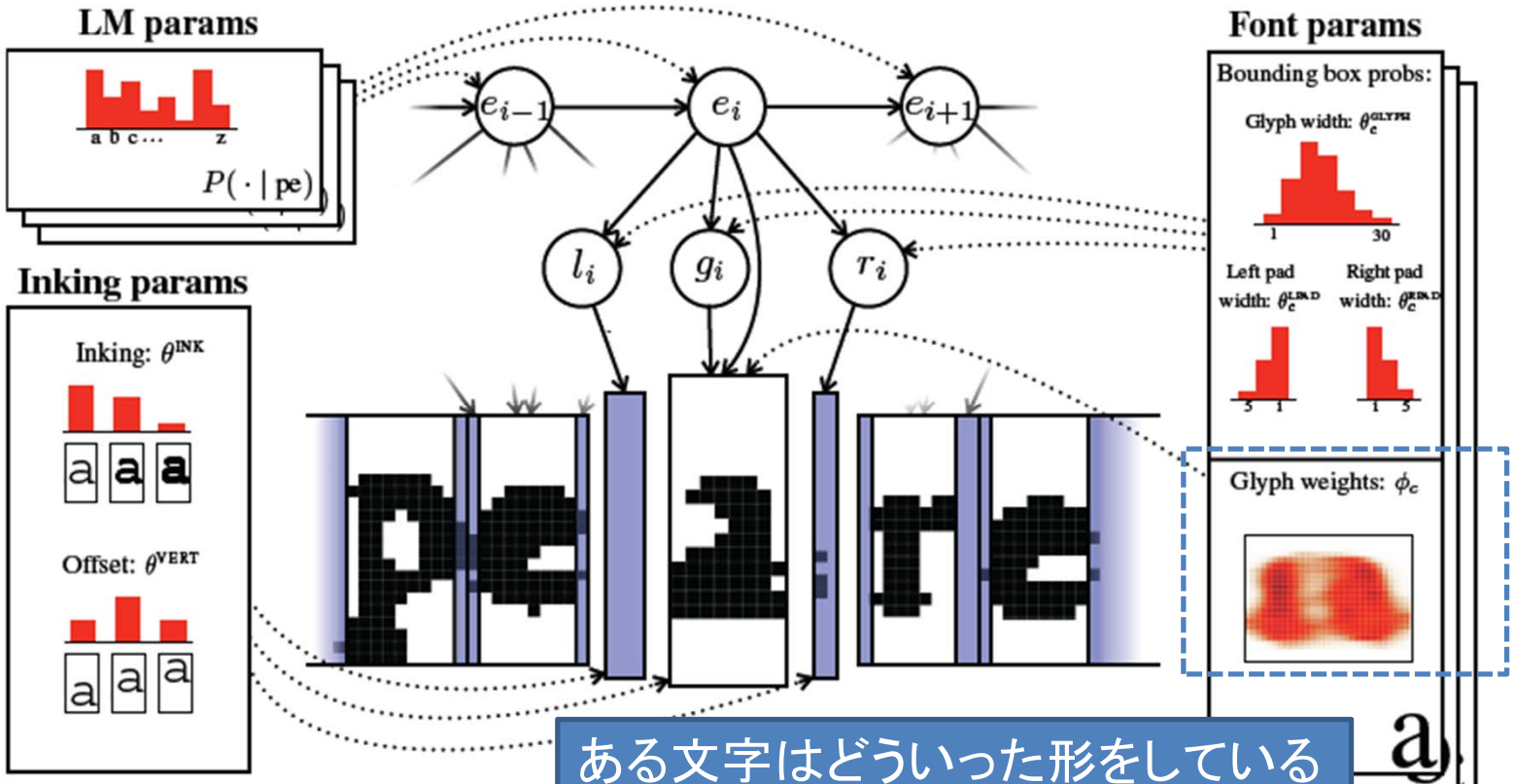


ある文字の文字が与えられたとき，
- それの横幅はどれくらいか
- 左側の余白はどれくらいか
- 右側の余白はどれくらいか

それぞれモデル（多項分布）

Berg-Kirkpatrick et al.,
Unsupervised Transcription
of Historical Documents,
Proc. of ACL, pp. 207—217,
2013.
Figure 3 より引用

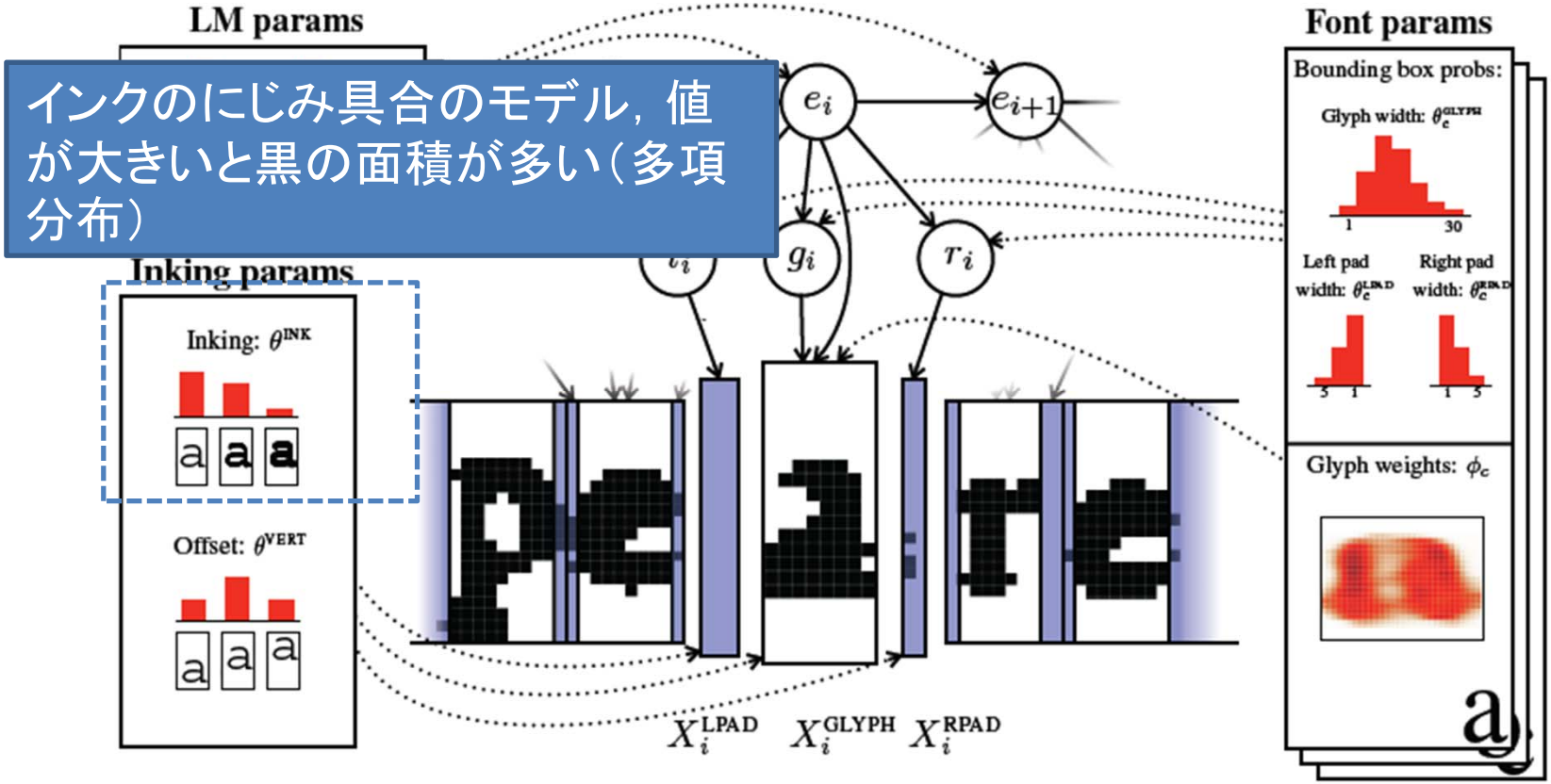
The big picture



ある文字はどういった形をしているかモデル（初期値は既存のフォントセットで与える）

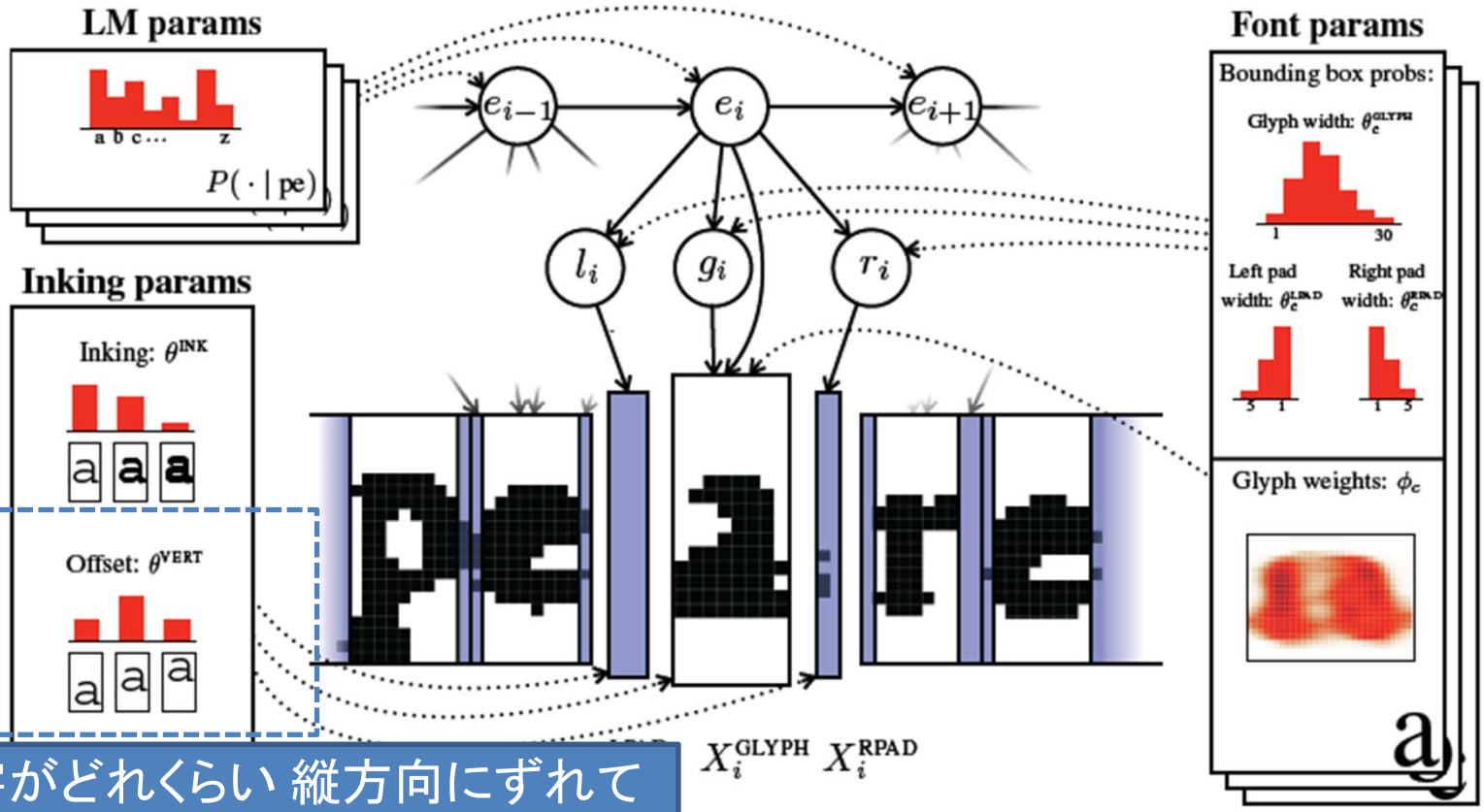
Berg-Kirkpatrick et al., Unsupervised Transcription of Historical Documents, Proc. of ACL, pp. 207—217, 2013.
Figure 3 より引用

The big picture



Berg-Kirkpatrick et al.,
Unsupervised Transcription
of Historical Documents,
Proc. of ACL, pp. 207—217,
2013.
Figure 3 より引用

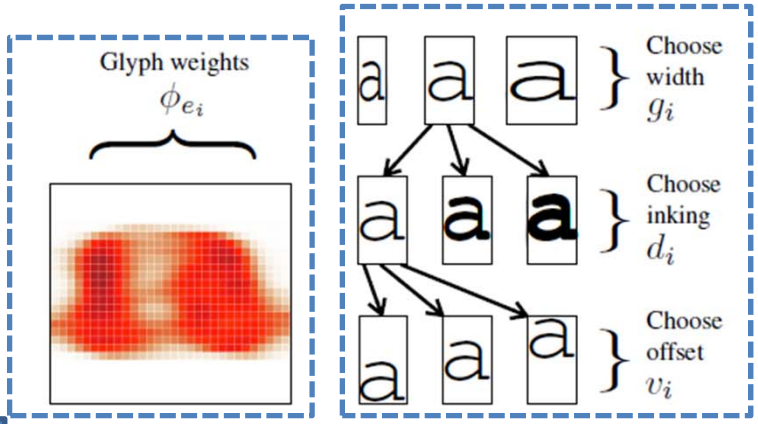
The big picture



文字がどれくらい 縦方向にずれているかモデル，値が大きいと縦にずれる（多項分布）

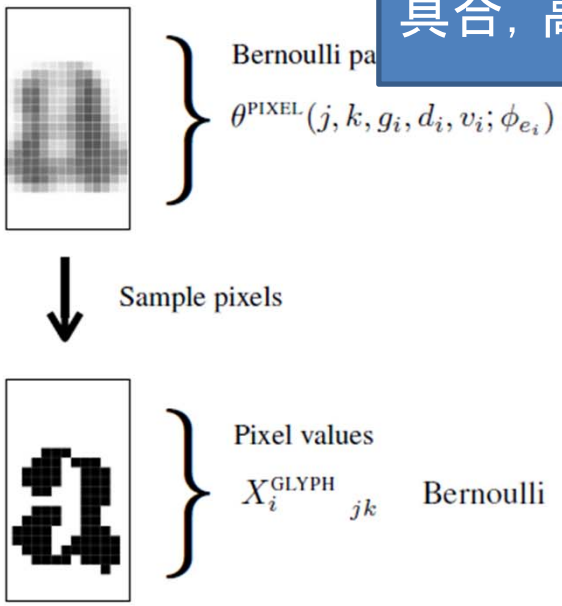
Berg-Kirkpatrick et al., Unsupervised Transcription of Historical Documents, Proc. of ACL, pp. 207—217, 2013.
Figure 4 より引用

文字の生成



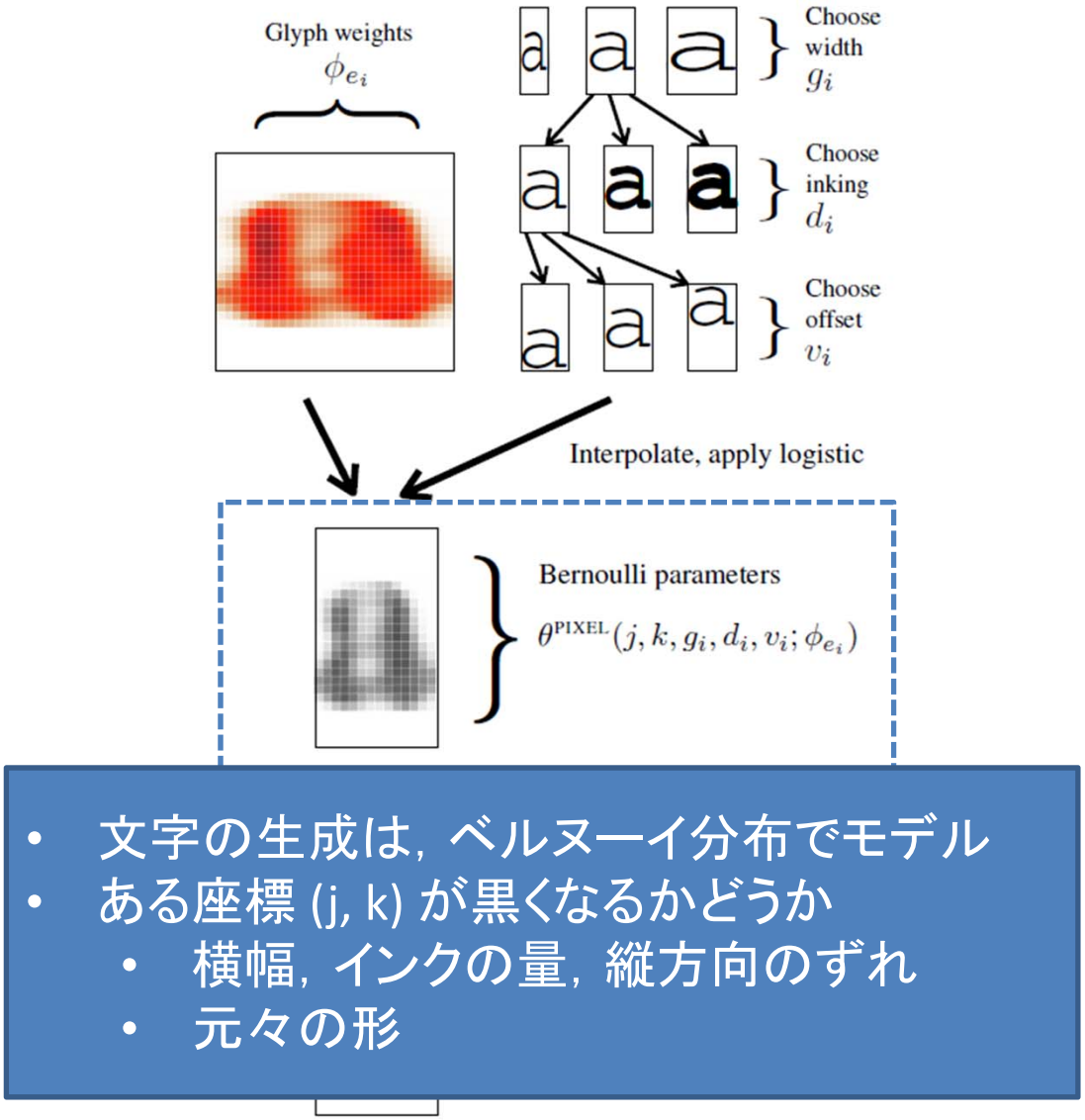
文字の形

ある文字の横幅，インクのにじみ具合，高さのずれ



Berg-Kirkpatrick et al., Unsupervised Transcription of Historical Documents, Proc. of ACL, pp. 207—217, 2013.
Figure 4 より引用

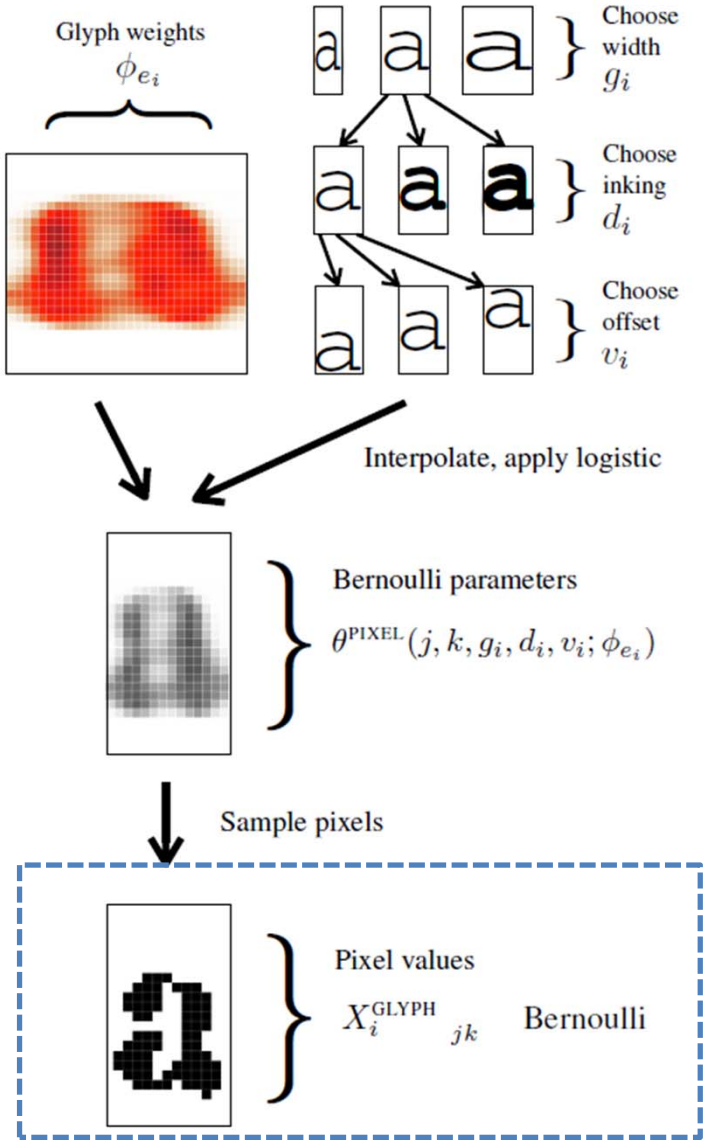
文字の生成



- 文字の生成は，ベルヌーイ分布でモデル
- ある座標 (j, k) が黒くなるかどうか
  - 横幅，インクの量，縦方向のずれ
  - 元々の形

Berg-Kirkpatrick et al., Unsupervised Transcription of Historical Documents, Proc. of ACL, pp. 207—217, 2013.
Figure 4 より引用

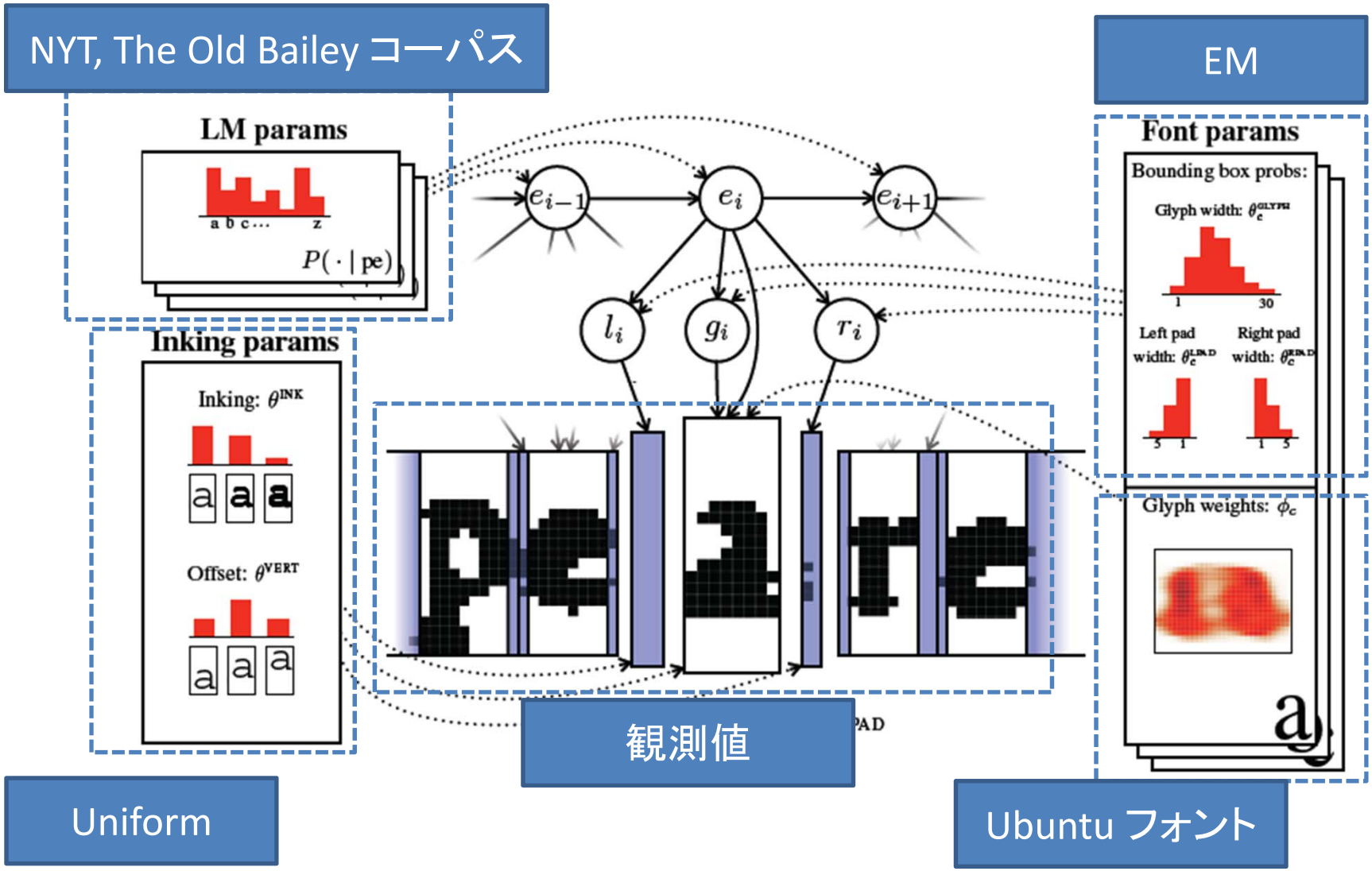
文字の生成



サンプリング（上のモデル（ベルヌーイ分布とそのパラメタ）が与えられたとき，この文字の生成されやすさ（確率）がわかる

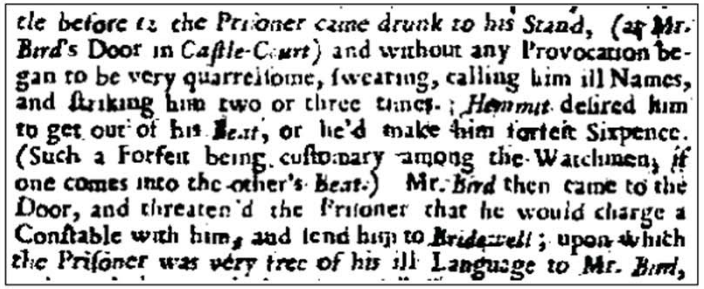
Berg-Kirkpatrick et al.,
Unsupervised Transcription of
Historical Documents, Proc. of ACL,
pp. 207—217, 2013.
Figure 3 より引用

Learning

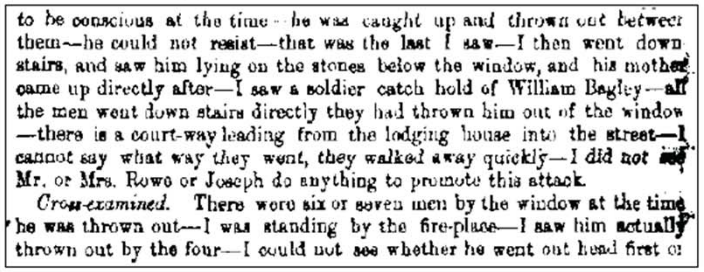


実験

(a) Old Bailey, 1725:

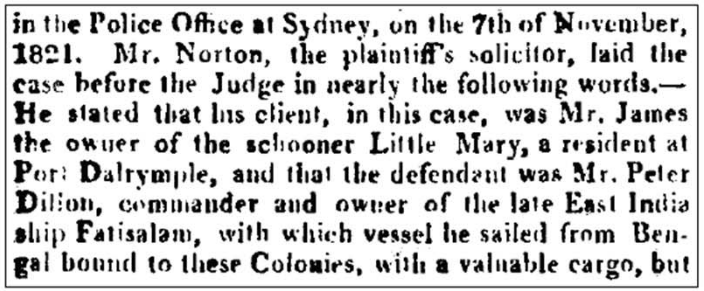

tle before the Prisoner came drunk to his Stand, (at Mr. Bird's Door in Castle-Court) and without any Provocation began to be very quarrelsome, swearing, calling him ill Names, and striking him two or three times; Hemmut desired him to get out of his Beat, or he'd make him forfeit Sixpence. (Such a Forfeit being customary among the Watchmen, if one comes into the other's Beat.) Mr. Bird then came to the Door, and threaten'd the Prisoner that he would charge a Constable with him, and send him to Bridewell; upon which the Prisoner was very free of his ill Language to Mr. Bird,

(b) Old Bailey, 1875:

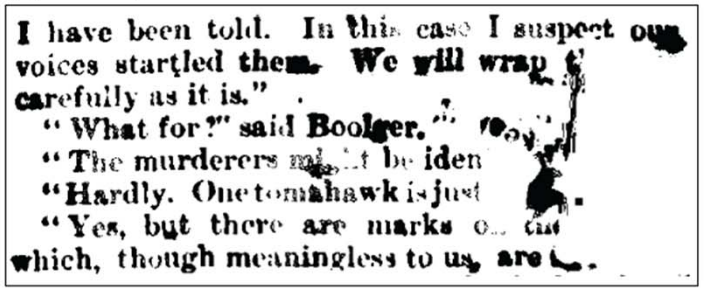

to be conscious at the time—he was caught up and thrown out between them—he could not resist—that was the last I saw—I then went down stairs, and saw him lying on the stones below the window, and his mother came up directly after—I saw a soldier catch hold of William Bagley—all the men went down stairs directly they had thrown him out of the window—there is a court-way leading from the lodging house into the street—I cannot say what way they went, they walked away quickly—I did not see Mr. or Mrs. Rowe or Joseph do anything to promote this attack.

Cross-examined. There were six or seven men by the window at the time he was thrown out—I was standing by the fire-place—I saw him actually thrown out by the four—I could not see whether he went out head first or

(c) Trove, 1823:


in the Police Office at Sydney, on the 7th of November, 1821. Mr. Norton, the plaintiff's solicitor, laid the case before the Judge in nearly the following words.—He stated that his client, in this case, was Mr. James the owner of the schooner Little Mary, a resident at Port Dalrymple, and that the defendant was Mr. Peter Dillon, commander and owner of the late East India ship Fatisalam, with which vessel he sailed from Bengal bound to these Colonies, with a valuable cargo, but

(d) Trove, 1883:


I have been told. In this case I suspect our voices startled them. We will wrap it carefully as it is."

"What for?" said Boolger."

"The murderers must be iden

"Hardly. One tomahawk is just

"Yes, but there are marks o... which, though meaningless to us, are ...

- コーパス
  - Old Bailey：英国の刑事法廷の議事録
  - Trove: オーストラリアの新聞
- 評価方法
  - CER, WER
- 1行単位で，どれくらいあたるか評価
- 行は自動的に検出

Berg-Kirkpatrick et al.,
Unsupervised Transcription of
Historical Documents, Proc. of ACL,
pp. 207—217, 2013.
Figure 6 より引用

結果

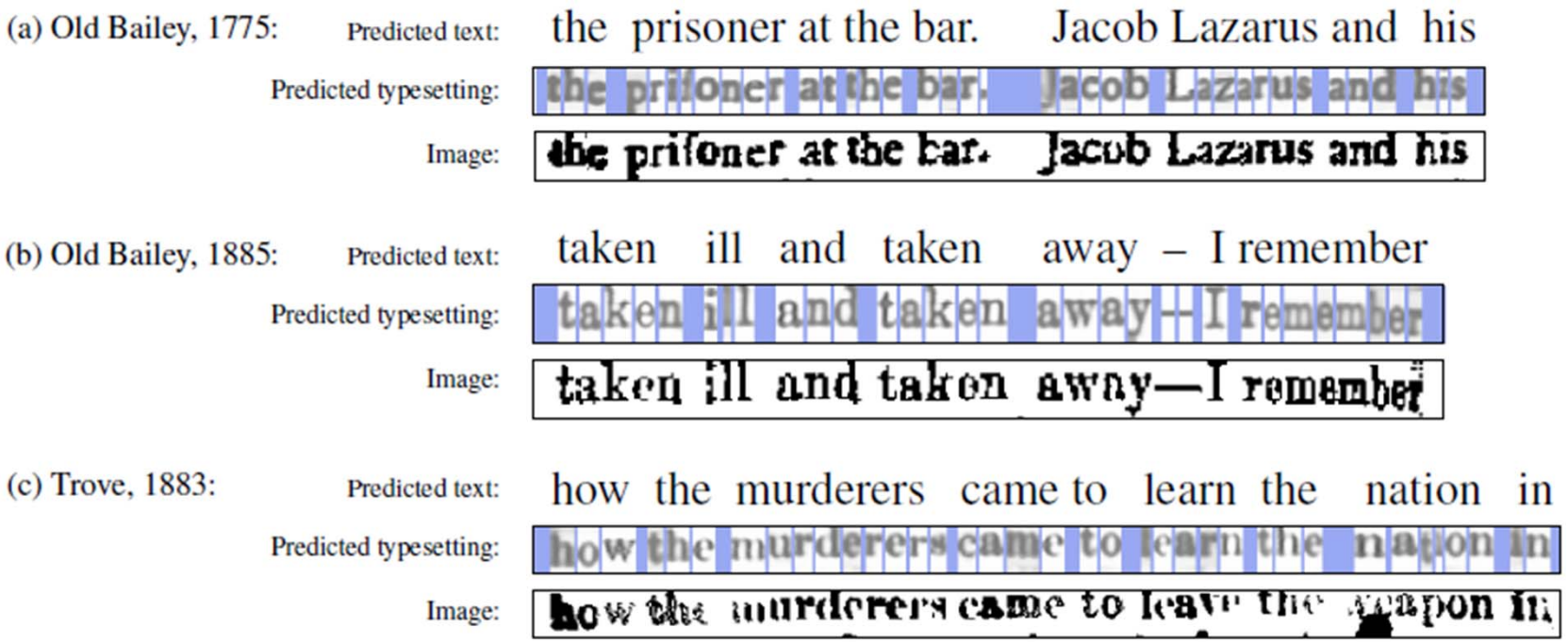
Berg-Kirkpatrick et al.,
Unsupervised Transcription of
Historical Documents, Proc. of ACL,
pp. 207—217, 2013.
Table 1 より引用

System	CER	WER
Old Bailey		
Google Tesseract	29.6	54.8
ABBYY FineReader	15.1	40.0
Ocular w/ NYT (this work)	12.6	28.1
Ocular w/ OB (this work)	**9.7**	**24.1**
Trove		
Google Tesseract	37.5	59.3
ABBYY FineReader	22.9	49.2
Ocular w/ NYT (this work)	**14.9**	**33.0**

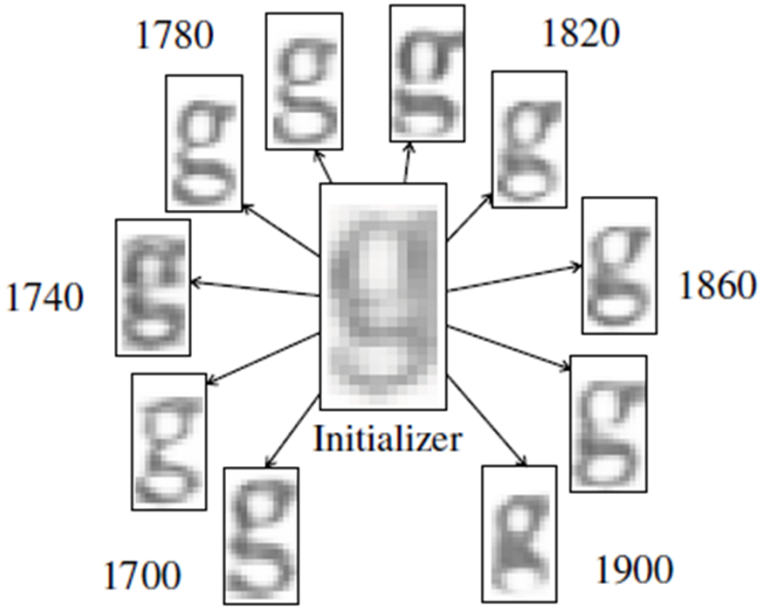
- 提案法が最もよい

例

Berg-Kirkpatrick et al.,
Unsupervised Transcription of
Historical Documents, Proc. of ACL,
pp. 207—217, 2013.
Figure 7 より引用

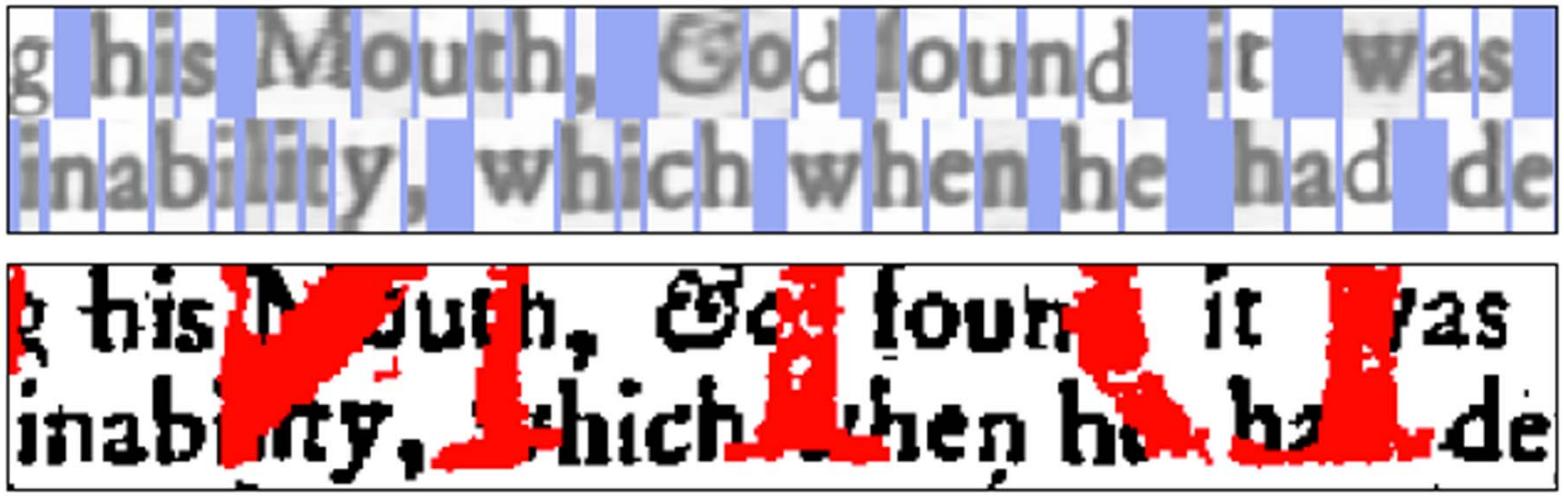


学習された文字



Berg-Kirkpatrick et al.,
Unsupervised Transcription of
Historical Documents, Proc. of ACL,
pp. 207—217, 2013.
Figure 8 より引用

強烈な染みにも頑健



Berg-Kirkpatrick et al.,
Unsupervised Transcription of
Historical Documents, Proc. of ACL,
pp. 207—217, 2013.
Figure 9 より引用

残存するエラー

- 句読点
- イタリック
- 本当にひどいインクの染み

まとめ

- 教師なし OCR で古い文献を認識
- 面白い
- 細かい問題点の分析とモデルへの反映
- ものすごい難しいことをしているわけではない

HEADY: News headline abstraction through event pattern clustering

- E. Alfonseca, D. Pighin (Google) and G. Garrido (UNED)
- 複数のニュース記事から，1つのヘッドラインを生成する問題
  - Google News で aggregate されている記事の集合に対して1つのヘッドラインを付与するような問題

ヘッドラインはたくさんある

- Carmelo and La La Party It Up with Kim and Ciara
- La La Vazquez and Carmelo Anthony: Wedding Day Bliss
- Carmelo Anthony, actress LaLa Vazquez wed in NYC
- Stylist to the Stars
- LaLa, Carmelo Set Off Celebrity Wedding Weekend
- Ciara rocks a sexy Versace Spring 2010 mini to LaLa Vasquez and Carmelo Anthony's wedding (photos)
- Lala Vasquez on her wedding dress, cake, reality tv show and fiancé, Carmelo Anthony (video)
- VAZQUEZ MARRIES SPORTS STAR ANTHONY
- Lebron Returns To NYC For Carmelo's Wedding
- Carmelo Anthony's stylist dishes on the wedding
- Paul pitching another Big Three with "Melo in NYC"
- Carmelo Anthony and La La Vazquez Get Married at Star-Studded Wedding Ceremony

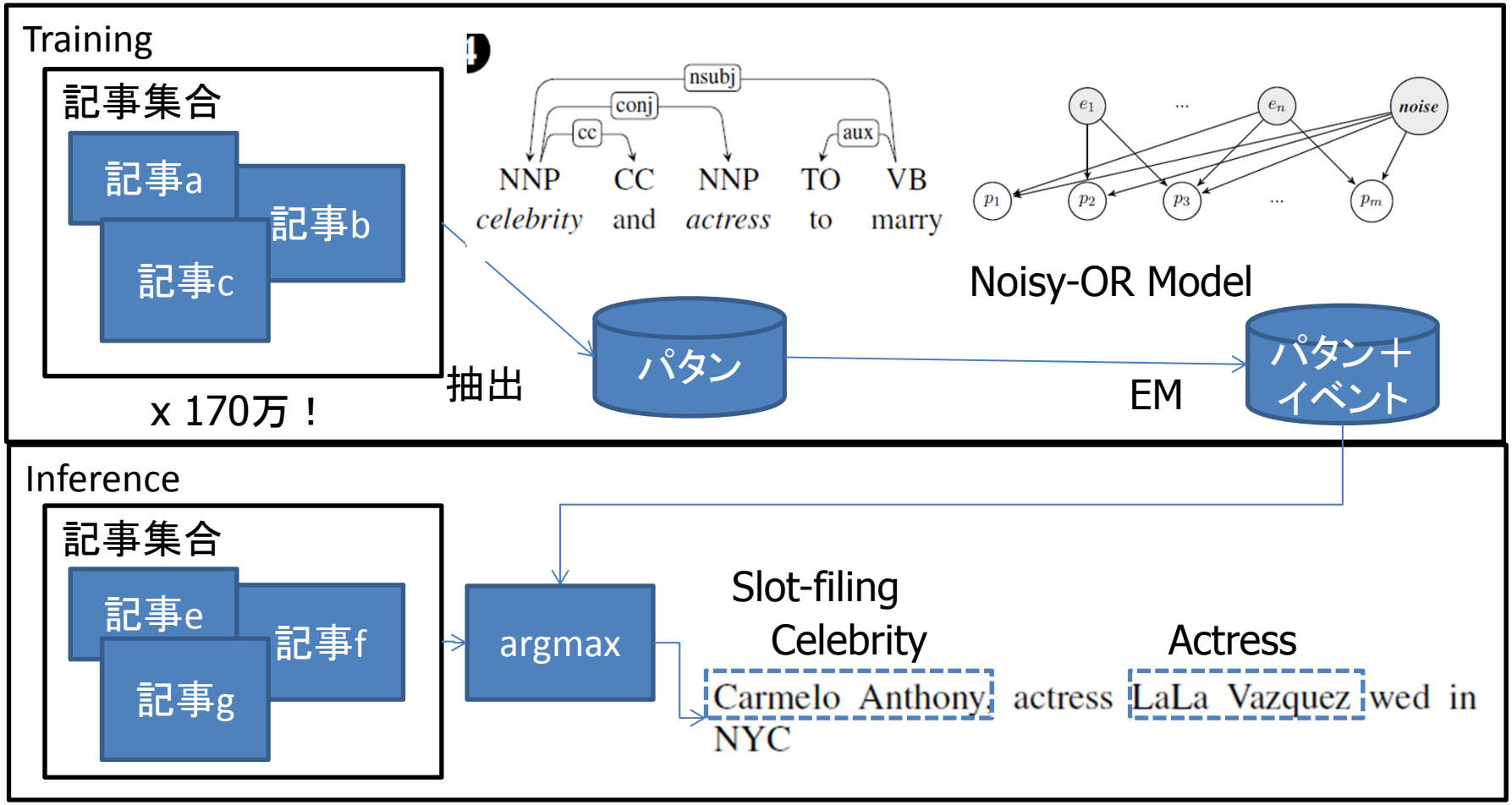
- 適当によさげなのを選べばいいんじゃないの？？？
- ダメ！
- 主観的でなく，センセーショナルでもないような見出しがない場合に困ってしまう（らしい）

Alfonseca et al., HEADY: News headline abstraction through event pattern clustering, Proc. of ACL, pp. 1243—1253, 2013.
Table 1 より引用

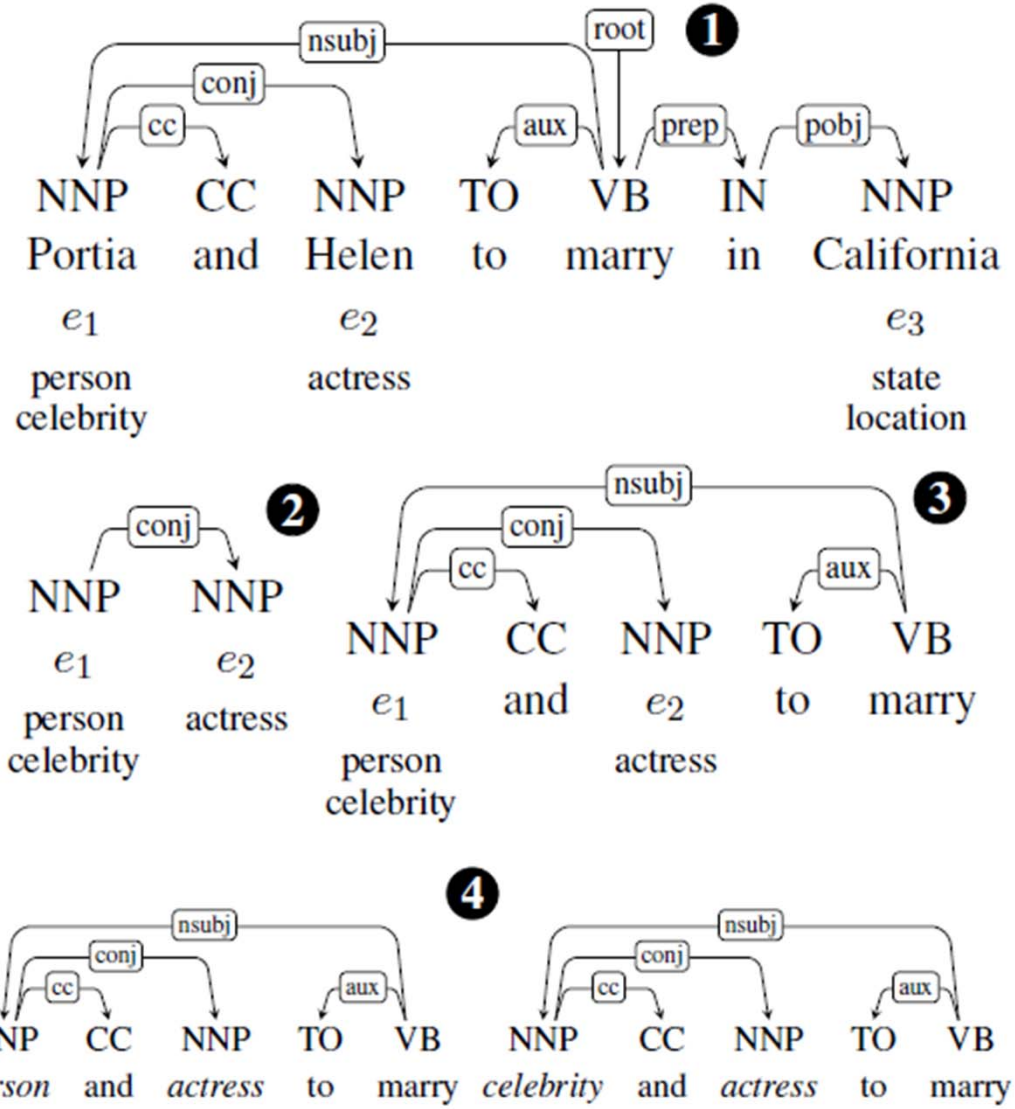
Alfonseca et al., HEADY: News headline abstraction through event pattern clustering, Proc. of ACL, pp. 1243—1253, 2013.
Figure 1, 2 より引用

The big picture

- パタンを事前に大量に持っておいて，新しいテキスト集合がやってきたときに良さげなパタンを選んで slot-filling
- 潜在変数として「イベント」を考え，イベントからパタンが生成されていると仮定

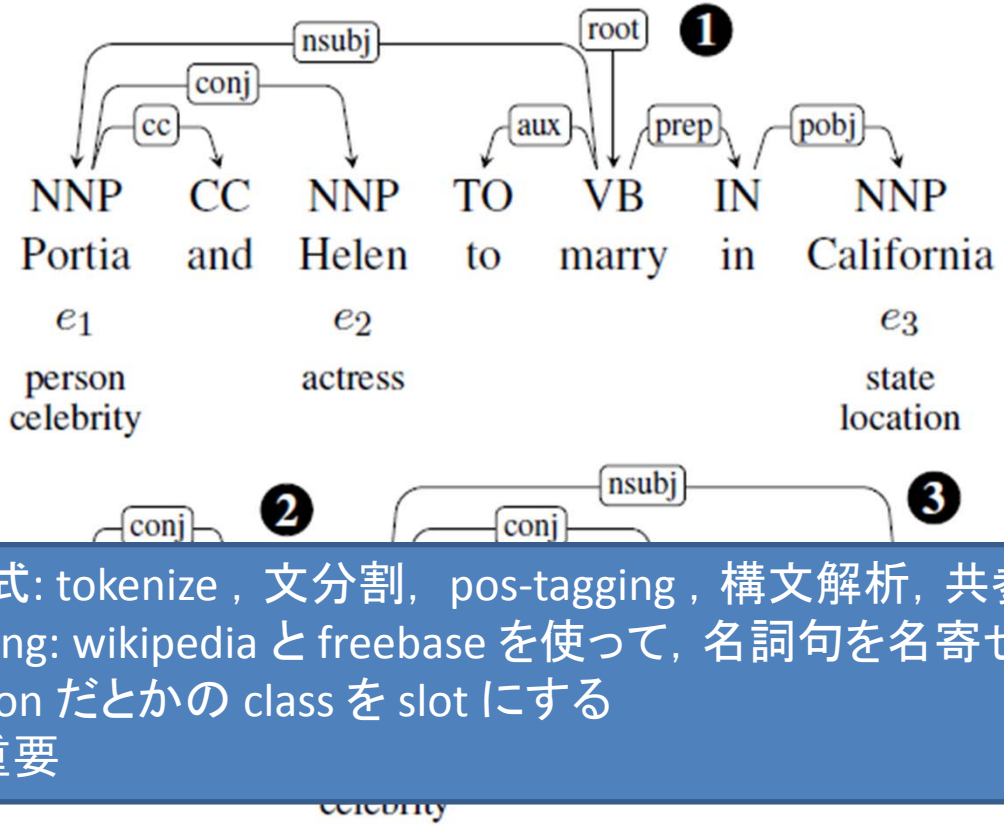


パタン抽出



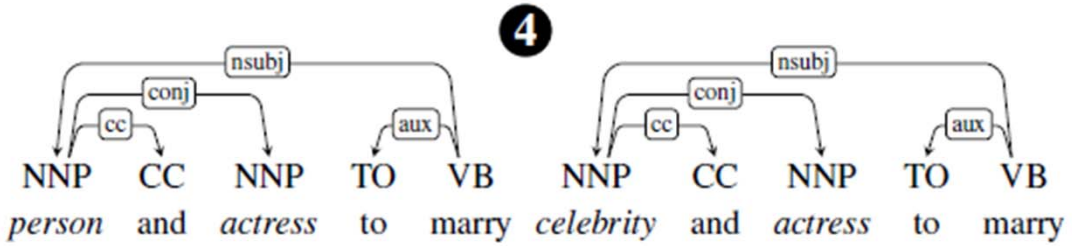
Alfonseca et al., HEADY: News headline abstraction through event pattern clustering, Proc. of ACL, pp. 1243—1253, 2013.
Figure 1 より引用

パタン抽出



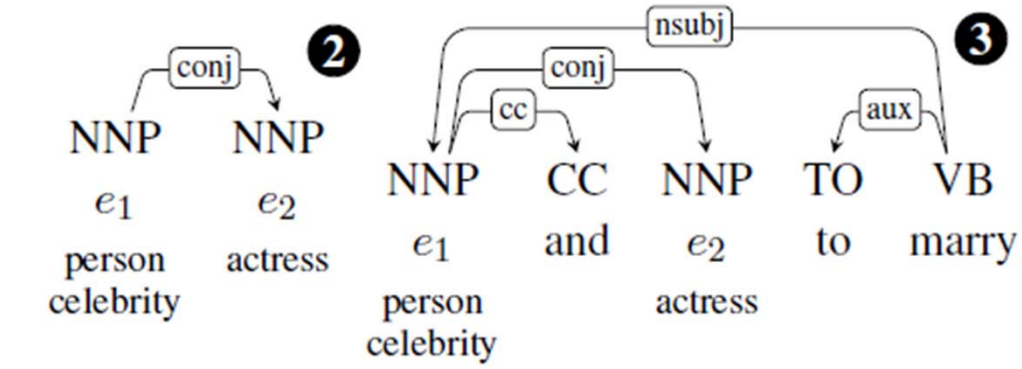
- 前処理一式: tokenize，文分割，pos-tagging，構文解析，共参照解析）
- entity-linking: wikipedia と freebase を使って，名詞句を名寄せ，person だとか location だとかの class を slot にする
  - ここ重要

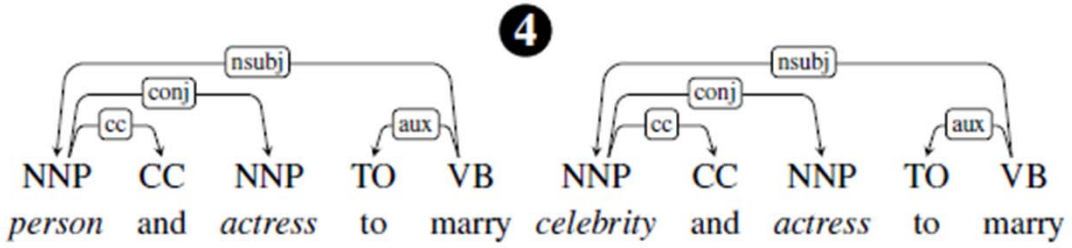
Alfonseca et al., HEADY: News headline abstraction through event pattern clustering, Proc. of ACL, pp. 1243—1253, 2013.
Figure 1 より引用



パタン抽出

- 記事中で一番よく出てくる entity（この例だと結婚した人とか）を選ぶ
- その entity と，同じ記事に出てくる他の entity の組み合わせ（集合）を生成
- それぞれ集合の要素間の MSP を同定（例では2つだが，3つの entity を含む msp まで取る）
- 単に MSP を取ると変な文が取れるので，ルールで head を足すなどする
- 出来た MSP をパタンとして利用

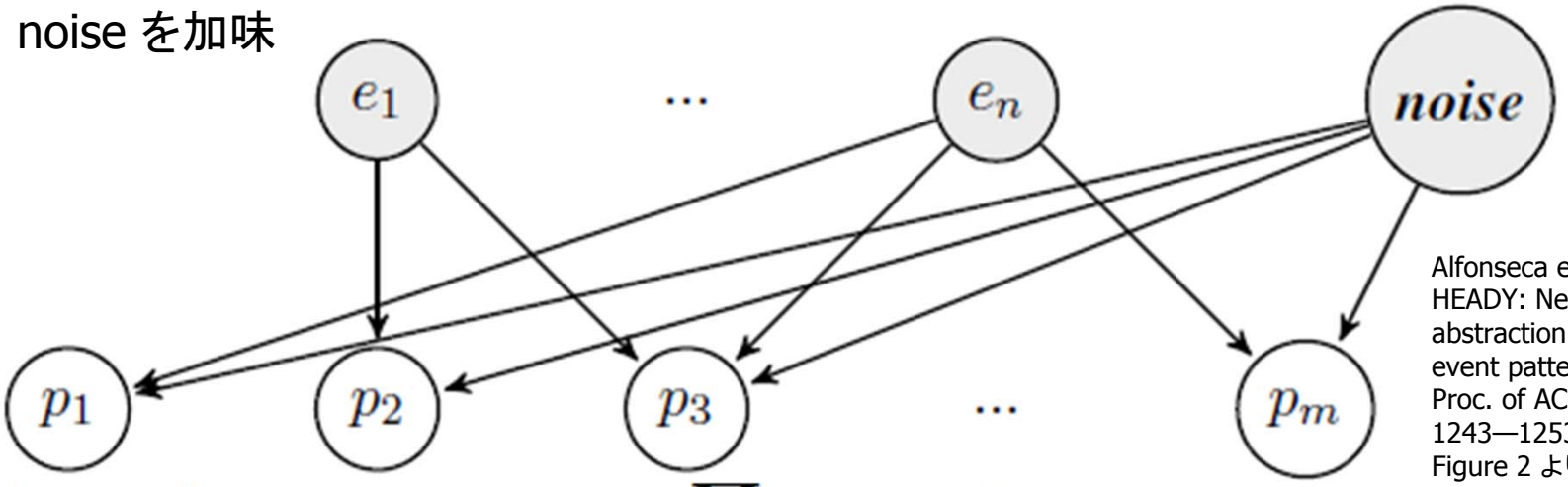




Alfonseca et al., HEADY: News headline abstraction through event pattern clustering, Proc. of ACL, pp. 1243—1253, 2013.
Figure 1 より引用

Noisy-OR model (Pearl, 1988)

- パタン p1..pm の裏にイベント e1..en があると仮定（記事集合は何かイベントを内包していて，そのイベントからパタンが生成された，と考える）
- ただし，もしかしたら全然関係なくパタンが出てきたりする可能性もあるので 陽に noise を加味



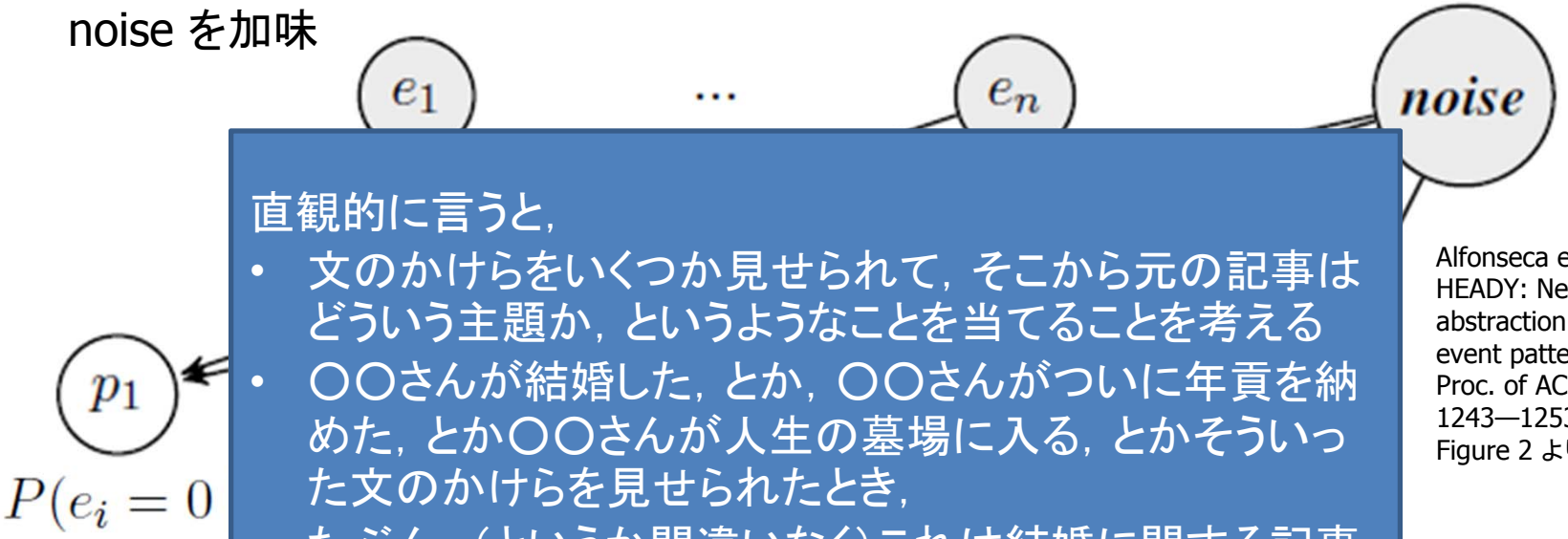
Alfonseca et al., HEADY: News headline abstraction through event pattern clustering, Proc. of ACL, pp. 1243—1253, 2013. Figure 2 より引用

$$P(e_i = 0 \mid \boldsymbol{p}) = (1 - q_{i0}) \prod_{j \in \pi_j} (1 - q_{ij})^{p_j}$$

$$= \exp\left(-\theta_{i0} - \sum_{j \in \pi_i} \theta_{ij} p_j \right) ,$$

Noisy-OR model (Pearl, 1988)

- パタン p1..pm の裏にイベント e1..en があると仮定（記事集合は何かイベントを内包していて，そのイベントからパタンが生成された，と考える）
- ただし，もしかしたら全然関係なくパタンが出てきたりする可能性もあるので 陽に noise を加味



直観的に言うと，

- 文のかけらをいくつか見せられて，そこから元の記事はどういう主題か，というようなことを当てることを考える
- ○○さんが結婚した，とか，○○さんがついに年貢を納めた，とか○○さんが人生の墓場に入る，とかそういった文のかけらを見せられたとき，
- たぶん，（というか間違いなく）これは結婚に関する記事からやってきたものであろう，ということがわかる
- ただ，出題者は意地悪なので，たまにダミーのかけらが混じっている（○○さんがそばを食べた，とか）

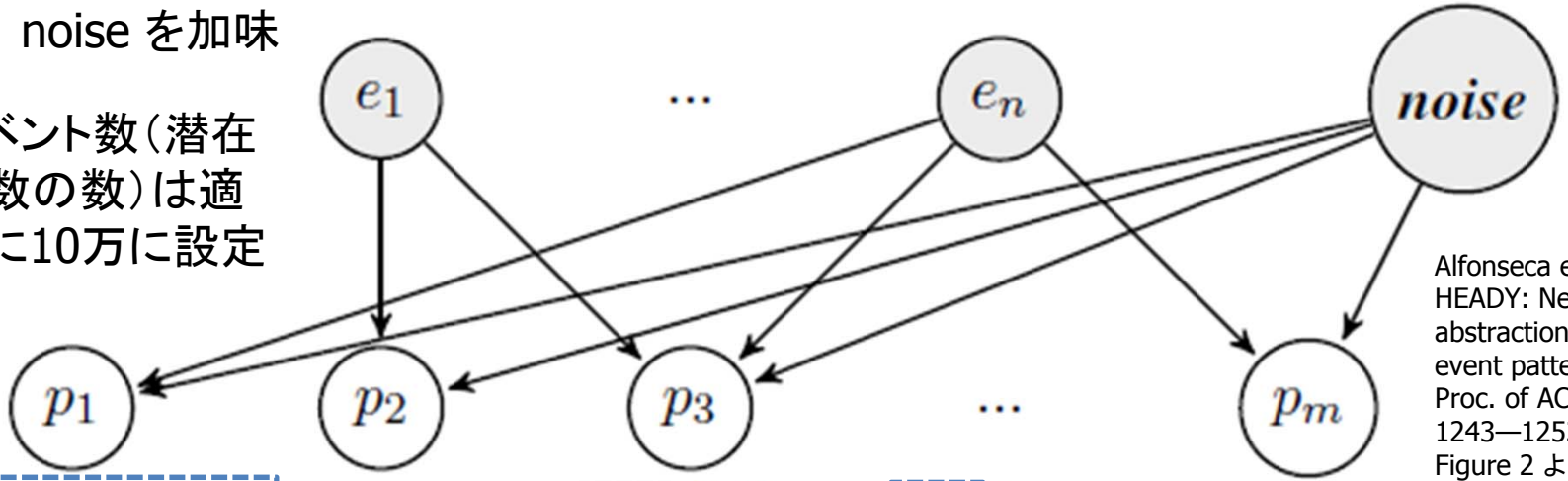
こんな感じのモデル

Alfonseca et al., HEADY: News headline abstraction through event pattern clustering, Proc. of ACL, pp. 1243—1253, 2013. Figure 2 より引用

Noisy-OR model (Pearl, 1988)

- パタン p1..pm の裏にイベント e1..en があると仮定（記事集合は何かイベントを内包していて，そのイベントからパタンが生成された，と考える）
- ただし，もしかしたら全然関係なくパタンが出てきたりする可能性もあるので 陽に noise を加味

イベント数（潜在変数の数）は適当に10万に設定



Alfonseca et al., HEADY: News headline abstraction through event pattern clustering, Proc. of ACL, pp. 1243—1253, 2013. Figure 2 より引用

$$P(e_i = 0 \mid \boldsymbol{p}) = (1 - q_{i0}) \prod_{j \in \pi_i} (1 - q_{ij})^{p_j}$$

$$= \exp\left(-\theta_{i0} - \sum_{j \in \pi_i} \theta_{ij} p_j \right),$$

あるイベントが記事集合に含まれていない確率

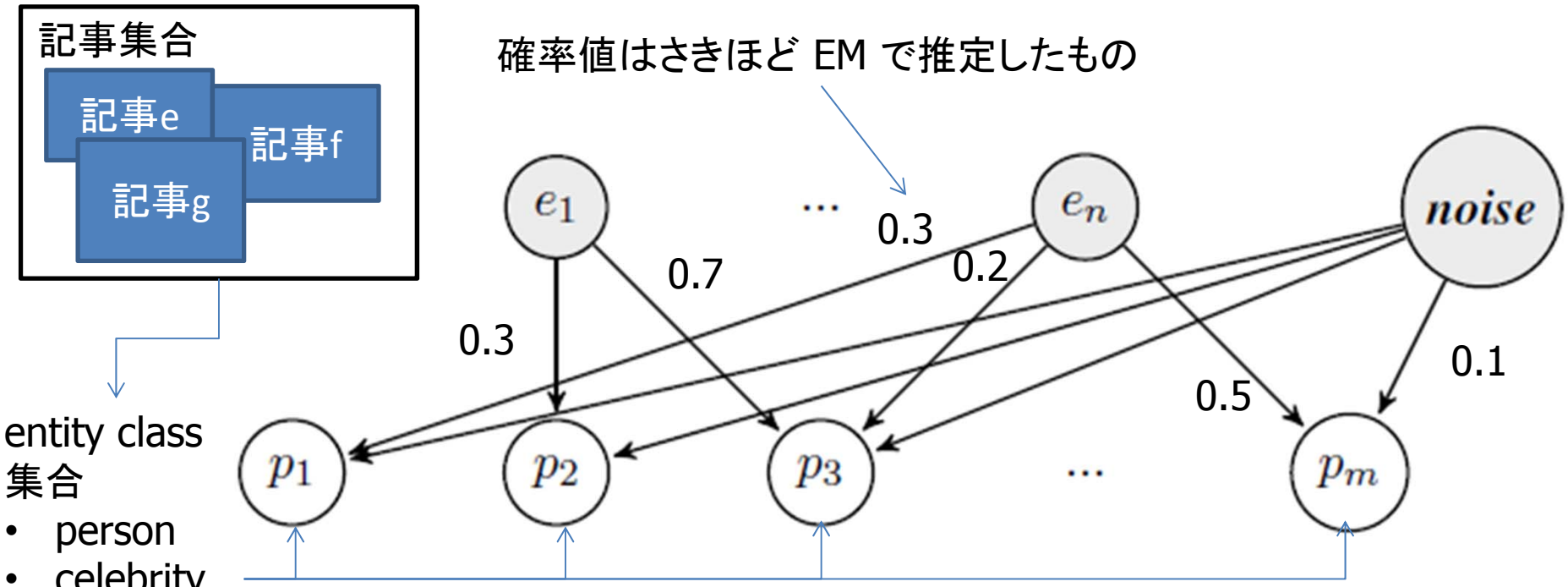
イベント e が未知のパタンから活性化させられる確率

パタン p がイベント e を活性化させる確率

こいつらを EM で推定

Inference （実際にヘッドラインを生成）

- 記事集合から entity 集合を取得し，それらと対応するパタン，それらのパタンと対応するイベントを特定
- ノード間の遷移確率がわかっているので，ランダムウォークさせると， pm の確率が出る



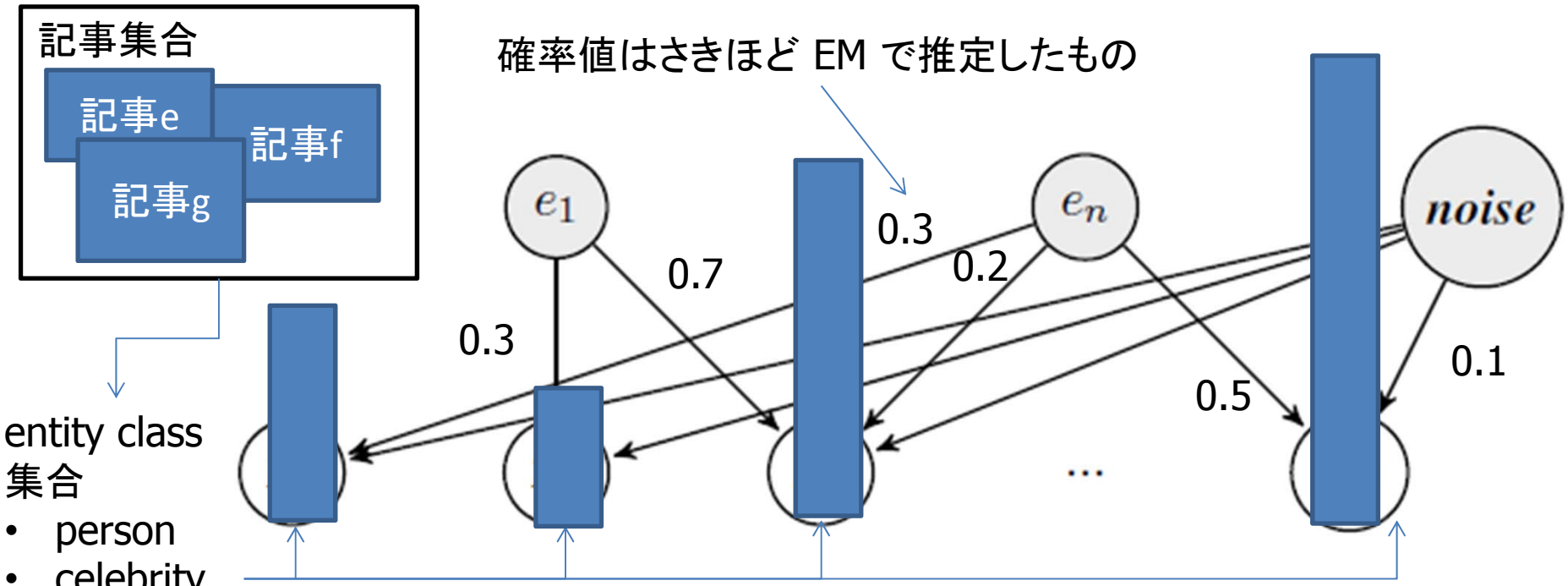
entity class 集合
- person
- celebrity
- actress
- location

etc

Alfonseca et al., HEADY: News headline abstraction through event pattern clustering, Proc. of ACL, pp. 1243—1253, 2013.
Figure 2 より引用

Inference （実際にヘッドラインを生成）

- 記事集合から entity 集合を取得し，それらと対応するパタン，それらのパタンと対応するイベントを特定
- ノード間の遷移確率がわかっているので，ランダムウォークさせると， pm の確率が出る



entity class 集合
- person
- celebrity
- actress
- location

etc

Alfonseca et al., HEADY: News headline abstraction through event pattern clustering, Proc. of ACL, pp. 1243—1253, 2013. Figure 2 より引用

評価

- ROUGE (人手で作ったヘッドラインとの類似度を比較)
  - HEADY：提案手法
  - Most frequent pattern:　入力文書集合中で一番よくでてきたパタン
  - TopicSum: 入力文書集合に含まれる人手によるヘッドラインのうち，よさそうなやつ（文書集合全体の中心にありそうなやつ）を同定，それを使う
  - MSC: Fillipova (2010) によるヘッドライン生成方法
  - Most f[illegible]番よくでてきたや
  - latest

	R-1	R-2	R-SU4
HEADY	0.3565	0.1903	0.1966
Most frequent pattern	0.3560	0.1864	0.1959
TopicSum	0.3594	0.1821	0.1935
MSC	0.3470	0.1765	0.1855
Most frequent headline	0.3177	0.1401	0.1668
Latest headline	0.2814	0.1191	0.1425

Alfonseca et al., HEADY: News headline abstraction through event pattern clustering, Proc. of ACL, pp. 1243—1253, 2013. Table 2 より引用

評価

- AMT で readability と informativeness を個別に評価
- 元々ついているヘッドラインを，TopicSum で選んだやつに負ける……
  - Readability になると自動生成は厳しい

	Readability	Informativeness
TopicSum	**4.86**	**4.63**
Most freq. headline	†‡4.61	†‡◇4.43
Latest headline	†‡4.55	† 4.00
HEADY	† 4.28	† 3.75
Most freq. pattern	† 3.95	† 3.82
MSC	3.00	3.05

自動生成

Alfonseca et al., HEADY: News headline abstraction through event pattern clustering, Proc. of ACL, pp. 1243—1253, 2013. Table 3 より引用

まとめと印象

- パタンを使ってヘッドライン生成
- 大規模なテキスト集合から潜在変数を加味してパタンを取得
- 人手で書いたやつからいいやつを選んだ手法には負けてしまう
- 名寄せは大切

まとめ

- 昨年度の Coling と同様，過去最大の投稿数となり，本分野の隆盛を痛感
- CL/NLP タスクの多様化と温故知新
  - マルチモーダル，論理形式， OCR ，暗号解読，プランニング
- テキストを媒体としたオープンなコミュニケーション（ソーシャルメディア）が発達し，それらを分析するためのより頑健な解析，および利用が求められている
- 技術的にはブレイクスルーはないが， DNN が徐々に CL/NLP 分野にも出現