大規模ウェブデータを用いた統計的自然言語処理

小町研究室（自然言語処理） 准教授 小町守 / 協力: 京都大学 森信介, PFI 徳永拓之, NTT 研究所 永田昌明, Apple 木田泰夫

大規模コーパスによる統計的自然言語処理の研究

できるだけ人手をかけないでメンテナンス

Google 日本語 N グラム・Wikipedia・etc…

頑健な深い自然言語処理解析技術の開発

大規模ウェブデータから、文の構造や意味を解析するための知識獲得・統計的モデルの学習

意味解析のツールを大規模なウェブテキストに適用

統計的かな漢字変換 ChaIME

P(かな漢字|入力)の降順に変換候補を提示

=P(入力|かな漢字)P(かな漢字)の降順に変換候補を提示(∵ベイズ則)

かな漢字モデル　言語モデル（200億文のGoogle日本語Nグラムから計算）

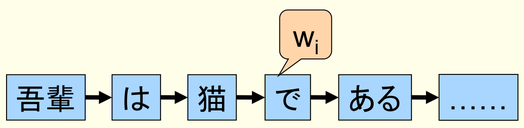
$$M_{kk}(y \mid w) = \prod_{i=1}^{h} P(y_i \mid w_i)$$

$$P(y_i \mid w_i) = \frac{f(y_i, w_i)}{f(w_i)}$$

こくめい？　かつあき？　克明

かな漢字モデル

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



言語モデル

名前	手法	コーパス	利点	欠点
ChaIME	単語表記2グラム	Google 日本語Nグラム	Google 日本語Nグラムに出現する単語なら自動で変換できる。コーパスが巨大なのでデータの過疎性の影響を受けにくい。自動単語分割を行うため、ユーザが単語分かち書きする必要がない。ブラウザ・uim から利用可能。	単語の表記で2グラムを作成しているので辞書サイズが巨大（2GB）になる。
Anthy	最大エントロピー法	独自コーパス（1万文）	機械学習による高精度な変換。文節の概念がある。ユーザの入力履歴からの予測入力が可能。Windows, Mac, Linux などさまざまなプラットホームで動作する。Linux でのユーザが多く、現在デファクトスタンダード。Emacs・uim・SCIM・ibus から利用可能。	モデルが複雑でありパラメータ推定がヒューリスティックである。コーパスの質・量ともに不十分たなめ、変換精度が悪い。
AjaxIME	品詞クラス2グラム	京大コーパス（4万文）	識別モデルによる高精度な変換。1文の変換結果のN-best解から文全体の変換結果を選択。ブラウザから使うことができるので、IMEがインストールされていない海外でも利用可能。uim でも動作。	コーパスのサイズが小さく、単語（文節）単位での変換をサポートしていない。かな漢字モデルが考慮されていない。学習しない。
Sumibi	単語表記2グラム	Webデータ（数GB）	ユーザが単語の分かち書きを指定するため、原理上単語分割ミスがない。分かち書きされたデータがあれば、任意のデータを学習に使うことができる。ブラウザ・Emacs・uim から利用可能。	連文節変換がサポートされておらず、単語分割を明示的に指定する必要がある。辞書にない単語は変換できない。学習しない。
Mana	品詞クラス2グラム	京大コーパス（4万文）	確率的言語モデルによる高精度な変換。単語単位での変換をサポート。ChaSen のコードを参考にしている。Emacs・uim から利用可能。	コーパスのサイズが小さい。辞書の情報が形態素解析用のままで、かな漢字変換用にチューニングされていない。学習しない。
Google 日本語入力 Mozc	品詞クラス2グラム	Google Web データ（200億文以上）	大規模なウェブデータを用いたかな漢字変換。ウェブから抽出した圧倒的な語彙。予測入力も可能。Windows と Mac でリリースされ、オープンソース版の Mozc は Linux でも ibus を用いることにより動作。	ウェブから学習しているので思いがけない単語が予測・変換される。長距離の単語の共起を扱えない（ただし上記のIMも同様）。

	ChaIME	ATOK 2007	Anthy 9100c	AjaxIME	Google 日本語入力 (Mozc)
せいきゅうしょのしはらいにちじ	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	請求書の支払い日時
ちかくしじょうちょうさをおこなう。	近く市場調査を行う。	知覚し冗長さを行う。	近く市場調査を行う。	近く市場調査を行う。	近く市場調査を行う。
そのございとないで	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	その後サイト内で
きょねんにくらべたかいすいじゅんだ。	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。	去年に比べ高い水準だ。
ひるいちまでにしょるいつくっといて。	昼イチまでに書類作っといて。	昼一までに書類津くっといて。	昼一までに書類作っといて。	肥留市までに書類作っといて。	昼一までに書類作っといて。
そんなはなししんじっこないよね。	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。	そんな話しんじっこないよね。
はじめっからもってけばいいのに。	初めっからもってけばいいのに。	恥メッカら持って毛羽いいのに。	恥メッカ羅持ってケバ飯野に。	始っから持ってけば良いのに。	はじめっから持ってけばいいのに。
あつあつのにくまんにぱくついた。	熱々の肉まんにぱくついた。	熱々の肉まん二泊着いた。	あつあつの肉まん2泊付いた。	熱熱の肉まんにぱくついた。	熱々の肉まんにぱくついた。

ATOK 2007 の誤変換例から抜粋

統計的機械翻訳

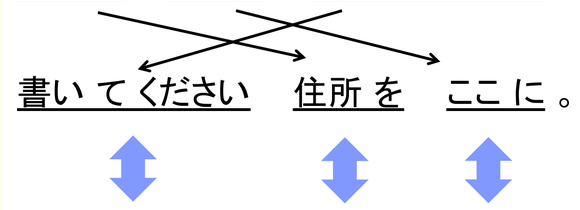
P(英語|日本語)の降順に翻訳候補を提示（※日英翻訳の場合）

=P(日本語|英語)P(日本語)の降順に翻訳候補を提示

翻訳モデル　言語モデル

$$P(e \mid j) = \frac{\exp\left(\sum_{m=1}^{M} \lambda_m h_m(e,j)\right)}{\sum_{e'} \exp\left(\sum_{m=1}^{M} \lambda_m h_m(e',j)\right)}$$

→対数線形モデルを用いると、$\sum_{m=1}^{M} \lambda_m h_m(e,j)$ の降順に翻訳候補を提示



Wikipedia からの対訳辞書構築

言語間リンクがある→対訳候補

専門用語をシードとして与えて分野適応

述語項構造解析を用いた語順の並べ替え

述語項構造解析器により日本語を SVO に並べ替え、句ベースの統計的機械翻訳器で学習