『現代日本語書き言葉均衡コーパス』に対する係り受け付与

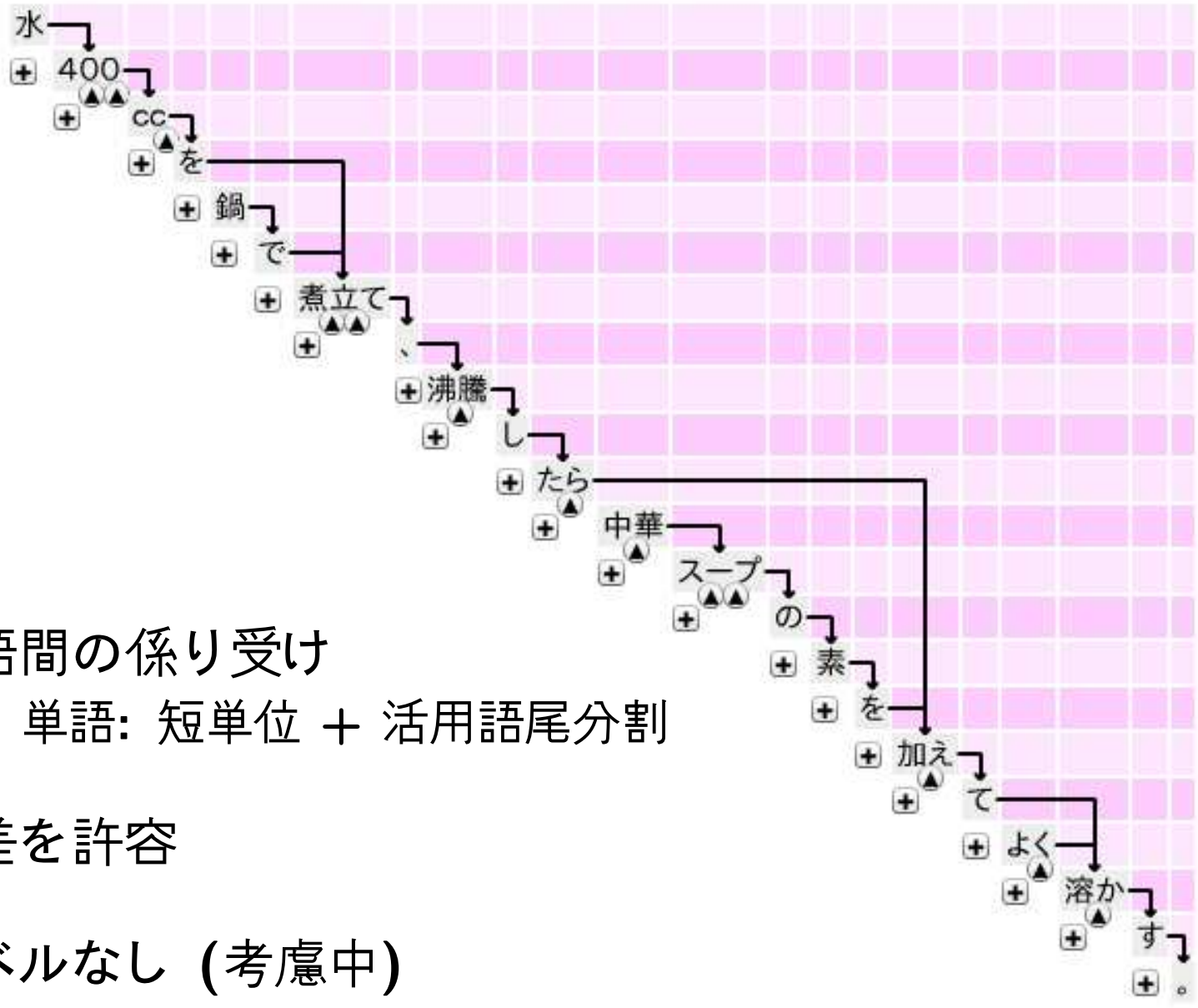
森　信介　　小椋　秀樹

2013 年 3 月 13 日

統語的構造の記述

- 言語理解に向かうステップ
  1. 単語分割 (品詞推定)
  2. (固有表現認識)
  3. 係り受け解析
  4. 述語項構造解析
  5. ...
- 様々な分野のテキスト
  - 『現代日本語書き言葉均衡コーパス』[前川 09]
    - 新聞, 雑誌, 書籍, 白書, ブログ, Web QA
  - 特許, レシピ, 論文抄録
  - 医療, 教科書, etc.

係り受け構造



- 単語間の係り受け
  - 単語: 短単位 + 活用語尾分割
- 交差を許容
- ラベルなし (考慮中)

アノテーション基準

- 主辞に係るとする
- 複数の妥当な候補があれば前側に係るとする
- アノテーションの過程で作業者が判断に迷う点などを中心に基準を整備

(1) 複合語: 語構造を判断 **(cf. BCCWJ の短単位規程)**

例) (交通→ (バリア → フリー)) → 法

(2) 複合語に係る連体修飾の係り先

- 全体に係る場合

  例) 総合 → 的 → な → (バリア → フリー → 化)

- 一部に係る場合 **(単語単位の長所)**

  例) 項 → 構造 → の → (曖昧 → 性) → 解消

アノテーション基準

(3) 括弧の対応

a 開き括弧は閉じ括弧に係る

例) [→ (1 →])

例) 「 → (京都 → 駅→ まで→ 」)

b 括弧が付された注釈的要素の扱い

例) 国際→ (原子 → 力 → 機関) → (（ → (I A E A →）) → の→ 調査

(4) 並列

a 最後の要素に並列マーカーがない場合

例) これ → と → あれ → を

例) 衆議 → 院 → と → (参議 → 院) → が

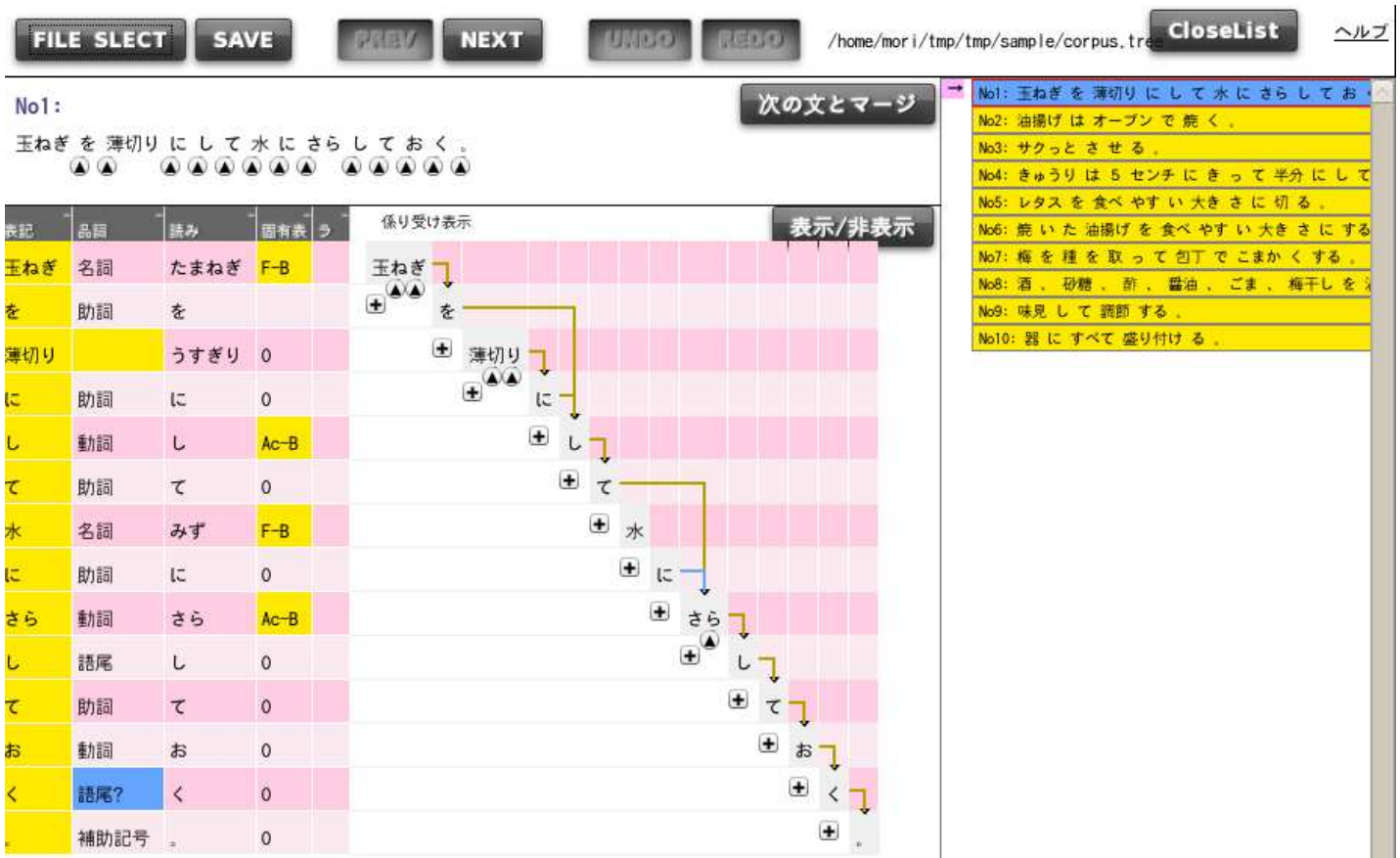
b 最後の要素に並列マーカーがあればそれに対応付ける

例) これ → と → (あれ → と) → を

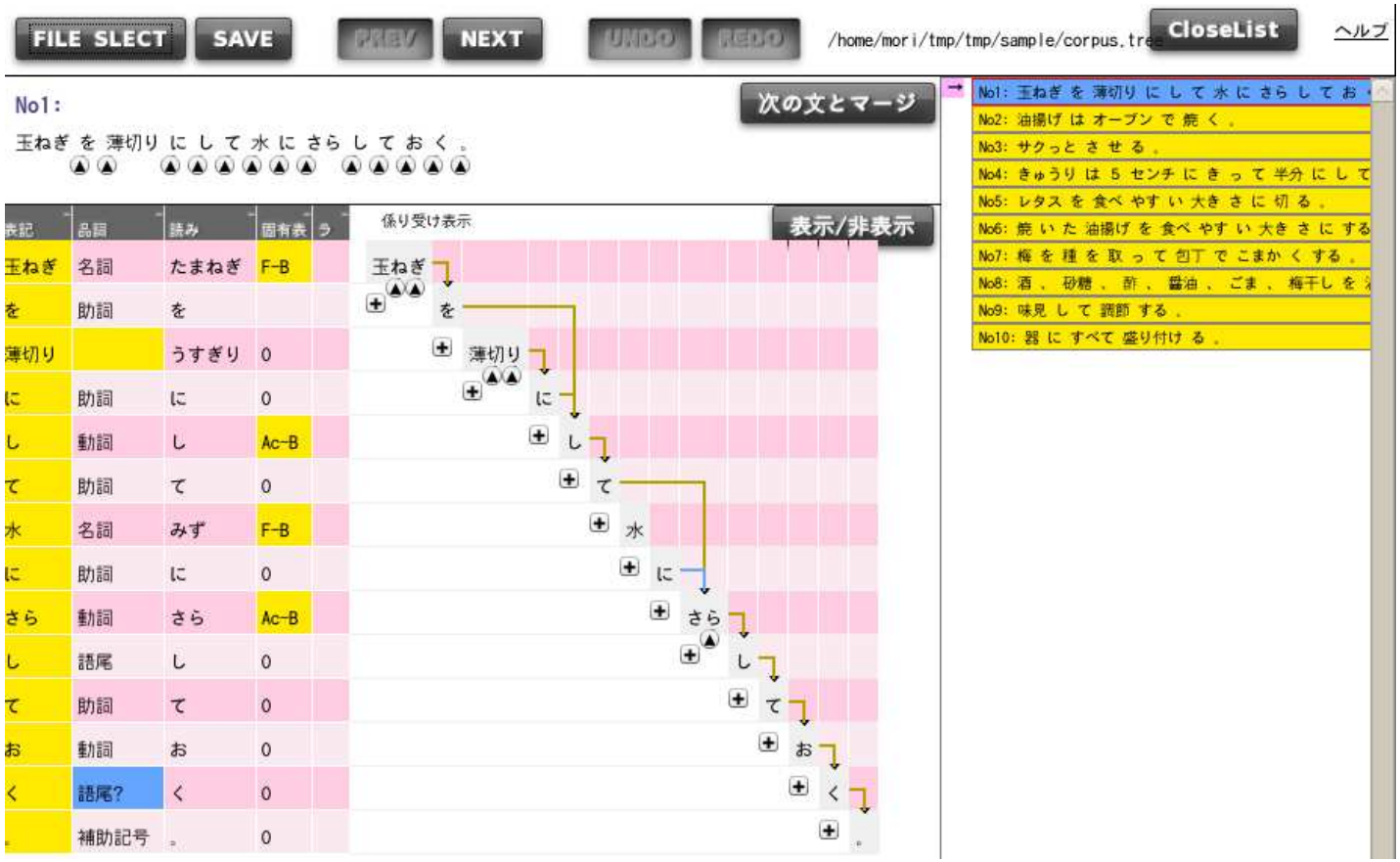
- その他多数の細かい問題

PNAT: Pointwise NLP Annotation Tool

- 単語分割
- 品詞
- 発音
- 固有表現タグ
- 係り受け



PNAT: Pointwise NLP Annotation Tool



- チェックすべき箇所 (能動学習などのため)
- 辞書にあるか否か
- 何らかの規則違反

係り受け解析 (EDA, 点予測) [Flannery 12]

- 係り受けの交差を許容
- 部分的アノテーションから学習可能
  - 文中の一部の単語にのみ係り先を付与
- 条件付き探索 (固有表現の保持)
- 配布中
  `http://plata.ar.media.kyoto-u.ac.jp/tool/`
  - 解析モジュール
  - 様々な分野に対応したモデル
  - 学習モジュール

点予測による係り受け解析

- 点予測による最大全域木 **(EDA) [Flannery 12]**
  1. 全ての単語間の係り受けスコアを計算

$$\sigma(\langle \mathbf{i}, \mathbf{d_i} \rangle, \vec{\mathbf{w}}), \qquad \text{ここで } \mathbf{w_i} \text{ は } \mathbf{w_{d_i}} \text{ に係る}$$

  2. エッジスコアの合計が最大になる全域木 **(MST)** を選択

$$\hat{\vec{\mathbf{d}}} = \underset{\vec{\mathbf{d}} \in D}{\mathbf{argmax}} \sum_{\mathbf{i}=1}^{\mathbf{n}} \sigma(\langle \mathbf{i}, \mathbf{d_i} \rangle, \vec{\mathbf{w}})$$

部分的アノテーションコーパスから学習可能
⇒ 柔軟なコーパス作成!
⇒ 迅速・安価な分野適応!

点予測による係り受け解析 (つづき)

- スコア計算の素性

			牡蠣	を	広島	に	食べ	に		行	く		
w_{i-3}	w_{i-2}	w_{i-1}	w_i		w_{i+1}	w_{i+2}	w_{i+3}						
				w_{d_i-3}	w_{d_i-2}	w_{d_i-1}	w_{d_i}	w_{d_i+1}	w_{d_i+2}	w_{d_i+3}			

F1 係り元 w_i と係り先 w_{d_i} の距離
F2 w_i と w_{d_i} の表記
F3 w_i と w_{d_i} の品詞
F4 w_i と w_{d_i} の前後 3 単語の表記
F5 w_i と w_{d_i} の前後 3 単語の品詞

作業状況 (フルアノテーション)

- 全ての単語に係り先を付与

ID	文数	単語数	文字数	備考
BCCWJ-Core (1/10)				
PN	1,713	35,802	52,798	新聞
PM	1,505	23,202	36,981	雑誌
PB	1,058	22,022	30,879	書籍
OC	615	11,819	16,414	Yahoo!知恵袋
OY	857	12,113	17,957	Yahoo!ブログ
OW	658	25,563	37,684	白書
EHJ	13,000	164,376	220,146	英語表現辞典
NKN	10,025	292,462	442,264	日経新聞記事
RCP	724	12,403	19,182	レシピ
NPT	500	20,145	31,631	NTCIR 特許翻訳
JNL	354	12,963	21,710	論文抄録

作業状況 (部分的アノテーション)

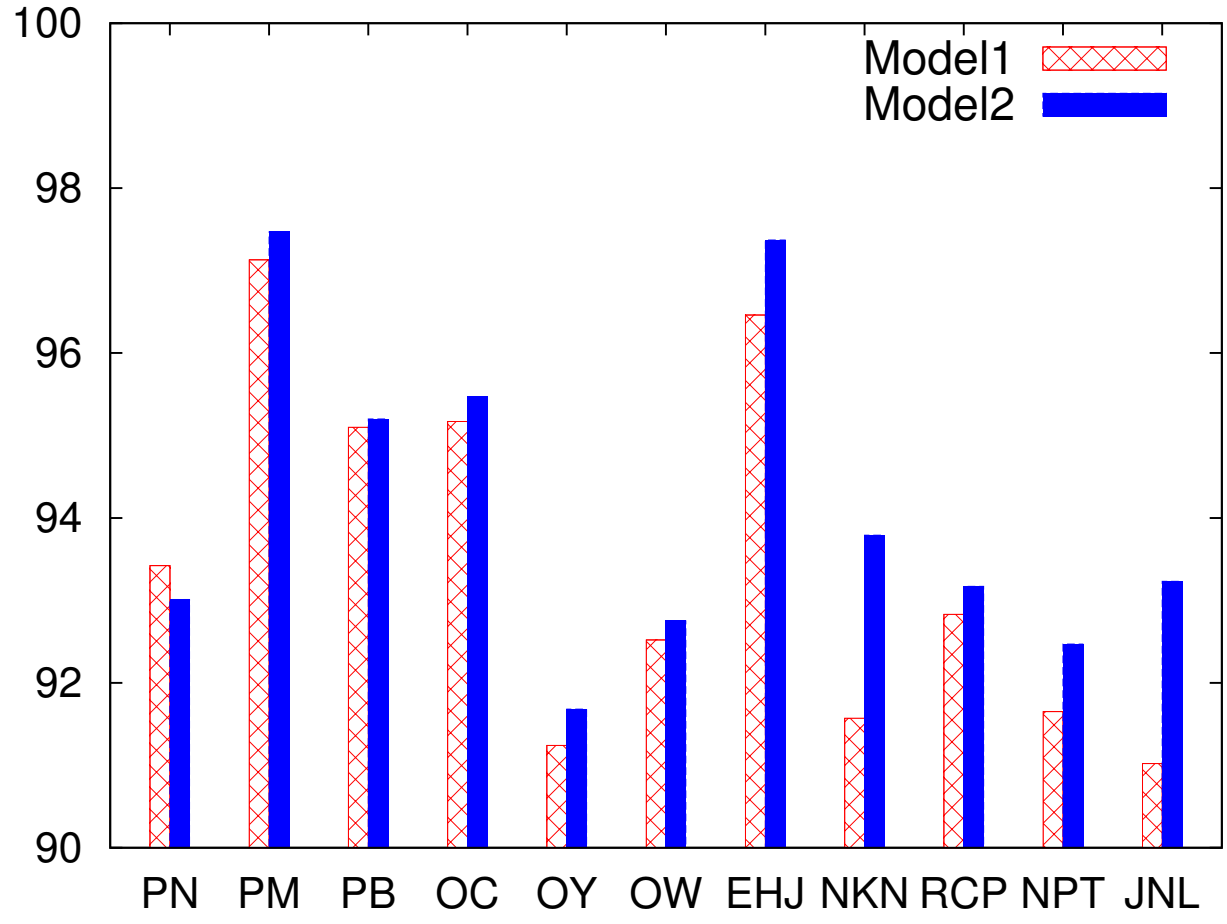
- 一部の単語にのみ係り先を付与
  - 効率的 (分野特有の語に集中，能動学習)
  - 主に学習用
    - **EDA** は学習に利用可能

ID	係り受け数	備考
EDR	**550,823**	**EDR** コーパスからの自動変換
PB	**doing**	**BCCWJ** 書籍，能動学習
RCP	**todo**	レシピ (2013/Apr.- ?)
NPT	**todo**	特許翻訳 (2013/Apr.- ?)
JNL	**todo**	論文抄録

係り受け解析実験

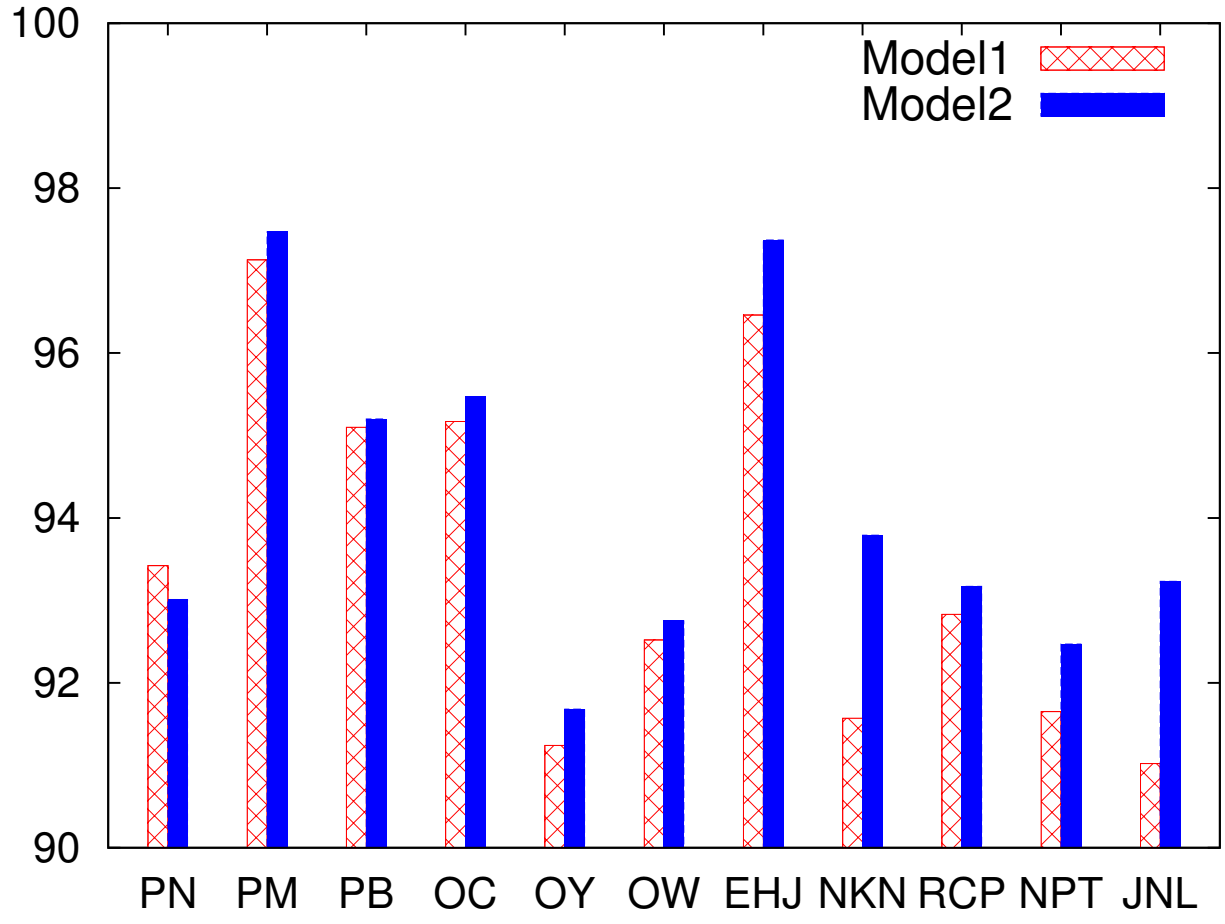
- テスト コーパス
  - **BCCWJ-Core**
    - **PN, PM, PB, OC, OY, OW (各 200 文)**
  - 英語表現辞典 **(1/10)**, 日経新聞記事 **(1/10)**
  - レシピ, 特許, 論文抄録
- 学習コーパス
  1. **Model 1: BCCWJ-Core** テスト 以外
  2. **Model 2:** テスト 以外全部

各分野の解析精度



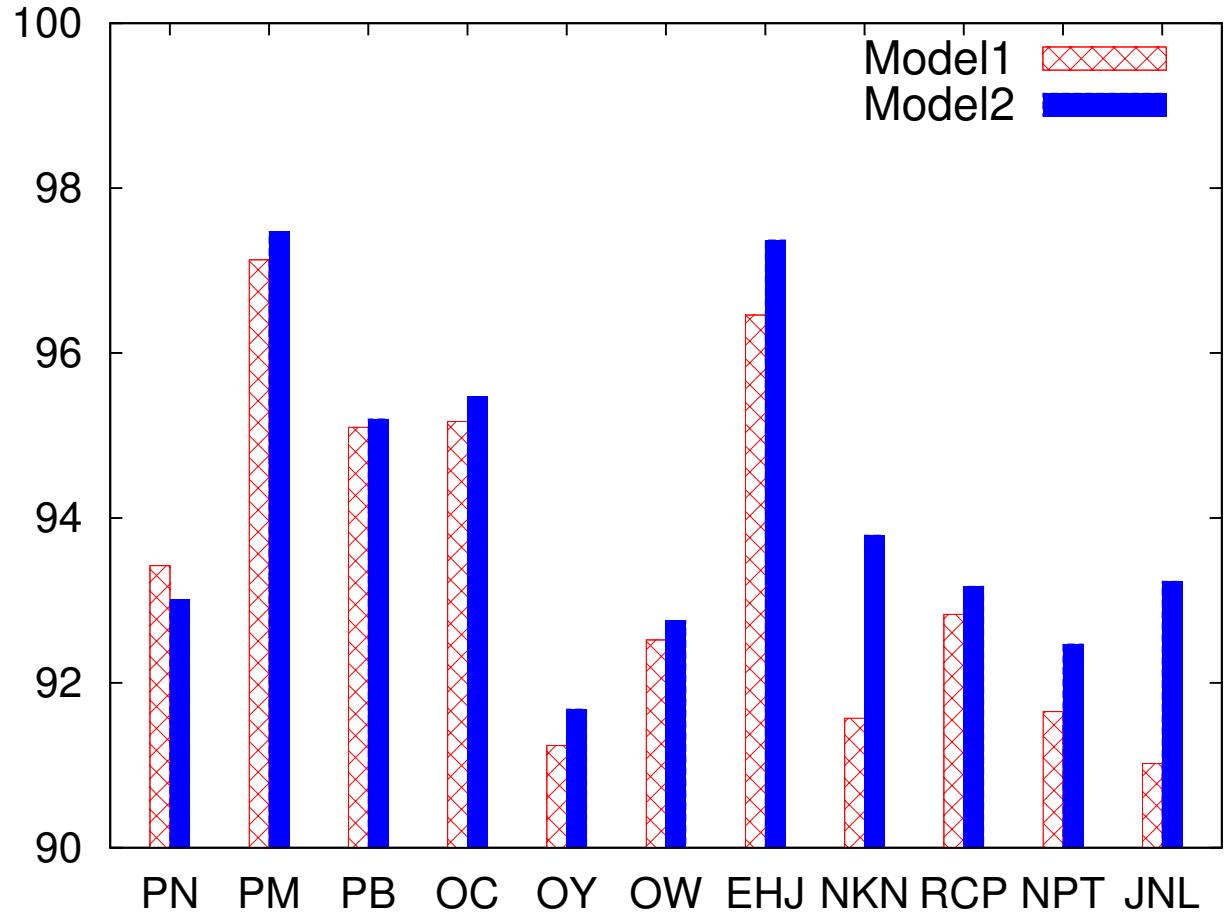
- **Model 1**
  - 雑誌 **(PM)** と辞書の例文 **(EHJ)** の精度が高い
  - ブログ **(OY)** と論文抄録 **(JNL)** の精度が低い

各分野の解析精度



- **Model 2**
  - 雑誌 **(PM)** と辞書の例文 **(EHJ)** は **98%**が見えてきた
  - ブログ **(OY)** と特許 **(NPT)** の精度が低い

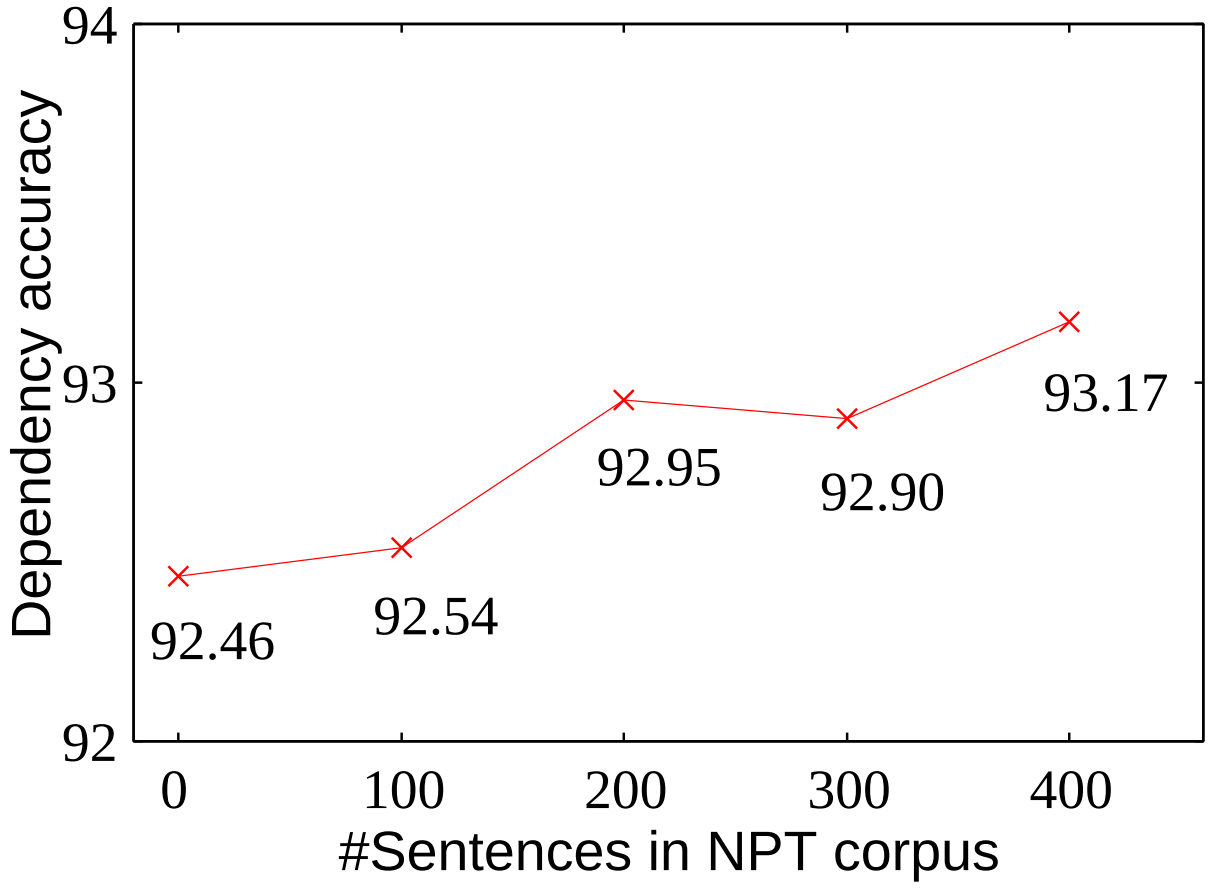
各分野の解析精度



- **Model 1 ⇒ Model 2**
  - 学習コーパスの増量により精度向上 **(PN は例外)**
  - 学習コーパスが追加された日経 **(NKN)** の精度向上幅大

特許への分野適応 (Cf. [森 12])

- テスト：特許文 **NPT** の **100** 文
- 学習: **Model 2 + NPT** の **0, 100, 200, 300, 400** 文



- 徐々に精度向上

おわりに

- 単語単位の係り受けコーパス (タグは配布可)
  - 『現代日本語書き言葉均衡コーパス』コアデータ
  - 英語表現辞典，日経新聞記事
  - レシピ，特許，論文抄録
- 精度評価
- 今後
  - レシピ，特許，**etc.** の増強
  - 教科書**?**
  - 部分的アノテーション

References

- **Flannery, D., Miyao, Y., Neubig, G., and Mori, S.: A Pointwise Approach to Training Dependency Parsers from Partially Annotated Corpora, Journal of Natural Language Processing, Vol. 19, No. 3 (2012)**
- 森 信介：自然言語処理における分野適応, 人工知能学会誌, **Vol. 27, No. 4 (2012)**
- 前川 喜久雄：代表性を有する大規模日本語書き言葉コーパスの構築, 人工知能学会誌, **Vol. 24, No. 5, pp. 616–622 (2009)**