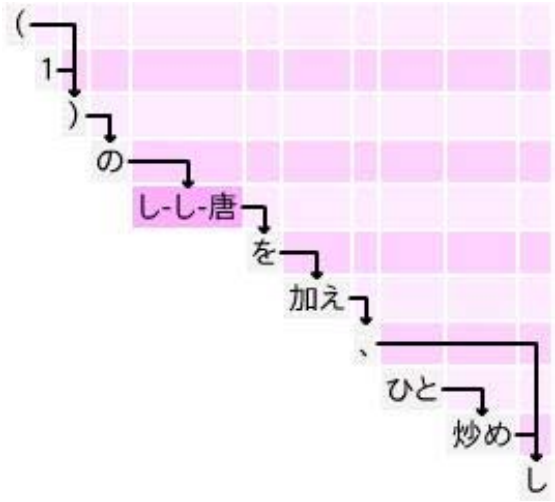
自然言語処理のレシピへの分野適応

森 信介 zelch

京都大学

2011 年 5 月 25 日



統計的手法による言語処理

- 現在主流の手法 (ChaSen, MeCab, ...)
  1. コーパスへの情報(アノテーション)付与
  2. 機械学習
  3. タスクのテキストの解析
- 問題点
  - タスクのテキスト (適応分野) に対する不十分な精度

    理由: 学習コーパス (一般分野) との差異
  - 自然言語処理の枠組みに制限されたアノテーション
    - アノテーション体系の理解困難 (ex. 活用型って何)
    - 目的には不要なアノテーション (ex. 品詞細分類)

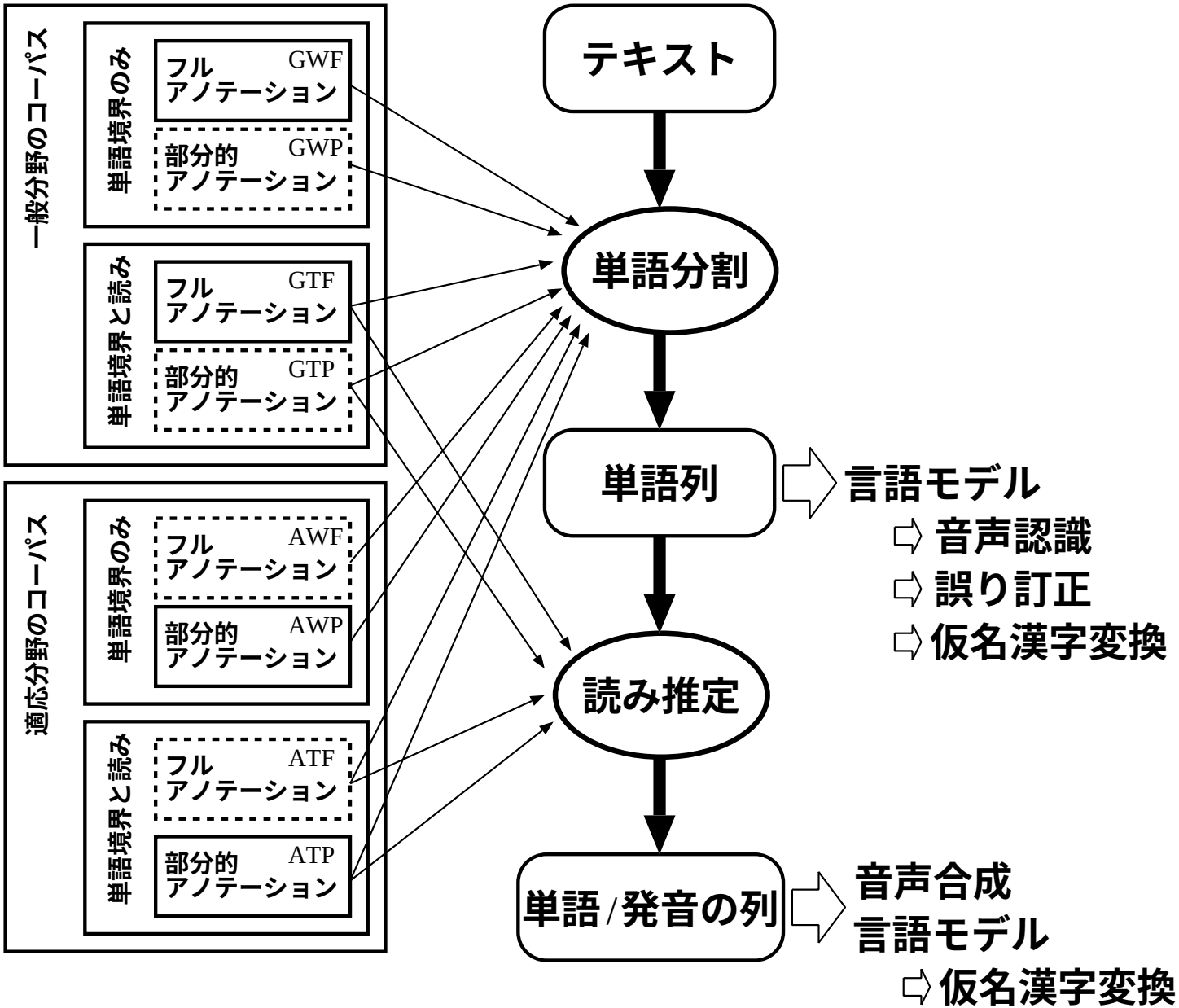
  $\Rightarrow$ 辞書追加程度

Pointwise NLP
点予測による言語処理

- 必要なときに必要なだけの曖昧性解消
- 部分的アノテーションによる迅速(Agile Adaptation)分野適応
- 推定値を参照しない

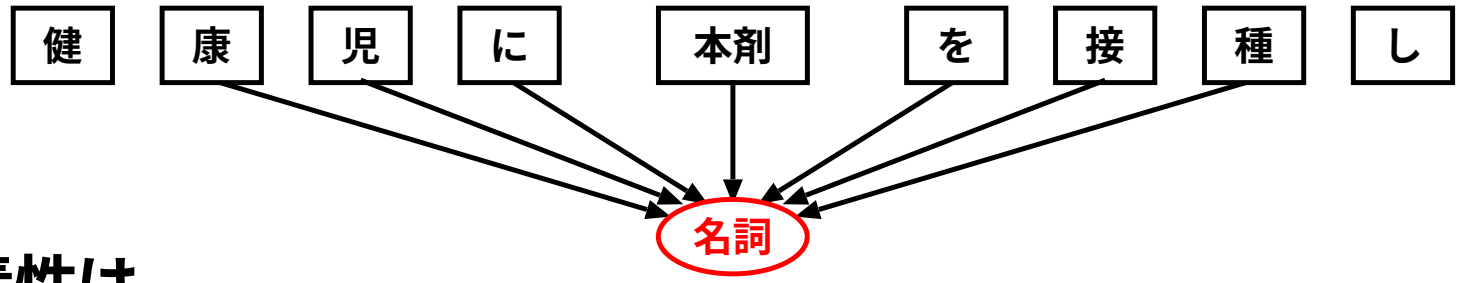
1. 単語分割 [LREC 2010, 情処論?]
2. 品詞推定 [ACL 2011, 情処 NL198] きゅーてぃー KyTea
3. 読み推定 [LREC 2010]
4. 係り受け解析 [情処 NL201, IJCNLP 2011?] 未公開
5. 述語項構造 実装中
   ex.) 上げ (かき-を, ざる-に)

必要なときに必要なだけの曖昧性解消



推定値を参照しない

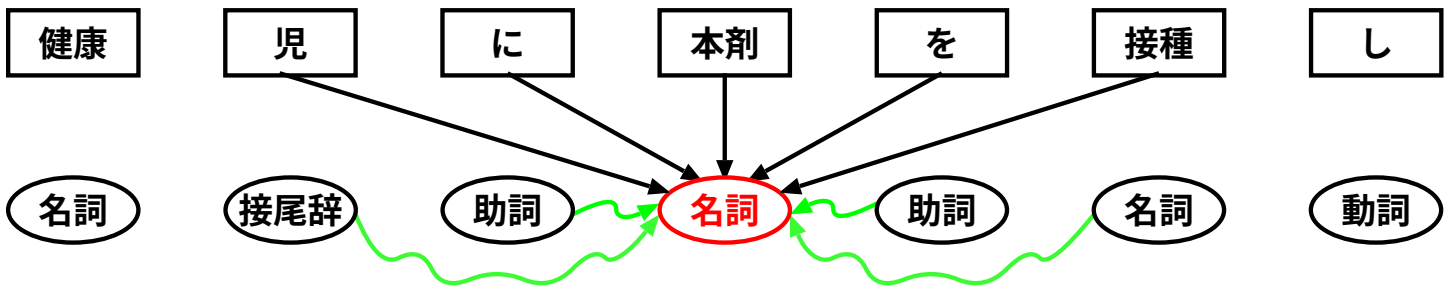
- 点予測による形態素解析



素性は

1. 注目単語
2. その単語境界
3. 前後の文字列

- Cf. 系列に基づく形態素解析 (CRF, n-gram モデル)



実験結果

- 言語資源: 国研のBCCWJ（コーパス） + UniDic(213,174 語)

出典	用途	文数	文字数
白書・書籍・新聞 (一般分野)	学習	27,338	1,131,317
	テスト	3,038	126,154
医薬品情報 (JAPIC) (適応分野)	テスト	1,236	67,828

- 結果

形態素解析手法	形態素解析精度	
	一般分野	適応分野
系列予測 (MeCab)	99.23	92.94
点予測 (KyTea)	98.86	93.70

部分的アノテーションによる迅速分野適応 Agile Adaptation

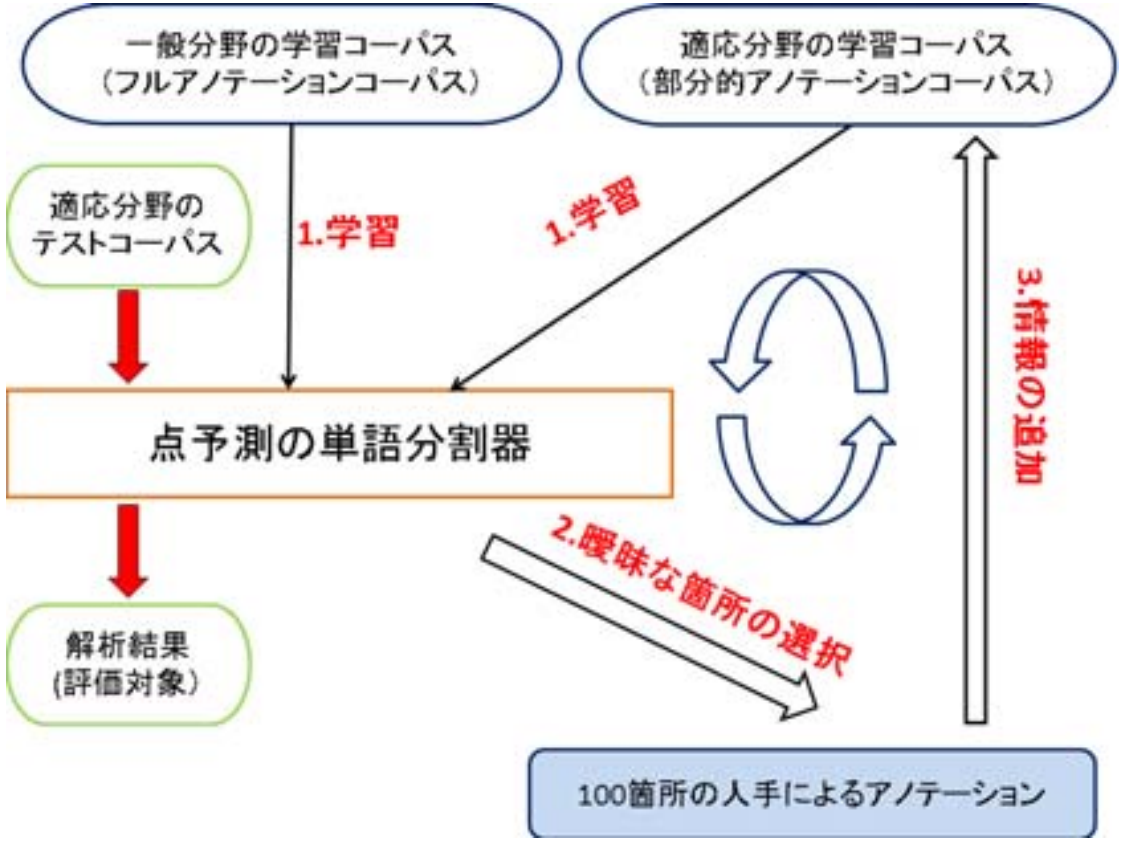
- 部分的アノテーションコーパスが利用可能
  - 分野特有の表現のみ情報付与

    例) （１）の ししとう/名詞 を加え、ひと炒めし

  - 注目単語の単語境界と品詞のみ
- 能動学習の活用
- Cf. フルアノテーションコーパス
  - 文のすべての単語境界と品詞

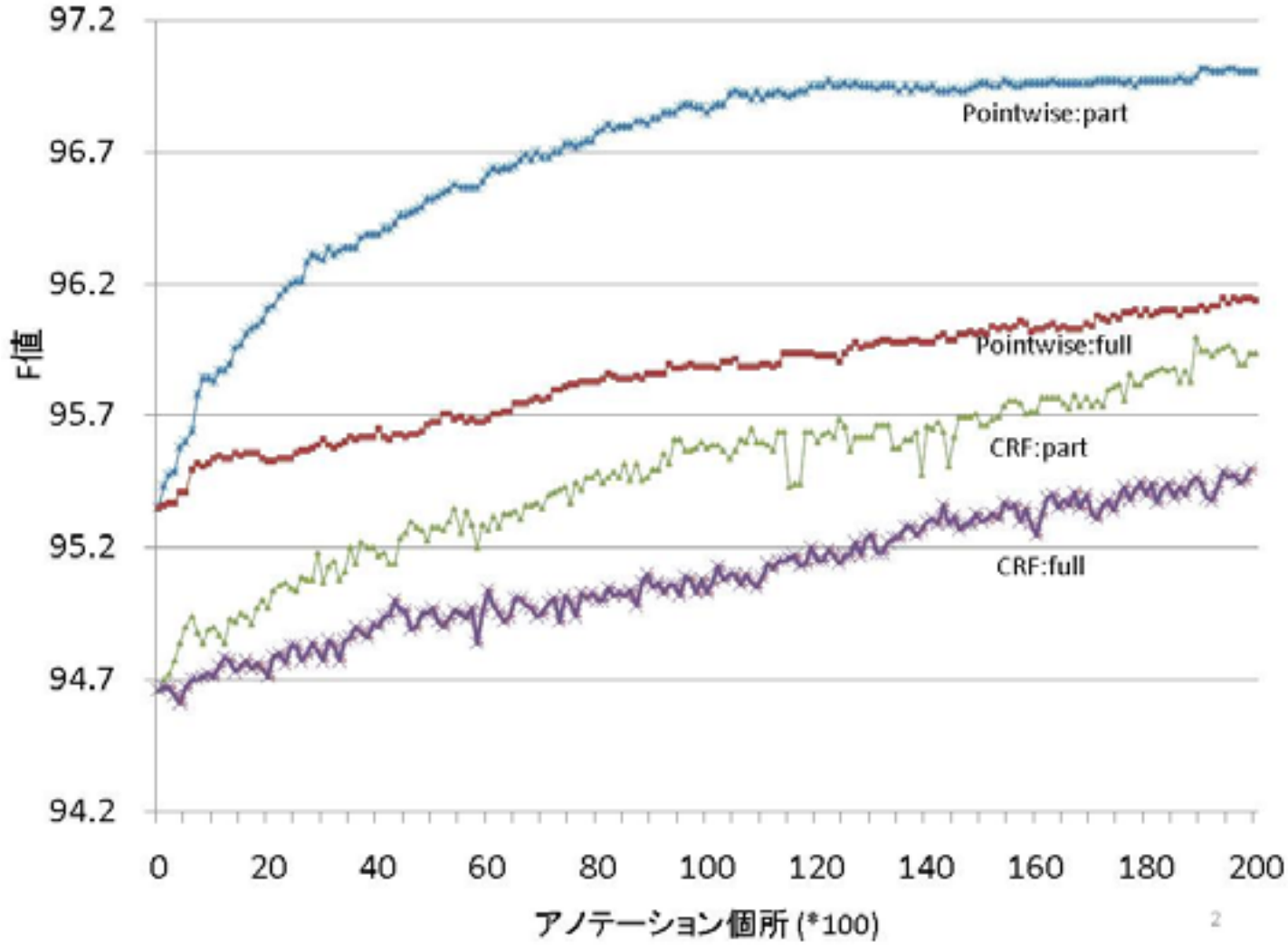
    例) （/記号 １/数字 ）/記号 の/助詞 ししとう/名詞
    を/助詞 加え/動詞 ひと/接頭辞 炒め/動詞 し/動詞

能動学習 (Active Learning)



- **東日本大震災の tweet の単語分割 (ANPI_NLP)**
  - **2 時間弱の作業 (4 iterations)**
  - **精度:** $97.31\% \overset{+\text{dict}}{\longrightarrow} 97.32\% \overset{+\text{AL}}{\longrightarrow} 97.74\%$
  - **もう少し時間をかければ 98.5%程度に到達**

能動学習の結果



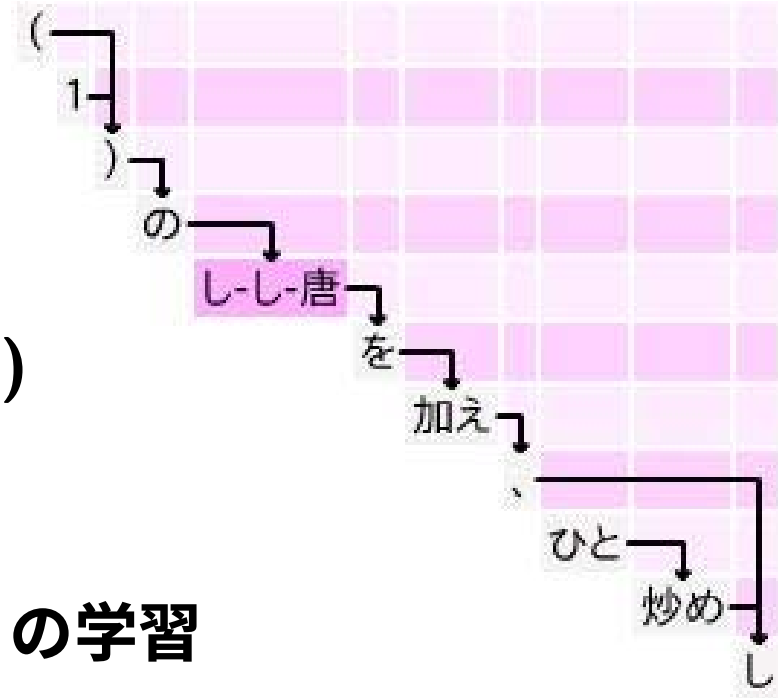
- 「点予測＋能動学習」は効果的

KyTea

係り受け解析 [情処 NL201]

係り受け木とは

1. **単語**をノードとし
2. 修飾関係をアークとする
   (いまのところラベルなし)
3. 木構造



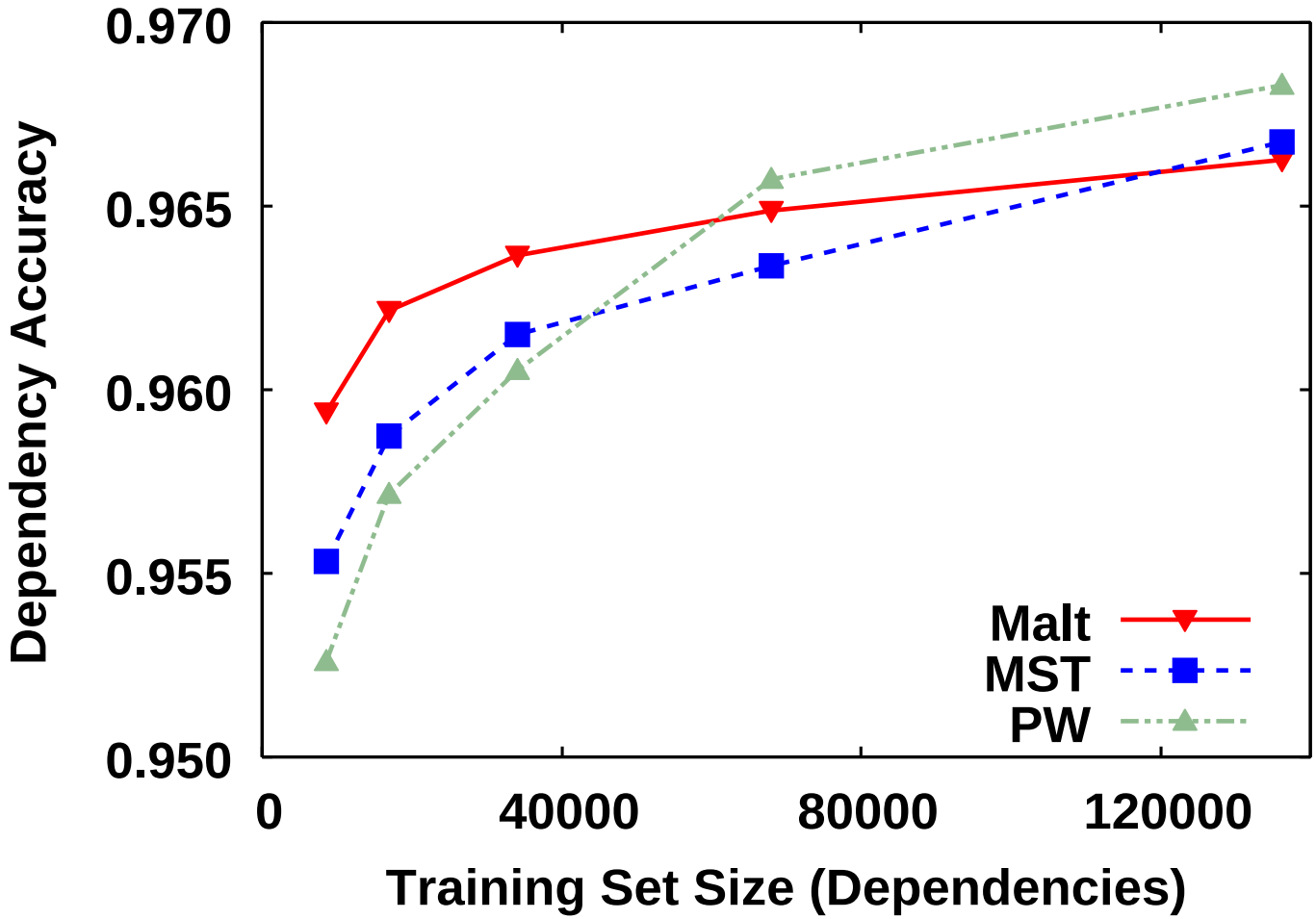
- 部分的アノテーションからの学習
- Pointwise 点予測の MST (Maximum Spanning Tree)
  - まもなく公開
  - レシピへの適応はまだ

コーパスの諸元

ID	出典	用途	文数	単語数	文字数
EHJ-train	辞書の	学習	11,700	145,925	197,941
EHJ-test	例文	テスト	1,300	16,348	22,207

- 辞書の例文
- 品詞なし (KyTea で自動付与)
- 比較手法
  - Malt: Nivre ほか (2006) の MaltParser (≈ CaboCha)
  - MST: McDonald ほか (2005) の MST Parser
  - PW: 提案手法

フルアノテーションコーパスサイズ v.s. 精度



- コーパスが大きくなると PW の精度が高くなる (11.47[係り受け/文])

述語項構造 (作成中)

- 述語とその項の同定

例文) かき は 塩水 で 洗 い 、 ざる に 上げ て
水気 を き る 。

⇓

上げ (調理者-が, かき-を, ざる-に)

- 手法
  1. 係り受け解析の結果を利用 (一般的)
  2. 単語列から直接 (考案中)
- 一般分野でも精度が低い
- レシピでの結果はない (?) [柴田ほか 情処論 2008, など]

レシピテキスト処理の課題

- 形態素解析の精度向上
  - 10 時間程度の作業 (すでに 1 時間作業済み)
- 係り受け解析の精度向上
  - 50 時間程度の作業
- 固有表現 (特定意味クラスの単語列) の抽出
  - 食材，調理器具，...
  - 様々な言語資源を活用
- 表記ゆれ
  - KyTea は読み推定ができるので容易??
  - JUMAN の代表表記の利用??

まとめ

- ◎ レシピに強い形態素解析
- ○ レシピに強い係り受け解析
- △ 表記ゆれ
- △ 固有表現
- × 述語項構造 (ゼロ代名詞など)

- 他にできること
  - 音声認識，仮名漢字変換，読み推定 (音声合成)
- 皆さんの要望??

mori.shinsuke.8u@kyoto-u.ac.jp or zelch 検索