点予測による自然言語処理

@zzzelch

2011 年 11 月 23 日

Outline

形態素解析とは

1. **文をある形態素に分割**

   例) 京都 に 行っ た

2. **それぞれ形態素の品詞を推定**

   例) 京都/名詞:地名 に/助詞:格助詞 行っ:動詞:カ変 た/助動詞:断定

3. **活用語の原形を推定**

   例) 行っ/動詞 → 行く

形態素解析の歴史

第 1 世代: ルールベース

- 精度向上手段(ユーザーによる)は辞書追加のみ

例) JUMAN

第 2 世代: 統計・機械学習

- 学習データと学習手法に分離
- 配布版モデルの学習データは利用不可
- 精度向上手段(ユーザーによる)は辞書追加のみ

例) 茶筌, MeCab

第 3 世代: 言語資源(or ユーザー)中心主義

- 柔軟な言語資源の追加による効率的精度向上
- 配布版モデルの学習データを含めて再学習

例) KyTea(きゅーてぃー)

言語資源中心主義

- 言語資源中心主義 = ユーザー中心主義
  - 実利用(ビジネス)では様々なテキストを処理
  - 十分な精度の実現には**言語資源の追加は必須**
- **理想**: 学習データ(BCCWJ)と同一の分野に対する精度 (F 値)

手法	単語分割	形態素解析
CRF(MeCab-0.98)	99.57	99.23
点予測 (KyTea-0.1.1)	99.37	98.86

- **現実**: 医薬品情報に対する精度 (F 値)

手法	単語分割	形態素解析
CRF(MeCab-0.98)	94.69	92.94
点予測 (KyTea-0.1.1)	95.17	93.70

1. 現実的状況では解析精度が大幅下落!!
2. 現実的状況では KyTea $\gg$ MeCab

自然言語処理への要求

- 目的のテキストに対する高い精度
- ラベル付与 (品詞大分類，読み，...)
- 速度 (学習，初期化，解析)
  - KyTea には高速化のアイデアあり

自然言語処理

= 機械学習 + 情報付与済みテキスト

1. 10% 機械学習 (inspiration)

研究者は好んでやる → 放っておけば OK

	単語分割精度 (F 値)
品詞 3-gram モデル (HMM)	だめだめ
単語 3-gram モデル	99.07
系列予測 (CRF)	99.16
点予測 (SVM)	99.37

学習: BCCWJ + UniDic, テスト: BCCWJ

2. 90% 情報付与済みテキスト (perspiration)

ここをいかに効率化するか

形態素解析とは

1. **文をある形態素に分割**

   例) 京都 に 行っ た

2. **それぞれ形態素の品詞を推定**

   例) 京都/名詞:地名 に/助詞:格助詞 行っ:動詞:カ変 た/助動詞:断定

細かい品詞は作業不能

3. **活用語の原形を推定**

   例) 行っ/動詞 → 行く (cf. 行う)

作業不能 & 原形は不要

入力文の読みが欲しい

言語資源 (≠ KyTea)

1. **形態素の定義 (KyTea とは独立)**
   - 国立国語研究所の短単位 (+ 活用語尾の分割)
   - 約 20 品詞ほど
   - 文脈依存の読み

例) 将棋/名詞/しょうぎ の/助詞/の 本/名詞/ほん
を/助詞/を 買/動詞/か っ/助詞/っ て/て き/動詞/き
ま/語尾/ま し/助動詞/し た/助動詞/た 。/記号/。

2. **言語資源**
   - 現代日本語書き言葉均衡コーパス (BCCWJ)
   - UniDic
   - 部分的アノテーション
     - 医薬品情報, レシピ, 特許文書, etc.
   - 辞書 (単語のみ, 品詞あり, 読みあり)
     - 人名辞書, 野球関連辞書, etc.

Pointwise NLP
点予測による自然言語処理

- 推定値を参照しない
- 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分野適応

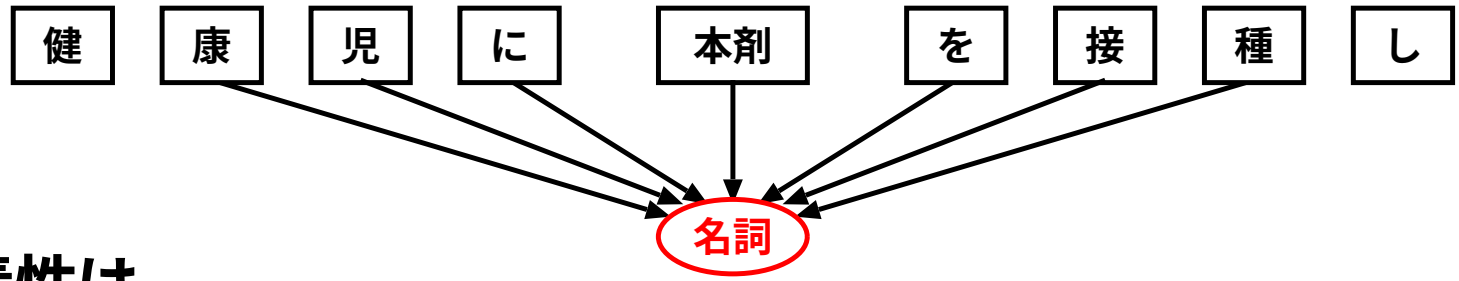
1. 単語分割 [LREC2010, 情処論 2011]
2. 品詞推定 [ACL2011, 自然言語処理 2011]
3. 読み推定 [LREC2010, InterSpeech2011]
4. 係り受け解析 [NL201, IJCNLP2011]

* Twitter の安否情報の解析 [IJCNLP2011] ANPI-NLP
   - 単語分割: 97.3% → 約 90 分の能動学習 →97.7%

推定値を参照しない

推定値は精度と同程度の信頼性

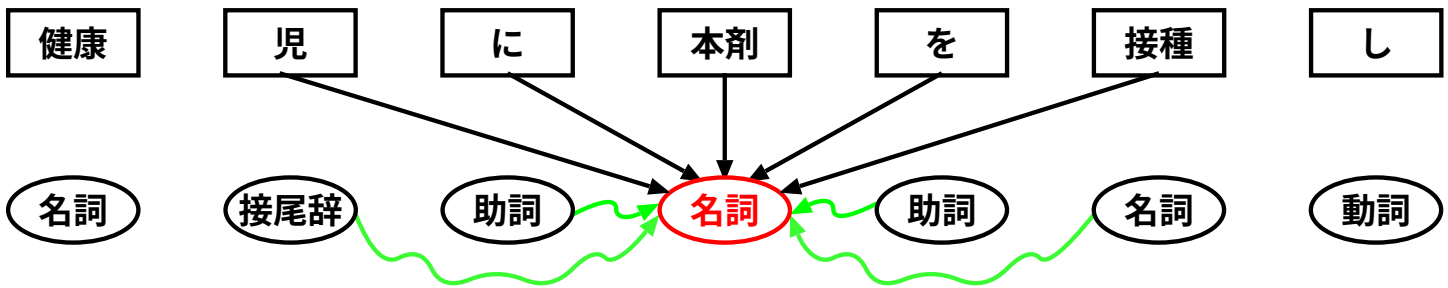
- 点予測による形態素解析



素性は

1. 注目単語
2. その単語境界
3. 前後の文字列

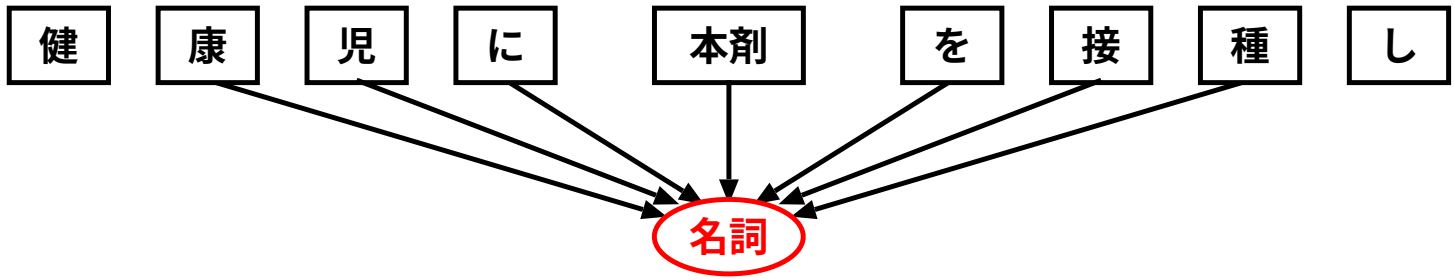
- Cf. 系列に基づく形態素解析 (CRF, n-gram モデル)



点予測 v.s. 系列予測

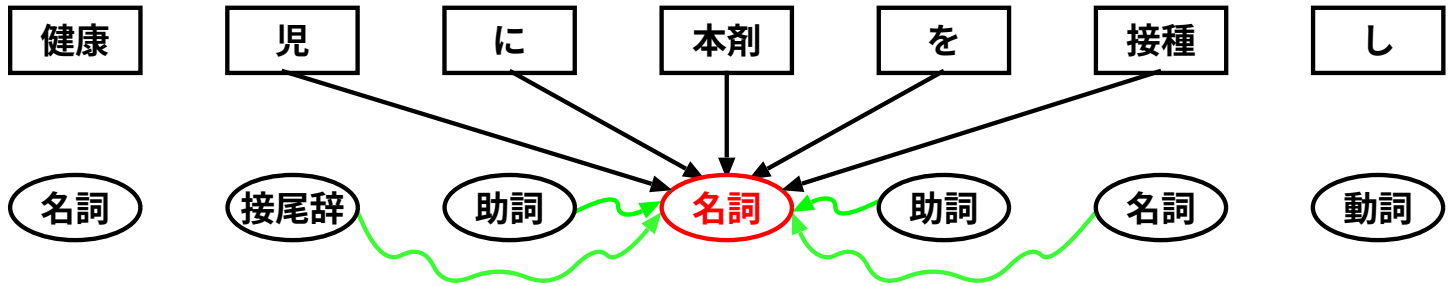
分類のための情報は入力のみ!!

- 点予測による自然言語処理



$$y_i = f(i, x_1 x_2 \cdots x_n) \quad 1 \leq i \leq m$$

- 系列予測に基づく自然言語処理



$$y_i = f(i, x_1 x_2 \cdots x_n, y_1 y_2 \cdots y_m)$$

※点予測と同じ素性ならより広い範囲を間接的に参照

部分的アノテーションによる迅速分野適応

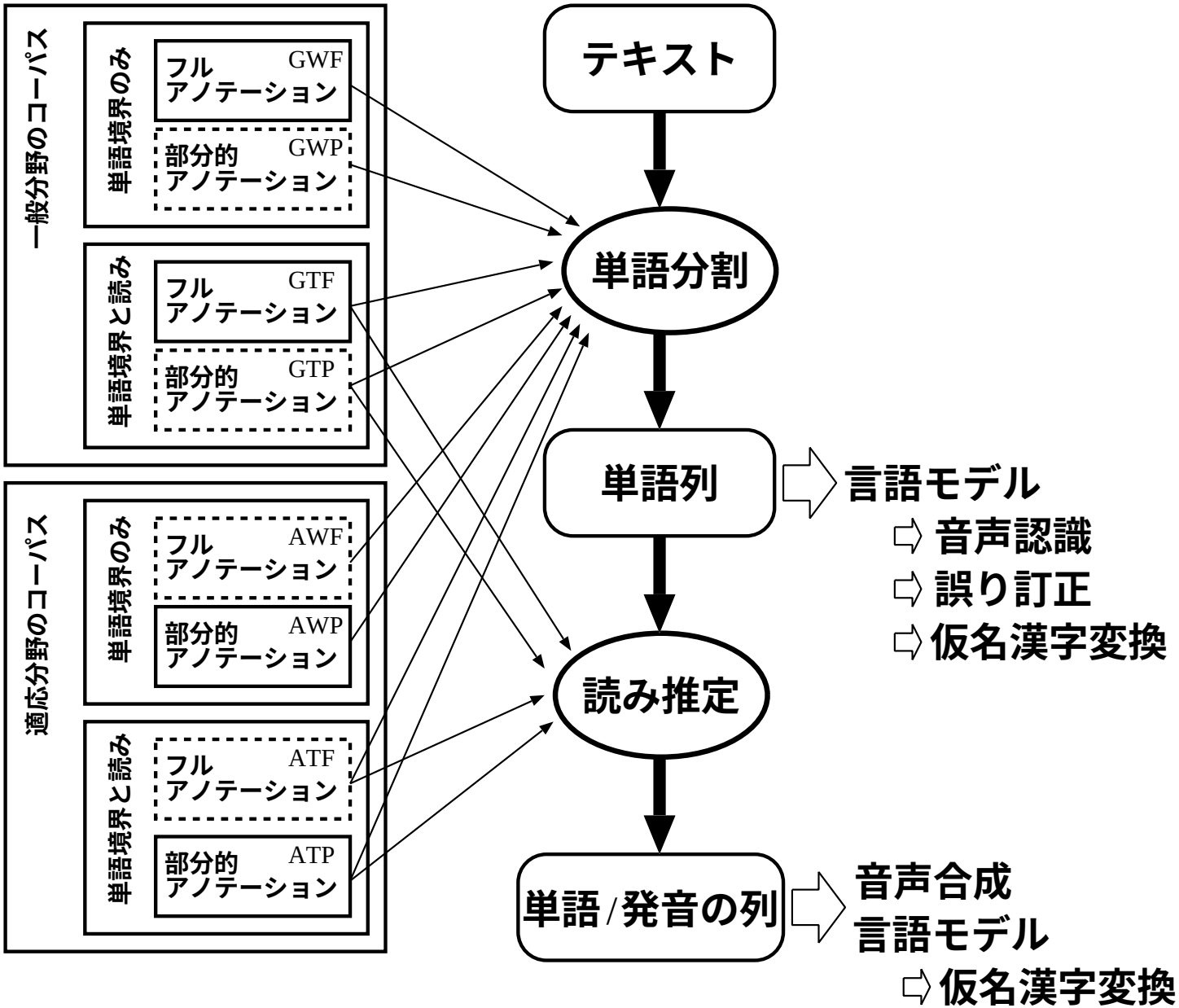
- **部分的アノテーションコーパス**が利用可能
  - **分野特有の表現**のみ情報付与

    例) ガーゼ等は **本剤/名詞** を吸着する

  - 注目単語の単語境界と品詞のみ
- 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) ガーゼ/名詞 等/接尾辞 は/助詞 本剤/名詞 を/助詞 吸着/名詞 する/動詞

必要なときに必要なだけの曖昧性解消



点予測による自然言語処理の歴史 (1)

2004/07 **確率的単語分割 [NL162, ICSLP2004]**

- **統計的仮名漢字変換の自動語彙拡張 [NL172, Coling-ACL2006, 情処論 2007]**
- **音声認識の自動語彙拡張 [ICASSP2007]**

2005/10 **ステーキ論文 [NL168, ICSLP2006]**

1. **リストに含まれる単語に対して**
2. **生コーパスにおける出現箇所に単語境界情報を付与**
   - **Raw: 生コーパスのまま (確率的単語分割)**
   - **Rare: リストの各単語に対し 2 箇所にのみ情報付与**
   - **Medium: リストの各単語に対し全箇所に情報付与**
   - **Well-done: 全文を人手で単語分割**
3. **言語モデルの作成**

※付与情報 (部分的アノテーション) から単語分割器を再推定すべき!!

とのつっこみが査読者から来るのではとビビってた

点予測による自然言語処理の歴史 (2)

2006/10 最大エントロピー (Maximum Entropy) 法による単語境界確率推定

2007/11 部分的アノテーションデータからの CRF の学習 [NL182, Coling2008] 学習速度が遅い (泣)

2009/07 PFI @hillbig さん et al. が LibLinear を推薦

2009/09 NL193/YANS のときに複数の学生に指令 (おねがい)

2009/11 KyTea 0.0.1 (単語分割&読み推定) by @neubig [LREC2010]

2010/10 KyTea 0.1.3 (品詞推定) [ACL2011, 情処論 2011]

2011/05 単語単位の係り受け解析 by @kansaidaniel [NL201, IJCNLP2011]

部分的アノテーション + 点予測

(v.s. フルアノテーション + 系列予測)

1. 言語処理システム　　速い!!
   - 実装が簡便
   - メンテナンス& 並列化が容易
   - 能動学習 に耐えるモデル学習速度
2. コーパス作成　　安い!?
   - 作業者の確保が容易
   - 能動学習によるアノテーション箇所の最少化
3. 解析精度　　旨い??
   - 一般分野での精度
   - 適応分野での精度

能動学習による精度向上

- 能動学習
  1. 曖昧な箇所をシステムが提示 (100 箇所)
  2. 人手で作業 (& 精度評価)
  3. goto 1.

- 初期状態 (再掲)
  - 学習データ (BCCWJ) と同一の分野に対する精度 (F 値)

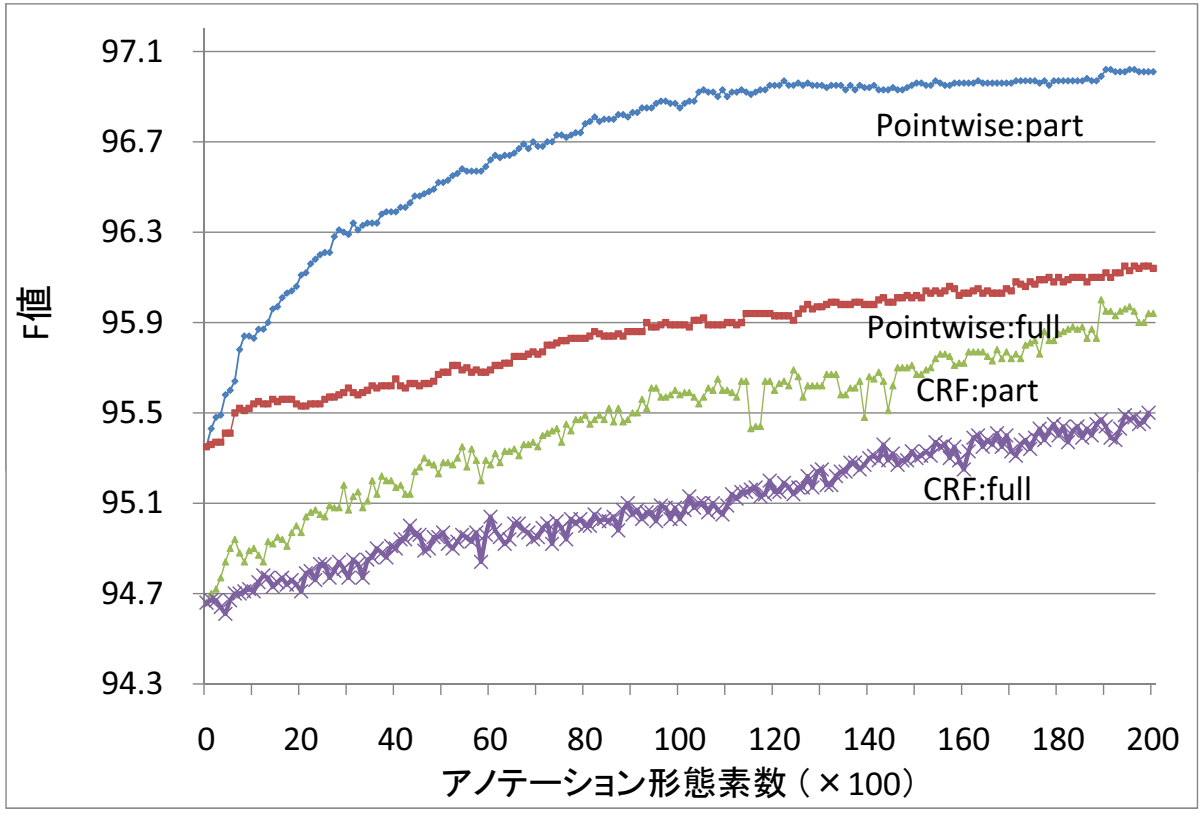
手法	単語分割	形態素解析
CRF(MeCab-0.98)	99.57	99.23
点予測 (KyTea-0.1.1)	99.37	98.86

  - 医薬品情報

手法	単語分割	形態素解析
CRF(MeCab-0.98)	94.69	92.94
点予測 (KyTea-0.1.1)	95.17	93.70

能動学習による精度向上

1. 曖昧な箇所をシステムが提示 (100 箇所)
2. 人手で作業 (& 精度評価)
3. goto 1.



Yahoo!知恵袋に対する結果

KWICによる未知語追加

1. **分布分析による自動単語抽出 [NL108, Coling1996]**
   (自動解析の結果からの未知語候補でも OK)
2. **KWIC (KeyWord In Context) を見て人手で修正**

→"target.part" **として保存**

3. **部分的アノテーションコーパスを追加して再学習**

```
% wget http://plata.ar.media.kyoto-u.ac.jp/*/fullLR.kfm
% train-kytea -feat fullLR.kfm -part target.part -model target.kbm
```

実用を考えるなら真面目に精度向上を

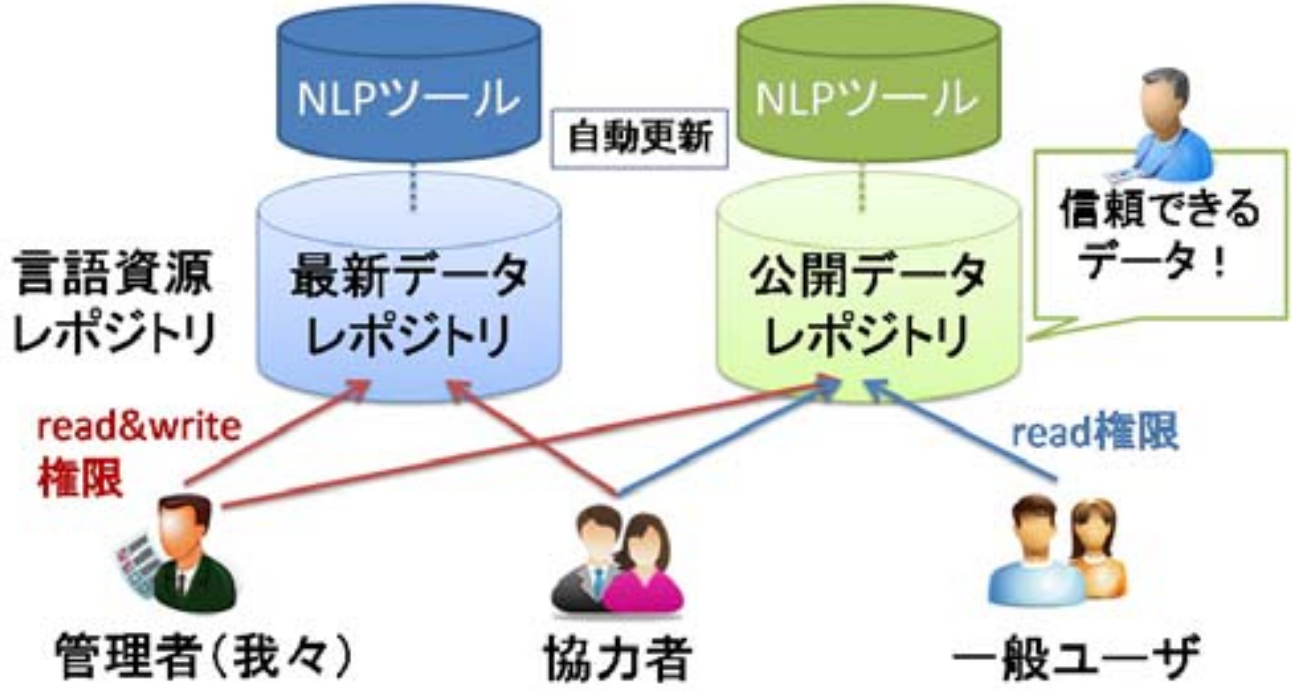
- 大学の研究ではツールをそのまま使う例がほとんど
  - Web からの知識獲得の研究をしてみた
    - Q. Web テキストの解析精度は?
    - A. ...
  - 辞書への単語追加をしてればましなほう
- 企業はコストを払って分野適応

例) 社内文書/医療所見/法律文書の
検索/テキストマイニング/音声認識

1. 対象分野のテキストの収集
2. テキストへの人手による部分的アノテーション
3. 自然言語処理システムの再学習
   - 元の学習データが必要 →Yes, KyTea can!

そうはいってもアノテーションは高コスト

- 部分的アノテーションテキストの共有 (この話は近々)



- ユーザーの言語活動の活用
  - 検索ログ
  - 仮名漢字変換ログ (研究中)
- ≈ Social IME @nokuno

おわりに

- 点予測による自然言語処理
- 部分的アノテーションによる学習データ作成の効率化
- ツールの公開
  - 単語分割，品詞推定，読み推定 KyTea(きゅーてぃー)
  - テキストからの未知語の自動抽出 PaLin(仮，未公開)
  - 単語単位の係り受け解析 (名前はまだ無い)
  - 言語資源共有の枠組み (名前はまだ無い)