

ChalME:大規模コーパスを用いた統計的かな漢字変換

小町守 (NAIST)・森信介 (京大)・徳永拓之 (PFI)

背景・目的

計算機の高速化・メモリや HDD の大容量化
→大規模コーパスが利用可能に
機械学習・統計的手法の発展
→コーパスからの変換規則(モデル)の学習
→極力人手に頼らないかな漢字変換システム

ヒューリスティックによるかな漢字変換
□規則ベース変換 (Canna)
□N文節最長一致 (Wnn, VJE, IBATOK)
□最小コスト法 (WXG, IBMS-IME)
→辞書のメンテナンス・アルゴリズム開発のハードル

大規模コーパスによる統計的自然言語処理の応用
人手をかけない (Google 日本語 N グラム・Wikipedia など)

複雑な未知語処理・品詞情報を用いないかな漢字変換
大規模データから変換確率を推定・一般の開発者向け

手法

統計的かな漢字変換 (森ら, 1998): $P(x|y)$ の降順に変換可能文字列を提示する (x: 文, y: 入力)

かな漢字モデル $P(y|x)$ × 言語モデル $P(x)$ によるランキング

$$\therefore P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

言語モデル (n-gram) の学習

Google 日本語 N グラム データ (200 億文) から単語の 1, 2 グラム を計算 (異なり 1 グラム 数: 250 万; 異なり 2 グラム 数: 8,000 万)

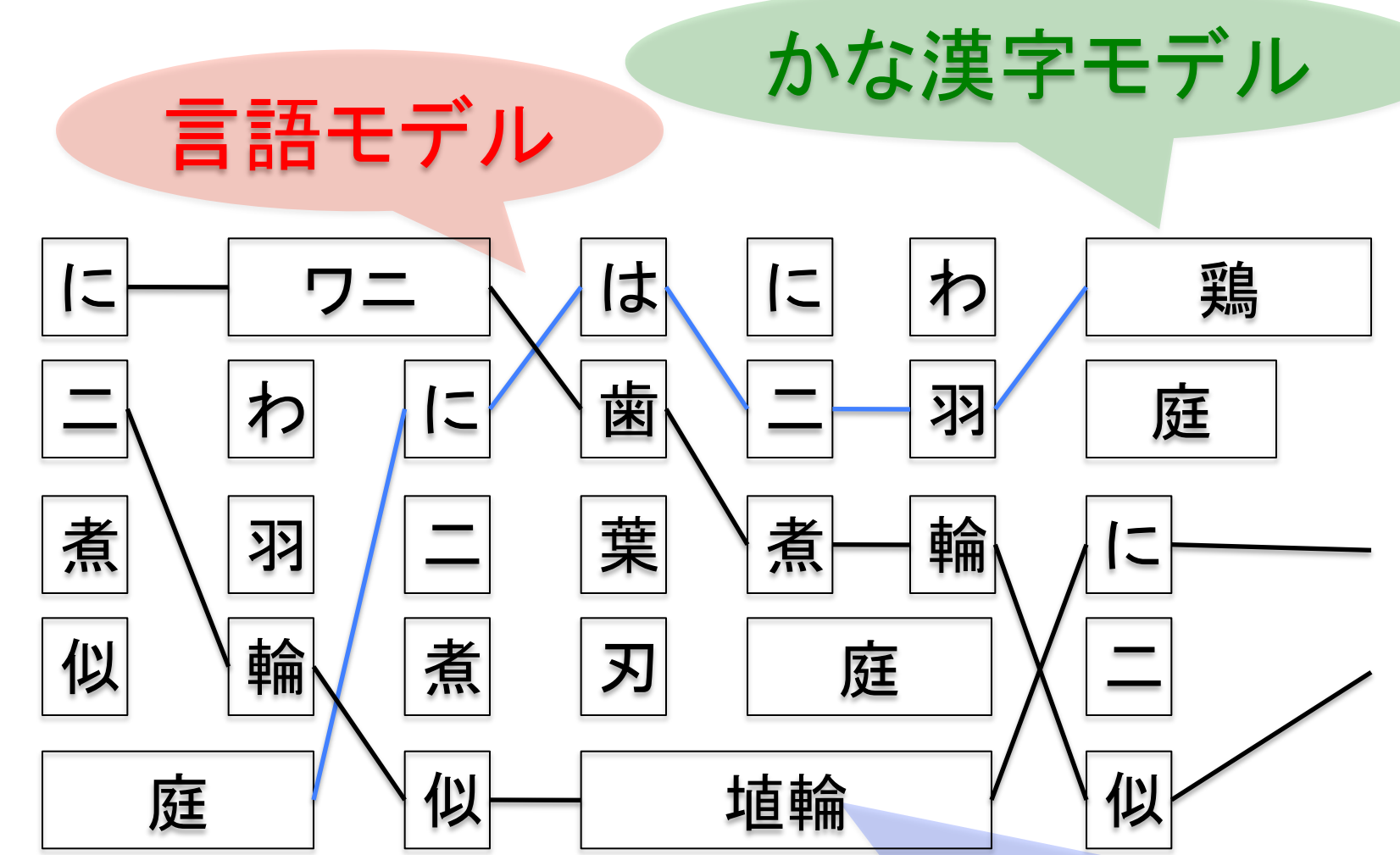
$$M_{w,n}(w) = \prod_{i=1}^{h+1} P(w_i | w_{i-n+1}^{i-1})$$

我が輩 → は → 猫 → である
↑ 文頭 ↓ 文末

かな漢字モデルの学習

出現する単語の読みはなにか推定するモデル
解析済みコーパスから頻度を計算して最尤推定
毎日新聞 13 年分を形態素解析器 MeCab で解析

$$M_{kk}(y|w) = \prod_{i=1}^h P(y_i | w_i)$$
$$P(y_i | w_i) = \frac{f(y_i, w_i)}{f(w_i)}$$



統計的かな漢字変換はヒューリスティック
最小コスト法の自然な拡張になっている

かつあき 小野克明さん (43)
こくめい 克明なやりとりが判明

変換結果 (デモ)・他の手法との比較

変換サンプル

ChalME	ATOK 2007	Anthy 9100c	AjaxIME
請求書の支払日時	請求書の市は来日時	請求書の支払い日時	請求書の支払いに知事
近く市場調査を行う。	知覚冗長さを行う。	近く市場調査を行う。	近く市場調査を行う。
その後サイト内で	その五歳都内で	その後サイト内で	その後再都内で
去年に比べ高い水準だ。	去年に比べた海水順だ。	去年に比べたかい水準だ。	去年に比べ高い水準だ。
屋イチまでに書類作っというて。	屋一までに書類津くっというて。	屋一までに書類作っというて。	肥留市までに書類作っというて。
そんな話信じっこないよね。	そんな話心十個内よね。	そんなはな視診時っこないよね。	そんな話神事っ子ないよね。
初めっから持ってけ方がいいのに。	恥メッカ持って毛羽いいのに。	恥メッカ羅持ってケバ飯野に。	始っから持ってけ方がいいのに。
熱々の肉まんにばくついた。	熱々の肉まん二泊着いた。	あつあつの肉まん2泊付いた。	熱熱の肉まんにばくついた。

※例文は <http://www.justsystems.com/jp/products/atok/> より取得

統計的手法

- 数学的モデルに基づいた理論的根拠
- 変換規則や辞書に当たる知識を自動で学習
- なにやっているのかよく分からない
- コーパスに激しく依存
- 変化の微調整が難しい

規則ベース

- なにをやっているのかはっきり分かる
- 柔軟な前処理・後処理
- 問題が簡単なきは規則ベースでもうまくいく
- アドホックなヒューリスティック (要言語学的直観)
- 「必殺パラメータ」

他の統計的手法

- Anthy
 - メンテナンスに言語学的知識が必要
- Sumibi
 - 単語分かち書きが必要・辞書にない単語は変換できない
- AjaxIME
 - 言語モデル・かな漢字モデルが貧弱

課題と今後の予定

- 単語入力履歴 (変換ログ) を用いた変換
- トピックモデルを用いた同音異義語の変換・単語のクラスタリングを用いた精度向上とモデル圧縮
- 予測入力との統合