部分的アノテーションによる言語処理

森　信介　Graham NEUBIG

京都大学

国語研サロン 2010 **年** 9 **月** 15 **日**

コーパスの統計による自然言語処理

- 入力文の断片を学習コーパスから探す
- 文脈を考えて最適な判断を下す

コーパスの統計による自然言語処理

- 入力文の断片を学習コーパスから探す
- 文脈を考えて最適な判断を下す

1. 適用分野の学習コーパスが必要

コーパスの統計による自然言語処理

- 入力文の断片を学習コーパスから探す
- 文脈を考えて最適な判断を下す

1. **適用分野の学習コーパスが必要**

2. **学習コーパスの質が重要**

取り組んでいる言語処理とその応用

- 単語分割
  - 検索
  - 仮名漢字変換 & 音声認識
  - その他多数
- 品詞推定
- 読み推定
  - 仮名漢字変換 & 音声認識
  - 音声合成
- 固有表現抽出
- 係り受け解析
- 機械翻訳

仮名漢字変換ログの利用

ちょっと寄り道

コーパスの部分文字列も列挙する仮名漢字変換

- 疑似確率的単語分割コーパス
  例) 単$_{0.2}$ 項$_{0.9}$ イ$_{0.1}$ デ$_{0.1}$ ア$_{0.1}$ ル$_{0.9}$ 整$_{0.3}$ 域$_{0.3}$ 上$_{0.9}$ の

試行	結果
1	単 項 / イ デ ア ル / 整 / 域 上 / の
2	単 項 / イ デ / ア ル / 整 域 / 上 / の
3	単 項 / イ デ ア ル / 整 域 上 / の
4	単 / 項 / イ デ ア ル / 整 域 / 上 / の

- 単漢字辞書から単語候補の読み推定
  例) 整域: せい いき，ととの いき, …
- 単語候補とその可能な読みも入れてモデルを作成

仮名漢字変換ログ

も/も 時間/じかん 文字/もじ 間/かん	注目/ちゅうもく 注目/ちゅうもく
例えば/たとえば 、/、 例えば/たとえば 、/、	場合/ばあい 場合/ばあい
列/れつ 列/れつ	場合/ばあい 場合/ばあい
複合/ふくごう 語/ご 複合/ふくごう 語/ご	じょうきおの/じょうきおの/UW じょうきおの/じょうきおの/UW
辞書/じしょ 辞書/じしょ	上記/じょうき の/の 上記/じょうき の/の
この/この 辞書/じしょ この/この 辞書/じしょ	○/れい 文/ぶん ○/れい 文/ぶん
連接/れんせつ/NW 連接/れんせつ/NW	○/れい 文/ぶん を/を 例文/れいぶん/NW を/を

言語処理への応用

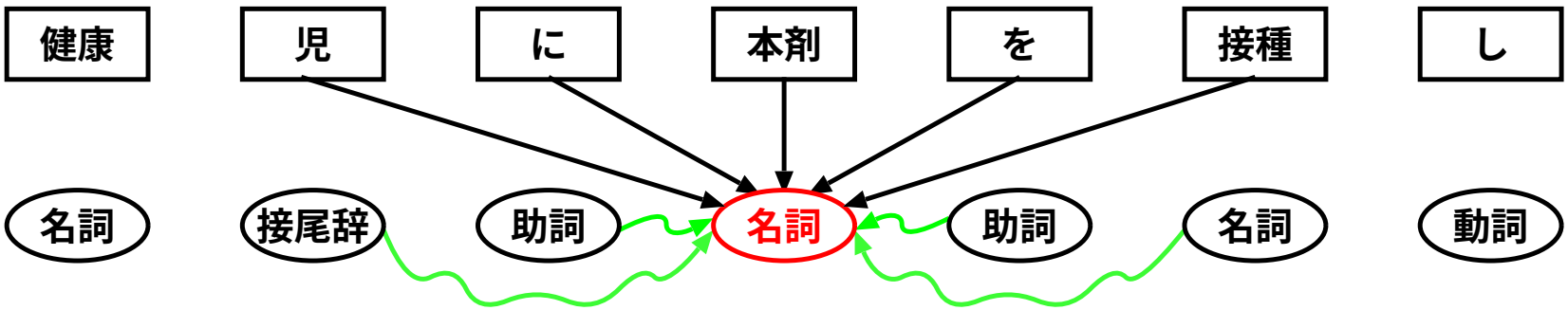
- 言語処理の精度向上をコストなしで実現
  - 仮名漢字変換
  - 単語分割
  - 読み推定
- 継続的に精度向上するのか?
- どう定式化するか?
  - ノイズありデータからの学習

系列予測と点予測

コーパス作成を意識した言語処理の設計

系列予測による方法 – 現在の主流 –

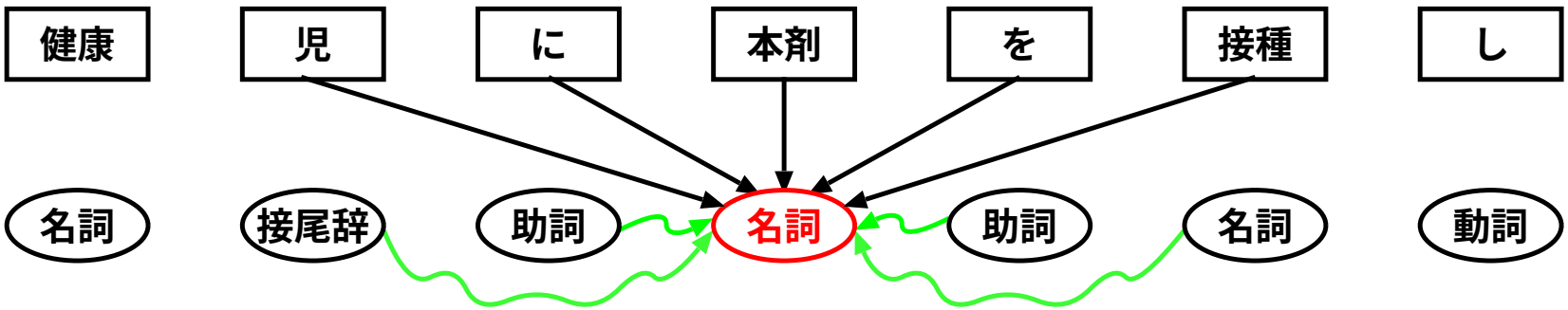
- 条件付き確率場 (CRFs; Condiational Random Fields)



- 1) 品詞を推定すべき単語と 2) その前の単語列と
  3) 後の単語列と 4) それらの品詞を参照する

系列予測による方法 – 現在の主流 –

- 条件付き確率場 (CRFs; Condiational Random Fields)



- 1) 品詞を推定すべき単語と 2) その前の単語列と
  3) 後の単語列と 4) それらの品詞を参照する

- 学習にはフルアノテーションコーパスが必要

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

アノテーションコストの最小化

- フルアノテーションコーパスは無駄が多い
  - 分野特有の表現以外は一般分野のコーパスでカバー済み

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

アノテーションコストの最小化

- フルアノテーションコーパスは無駄が多い
  - 分野特有の表現以外は一般分野のコーパスでカバー済み

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

- 部分的アノテーションコーパスを利用したい
  - 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

アノテーションコストの最小化

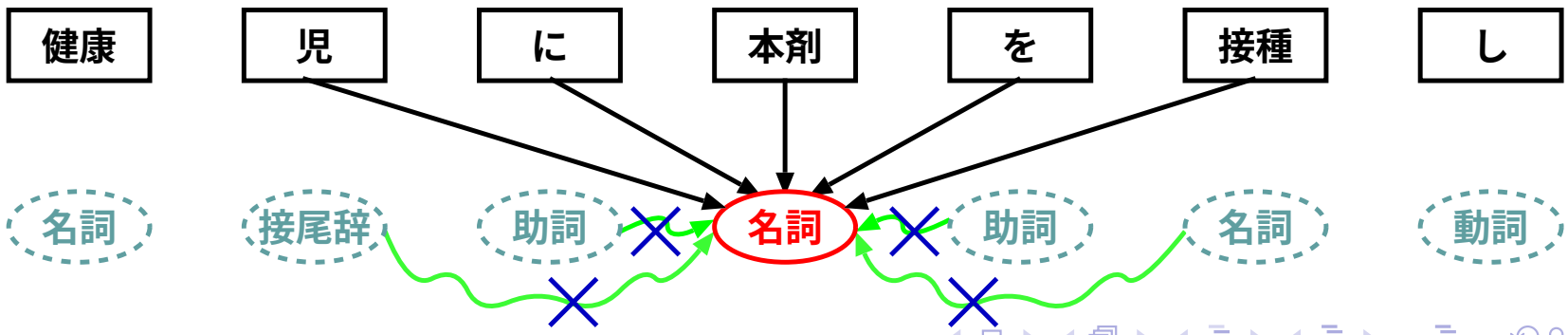
- フルアノテーションコーパスは無駄が多い
  - 分野特有の表現以外は一般分野のコーパスでカバー済み

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

- 部分的アノテーションコーパスを利用したい
  - 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

- 形態素解析をどう設計するか



アノテーションコストの最小化

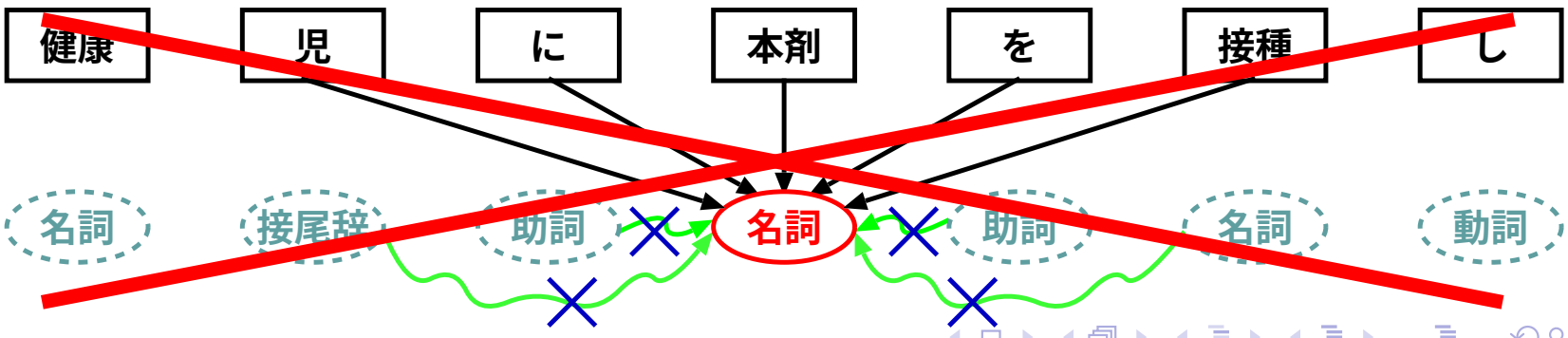
- フルアノテーションコーパスは無駄が多い
  - 分野特有の表現以外は一般分野のコーパスでカバー済み

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

- 部分的アノテーションコーパスを利用したい
  - 分野特有の表現のみ情報付与

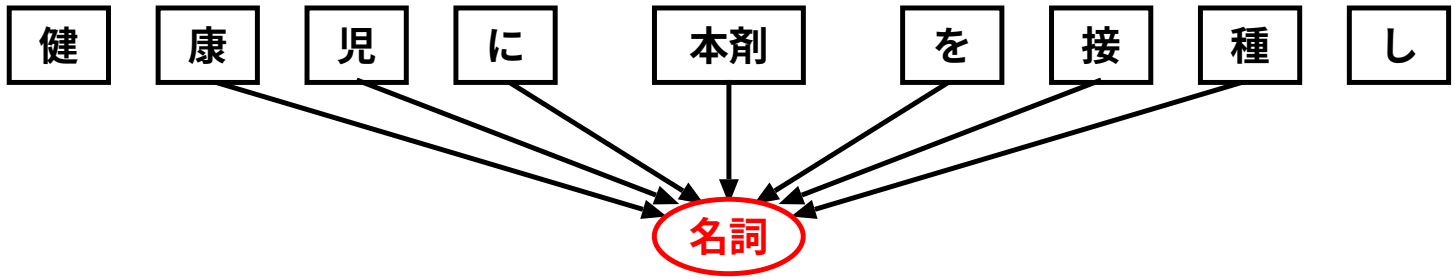
例) ガーゼ等は 本剤/名詞 を吸着する

- 形態素解析をどう設計するか



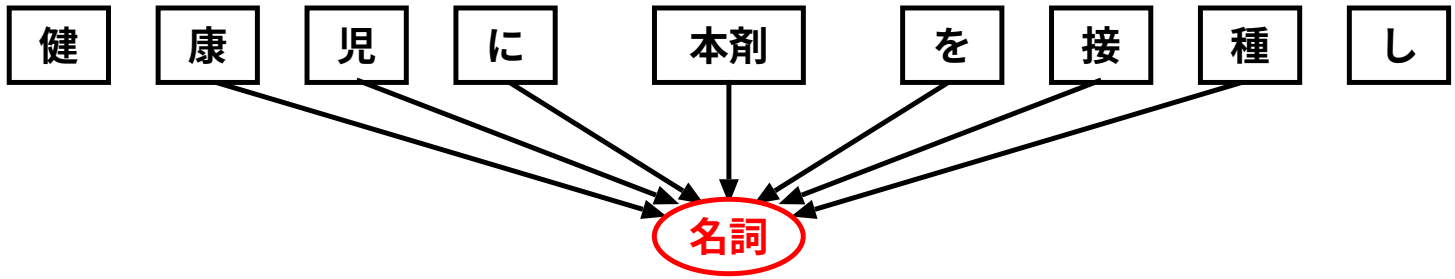
点予測による方法

- 各点の判断を独立と仮定して分類器を利用 (SVM 等)



点予測による方法

- 各点の判断を独立と仮定して分類器を利用 (SVM 等)

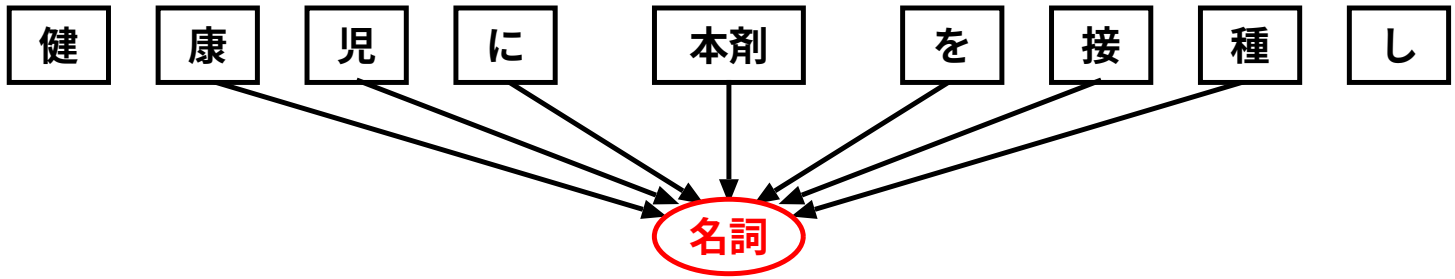


- 1) 注目単語と 2) その前後の**文字列**のみ参照

周辺の予測結果を参照しない!

点予測による方法

- 各点の判断を独立と仮定して分類器を利用 (SVM 等)



- 1) 注目単語と 2) その前後の**文字列**のみ参照

  周辺の予測結果を参照しない!

- **部分的アノテーションコーパス**が利用可能になる

  例) ガーゼ等は **本剤/名詞** を吸着する

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**

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

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

今後の展望

固有表現抽出

- 点予測による単語単位のタグ推定
  - 不適切なタグ系列が生成し得る
- 統一的解の探索

皮膚	粘膜	眼	症候	群	が	現れ
Dbeg 1.0	Dbeg 0.8	Dmid 1.0	Dmid 1.0	Dend 1.0		
	Dmid 0.2					

D: 病名, beg: 左端, mid: 中, end: 右端

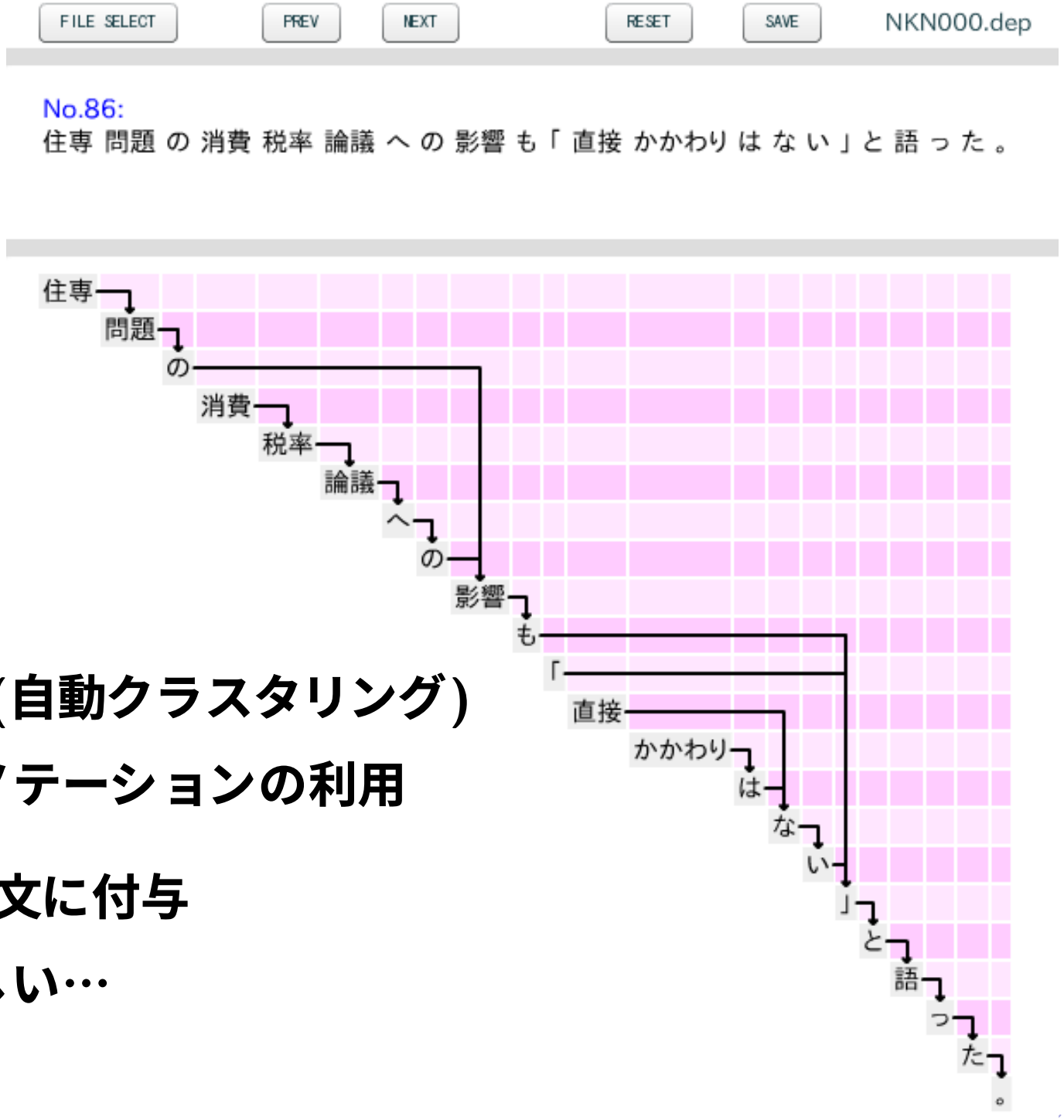
固有表現抽出

- 点予測による単語単位のタグ推定
  - 不適切なタグ系列が生成し得る
- 統一的解の探索

皮膚	粘膜	眼	症候	群	が	現れ
Dbeg 1.0	Dbeg 0.8	Dmid 1.0	Dmid 1.0	Dend 1.0		
	Dmid 0.2					

D: 病名, beg: 左端, mid: 中, end: 右端

構文解析



1. 単語単位
2. 品詞あり
   or なし (自動クラスタリング)
3. 部分的アノテーションの利用

- 約 25,000 文に付与
- 基準が難しい…

照応解析

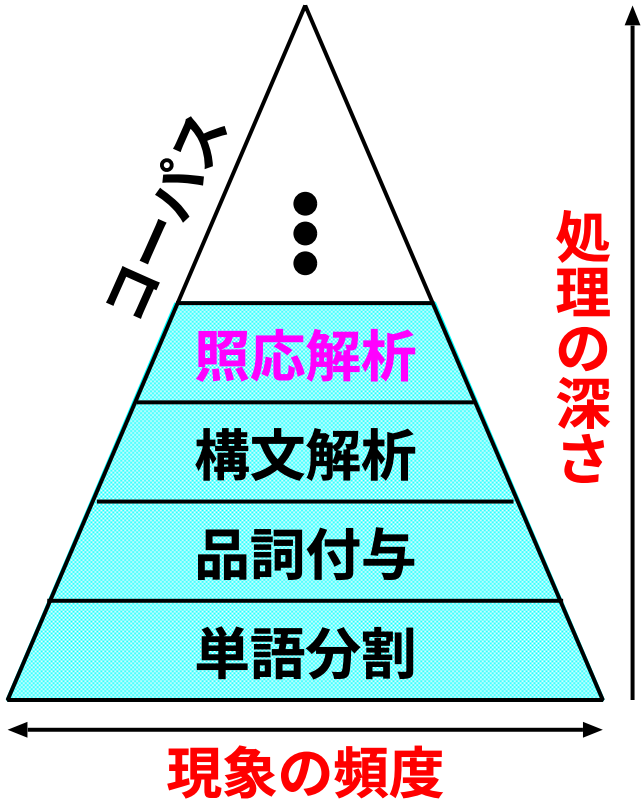
- 照応詞, ゼロ代名詞, …
- 照応詞と先行詞のみアノテーション (単語境界, 品詞?)

応用

1. 自動要約
2. 映像と言語のマッチング
3. 機械翻訳

フルアノテーションによる深い言語処理

- 前段までのフルアノテーションが必要

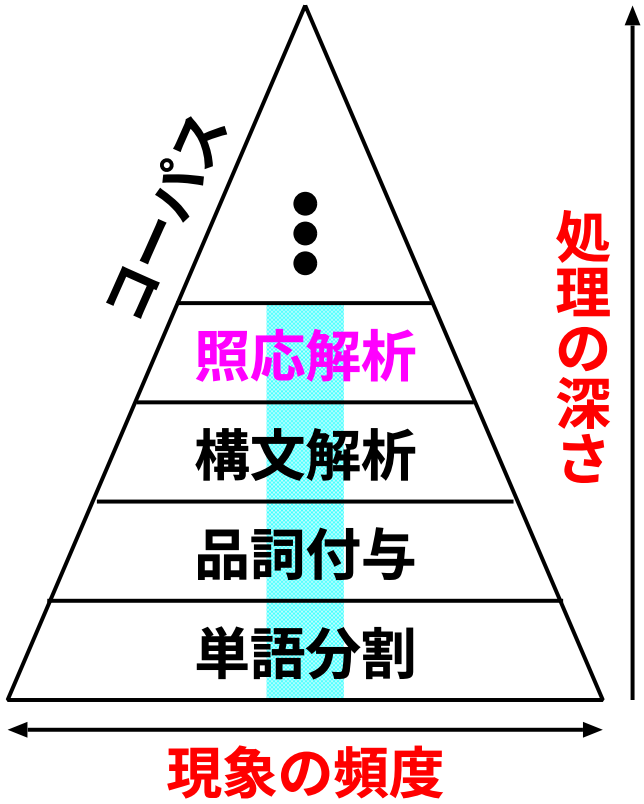


おたふくかぜに対して免疫のない健康児に本剤を接種した場合、接種後２～３週間頃、発熱、耳下腺腫脹、嘔吐、咳、鼻汁等を認めることがある。しかしこれらの症状は自然感染に比べ軽度であり、かつ一過性で、通常数日中に消失する。

- 照応が 1000 回出現するコーパスの単語数は?

深い言語処理には部分的アノテーション

- 深い言語処理のアノテーションに必須



おたふくかぜに対して免疫のない健康児に本剤を接種した場合、接種後２～３週間頃、発熱、耳下腺腫脹、嘔吐、咳、鼻汁等を認めることがある。しかしこれらの症状は自然感染に比べ軽度であり、かつ一過性で、通常数日中に消失する。

- 先行詞と照応詞のみアノテーション

1000 回 × 多くて 10 単語

部分的アノテーション + 点予測

1. **コーパス作成が容易**
2. **今までと同等以上の解析精度**
3. **簡便な言語処理システム**

部分的アノテーション + 点予測

1. コーパス作成が容易
2. 今までと同等以上の解析精度
3. 簡便な言語処理システム

これからはこれで!!