学術論文執筆のための仮名漢字変換システム

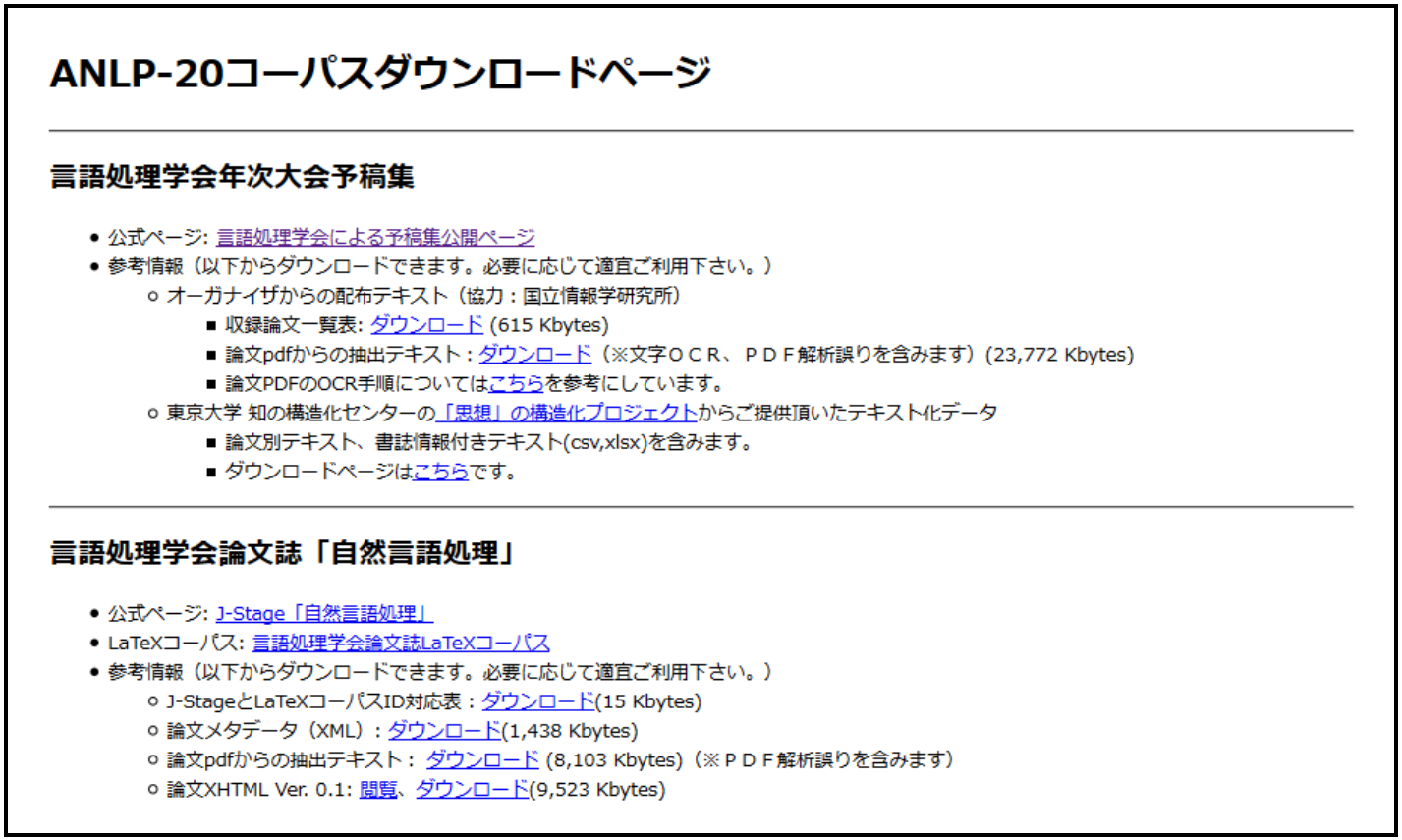
高橋 文彦　京都大学
前田 浩邦　サイボウズ
森 信介　京都大学

2015 年 3 月 19 日

特別セッション

のはずが…

- 言語処理学会設立 20 周年記念事業



ANLP-20コーパスダウンロードページ

言語処理学会年次大会予稿集

- 公式ページ: 言語処理学会による予稿集公開ページ
- 参考情報（以下からダウンロードできます。必要に応じて適宜ご利用下さい。）
  - オーガナイザからの配布テキスト（協力：国立情報学研究所）
    - 収録論文一覧表: ダウンロード (615 Kbytes)
    - 論文pdfからの抽出テキスト：ダウンロード（※文字ＯＣＲ、ＰＤＦ解析誤りを含みます）(23,772 Kbytes)
    - 論文PDFのOCR手順についてはこちらを参考にしています。
  - 東京大学 知の構造化センターの「思想」の構造化プロジェクトからご提供頂いたテキスト化データ
    - 論文別テキスト、書誌情報付きテキスト(csv,xlsx)を含みます。
    - ダウンロードページはこちらです。

言語処理学会論文誌「自然言語処理」

- 公式ページ: J-Stage「自然言語処理」
- LaTeXコーパス: 言語処理学会論文誌LaTeXコーパス
- 参考情報（以下からダウンロードできます。必要に応じて適宜ご利用下さい。）
  - J-StageとLaTeXコーパスID対応表：ダウンロード(15 Kbytes)
  - 論文メタデータ（XML）：ダウンロード(1,438 Kbytes)
  - 論文pdfからの抽出テキスト：ダウンロード (8,103 Kbytes)（※ＰＤＦ解析誤りを含みます）
  - 論文XHTML Ver. 0.1: 閲覧、ダウンロード(9,523 Kbytes)

http://nlp20.nii.ac.jp/resources/

- 主に論文執筆のための仮名漢字変換を作ろう!

自分で使う NLP

- 究極の主観評価
  1. 解くべき問題か
  2. 本当に問題が解けているか
  3. 周辺も含め使えるレベルか
- ログ収集で言語資源の自動獲得
  - ログは自然言語の内部情報を反映

仮名漢字変換 – 我的歴史/wodelishi –

- 1978 年 9 月 東芝ワープロ仮名漢字変換, 630 万円!
  ルールベース (内省によるスコア)
- 1998 年 「確率的モデルによる仮名漢字変換」 [森+ IPSJ-NL]
  中国語 IM [Chen ACL2000]
- 2002 年 Anthy (未踏 by tabata 氏)
  実はコーパスが重要だった! (国語研さまさま!)
- 2006 年 「無限語彙の仮名漢字変換」 [Mori+ COLING]
  テキスト の全部分文字列を候補に列挙
- 2009 年, Kagami, Google IME, Baidu IME, (KyTea)

OS には IME が必要

統計的仮名漢字変換 [森+ 1998]

- 単語と入力記号列の組 $u = \langle w, \vec{y} \rangle$ を言語モデルの単位

$$P(\vec{w}|\vec{y}) = P(\vec{w}, \vec{y})/P(\vec{y}) = P(\vec{u})/P(\vec{y})$$

分母 $P(\vec{y})$ は出力に依らないので分子だけモデル化

$$P(\vec{u}) = \prod_{i=1}^{h} P(u_i|\vec{u}_{i-n+1}^{i-1})$$

$$P(u_i|\vec{u}_{i-n+1}^{i-1}) = \begin{cases} P(u_i|\vec{u}_{i-n+1}^{i-1}) & \text{if } u_i \in \mathcal{U} \\ P(\mathtt{UU}|\vec{u}_{i-n+1}^{i-1}) M_{v,n}(u_i) & \text{if } u_i \notin \mathcal{U} \end{cases}$$

$\mathcal{U}$: 言語モデルの語彙 (単語と入力記号列の組の集合)

パラメータ推定

- 単語と読みの組を単位とする n-gram 確率

$$P(u_i|\vec{u}_{i-n+1}^{i-1}) = \frac{f(\vec{u}_{i-n+1}^{i})}{f(\vec{u}_{i-n+1}^{i-1})}$$

単語分割済み・入力記号列付与済みのコーパスが必要

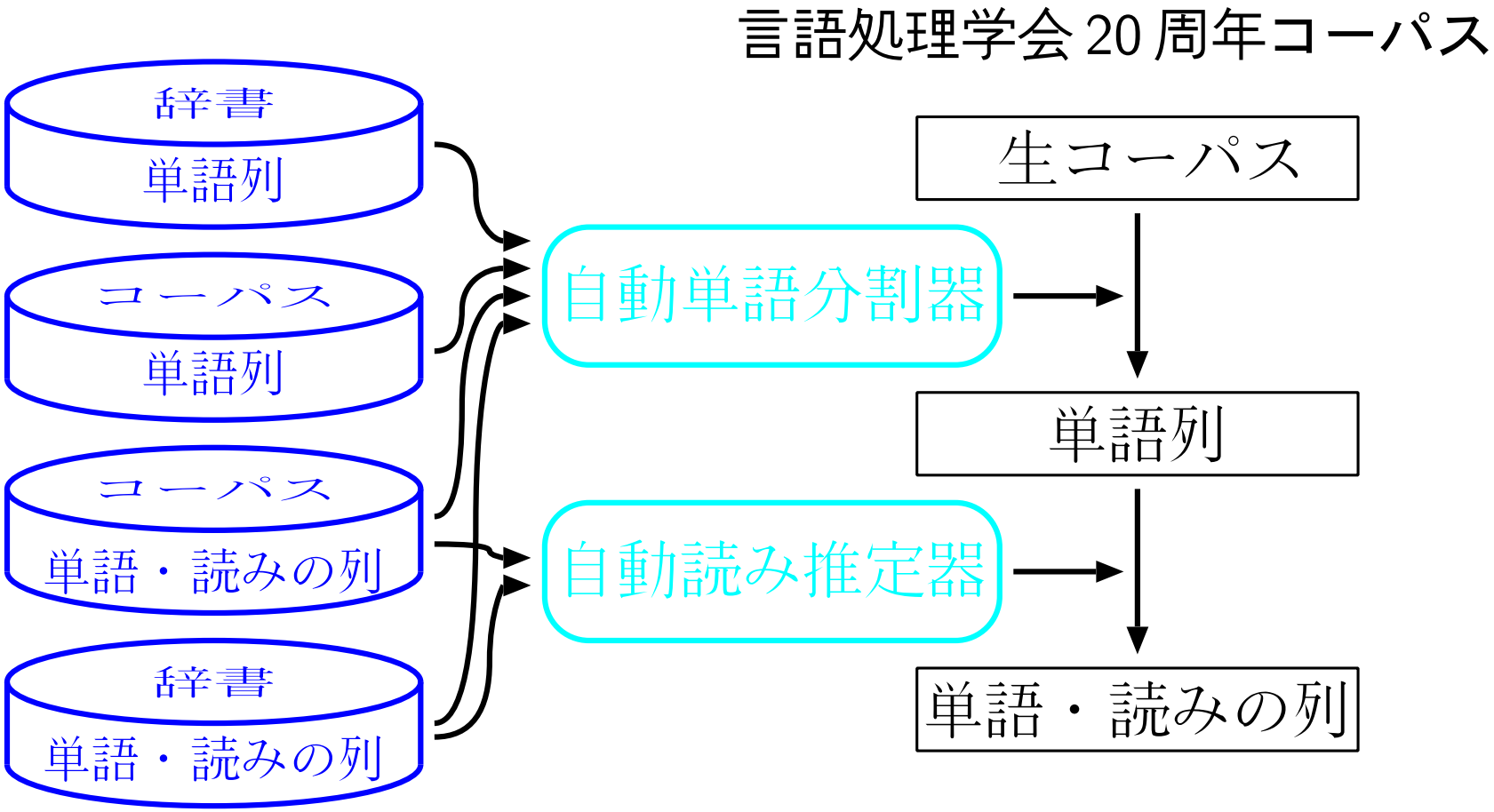
⇒ 京都テキスト解析システム KyTea

`http://www.phontron.com/kytea/`

Cf. JUMAN, MeCab でも OK

コーパスの処理

- 単語境界と読み付与を多段に処理



テキストの全部分文字列を候補に列挙

テキスト

グルタミン酸デカルボキシラーゼに対する抗体が見つかりますが、この抗体が症状を起こしているのかどうかは不明です。しかし、関節や神経異常などがみられることはまれです。

⇒

部分文字列: ヨミ, ヨミ, …

グ　　: グ

グル　: グル

グルタ: グルタ

⋮

抗体　: コウタイ, アラガタイ, …

⋮

神経異: コウタイ, シンケイイ, …

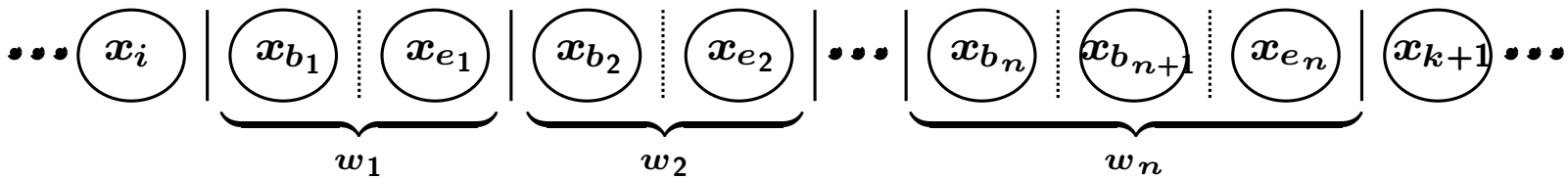
⋮

入力: イジョウナコウタイガジブンジシンヲ

出力: 異常な ? が自分自身を

確率的単語分割 [Mori+ 2004]

- 文字間に単語境界確率 P_i を設定
- 期待頻度を計算 $\Rightarrow$ 単語 n-gram モデル



$$f_r(w_1^n) = P_i(1-P_{b_1})P_{e_1}(1-P_{b_2})P_{e_2}\cdots(1-P_{b_n})(1-P_{b_{n+1}})P_{e_n}$$

- あらゆる部分文字列の出現確率が 0 より大きくなる
- 確率的言語モデルになっている (証明 [森+ 2004])

自動単語分割 [Neubig+ 2011]

- 直前の文字列と直後の文字列から単語境界確率を返す

$$\vec{x}_- \vec{x}_+ \mapsto P_i$$

- KyTea (京都テキスト 解析ツールキット)
  `http://www.phontron.com/kytea/`

確率的単語分割と相性が良い

CRF では周辺化が必要

確率的タグ付与 [森+ 2011]

- 単自動読み推定の結果には一定量の誤り
  ⇒ 言語モデルや仮名漢字モデル (発音辞書) に悪影響
- 各単語にタグ t と確率値 p の組の列を付与

$$w \mapsto \langle t_1, p_1 \rangle, \langle t_2, p_2 \rangle, \cdots$$

例)

両国	⟨りょうこく, **0.91**⟩, ⟨りょうごく, **0.09**⟩
の	⟨の, **1.00**⟩
代表	⟨だいひょう, **1.00**⟩
に	⟨に, **1.00**⟩
…	

cf. 決定的タグ付与 $p_i = 1 \wedge p_j = 0,\ \forall j \neq i$

自動読み推定 [Mori+ 2012]

- 注目単語 w_i と文脈から w_i の読みと確率の組の列を返す

$$\vec{x}_- \, w_i \, \vec{x}_+ \mapsto (\langle \vec{y}_1, p_1 \rangle, \langle \vec{y}_2, p_2 \rangle, \cdots)$$

- KyTea (京都テキスト 解析ツールキット)
  `http://www.phontron.com/kytea/`

確率的タグ付与と相性が良い

CRF では周辺化が必要

確率的タグ付与済みコーパス

- 1 文は w: 単語, t: 読み, p: 確率 として

$$\begin{aligned}
&\langle w_1, (\langle t_{1,1}, p_{1,1}\rangle, \langle t_{1,2}, p_{1,2}\rangle, \cdots, \langle t_{1,k_1}, p_{1,k_1}\rangle)\rangle \\
&\langle w_2, (\langle t_{2,1}, p_{2,1}\rangle, \langle t_{2,2}, p_{2,2}\rangle, \cdots, \langle t_{2,k_2}, p_{2,k_2}\rangle)\rangle \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\
&\langle w_h, (\langle t_{h,1}, p_{h,1}\rangle, \langle t_{h,2}, p_{h,2}\rangle, \cdots, \langle t_{h,k_h}, p_{h,k_h}\rangle)\rangle
\end{aligned}$$

- 単語とタグの組の n-gram の 1 回の出現あたりの頻度

$$f_1(\langle w_1, t_{j_1}\rangle\langle w_2, t_{j_2}\rangle \cdots \langle w_n, t_{j_n}\rangle) = \prod_{i=1}^{n} p_{j_i}$$

- 期待頻度はこの値の全出現箇所に渡って合計

擬似確率的タグ付与済みコーパス

- 期待頻度計算は計算コスト大
- モンテカルロ法による通常コーパスへの変換

  foreach 単語
  1. タグの確率と乱数を比較
  2. 単語のタグを唯一に決定

  これを繰り返して m 倍の決定的読み付与コーパス
- 収束に関して

モンテカルロ法による d 次元の単位立方体 $[0,1]^d$ 上の定積分 $I = \int_{[0,1]^d} f(x)dx$ の数値計算法と同じ

誤差 $|I_N - I|$ は次元 d によらずに $1/\sqrt{N}$ に比例

n によらずに $1/\sqrt{Fm}$ に比例する程度の速さで減少

テキストの全部分文字列を候補に列挙

テキスト

グルタミン酸デカルボキシラーゼに対する抗体が見つかりますが、この抗体が症状を起こしているのかどうかは不明です。しかし、関節や神経異常などがみられることはまれです。

⇒

部分文字列: ヨミ, ヨミ, …

グ　　: グ

グル　: グル

グルタ: グルタ

⋮

抗体　: コウタイ, アラガタイ, …

⋮

神経異: コウタイ, シンケイイ, …

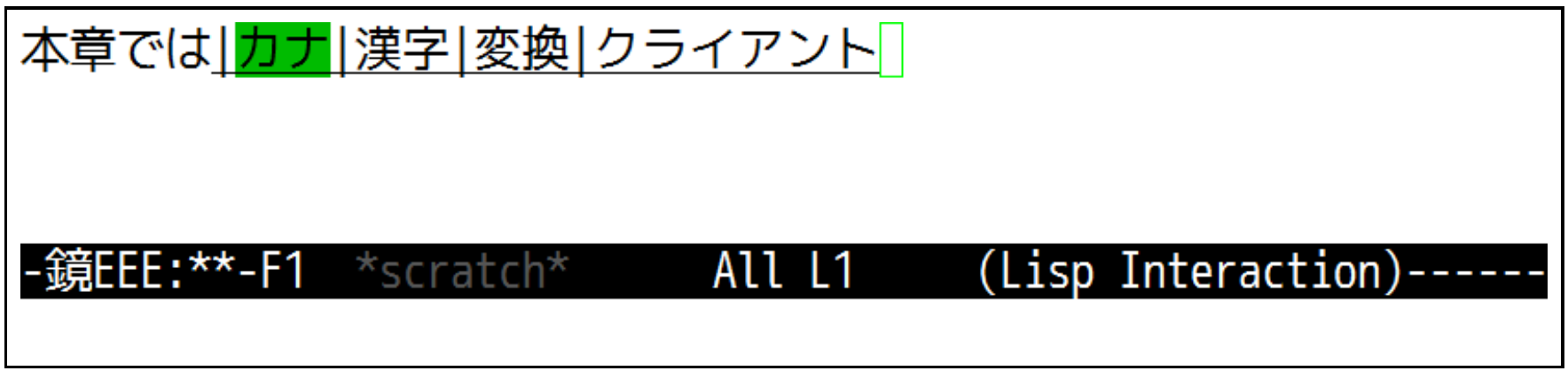
⋮

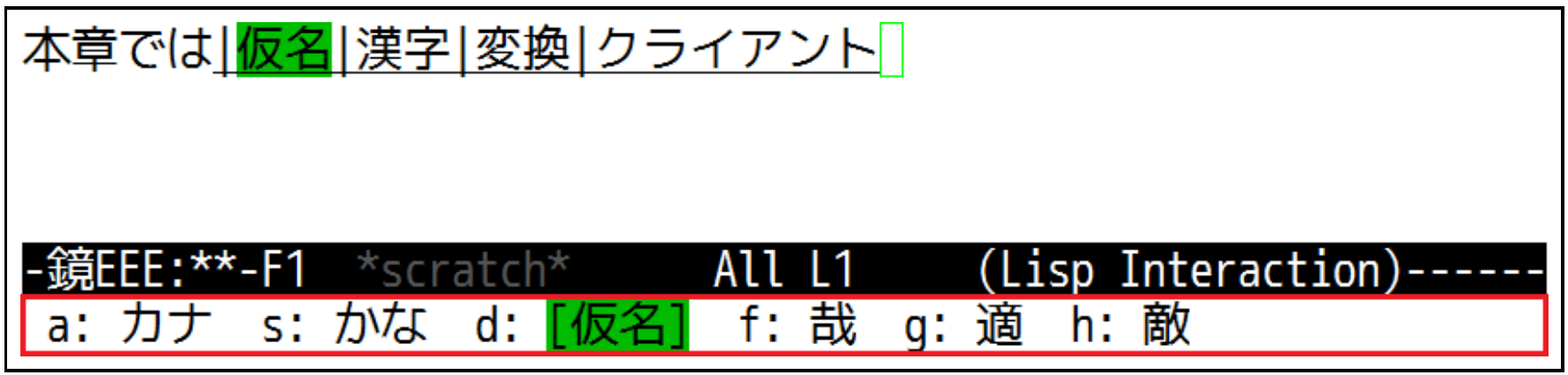
入力: イジョウナコウタイガジブンジシンヲ

出力: 異常な ? が自分自身を

Emacs のクライアント (Kagami)

- 論文や堅めの文書の執筆 (Cf. Kagami for Firefox add-on)
- クライアント (kagami.el)





Kagami 主要諸元 (足回り)

- KyTea の学習の言語資源

学習コーパス	文数	単語数	文字数
BCCWJ Core	56,753	1,324,951	1,911,660
日経新聞	8,164	240,097	361,843
和英辞書の例文	11,700	147,809	197,941
辞書		単語数	
UniDic			234,652
姓名			197,552
数字			280

Kagami 主要諸元 (エンジン)

- 言語モデルの学習コーパス

学習	分野	文数	単語数	文字数
BCCWJ-manu	一般分野	56,753	1,324,951	1,911,660
BCCWJ-raw	一般分野	716,154	16,749,959	23,782,812
NLP-raw	自然言語処理	43,173	1,552,650	2,504,356
MED-raw	医学	50,915	1,561,245	2,141,620
テスト	分野	文数	単語数	文字数
BCCWJ-test	一般分野	6,025	148,929	212,261
NLP-test	自然言語処理	265	29,368	41,738
MED-test	医学	1,000	8,666	12,775

Kagami 走行テスト

- 1 文一括変換の F 値

モデル	学習コーパス			テスト コーパス		
	BCCWJ	NLP	MED	BCCWJ	NLP	MED
G--	o			93.70	89.45	93.55
GN-	o	o		93.69	96.33	93.54
G-M	o		o	93.71	89.71	97.08
GNM	o	o	o	**93.73**	**96.52**	**97.10**

- 自然言語処理 (の文) は意外に難しい!!
- 適応すると大きく精度向上
- 各分野への適応は加法的に作用 $\Rightarrow$ ワンモデルで OK

おわりに

- 使って楽しい NLP!!

時々イラっと

- 私は 2009 年夏から使ってます
  - 論文を書くのにすこぶる便利です
  - いろいろな改良を思いつく
- よろしければどうぞ

  `https://github.com/ShinsukeMori/kagami`
- 時々NLP に寄りすぎます (笑)

  例) 語彙論ありませんか → ご異論

To Do

- 機能追加・改良したい人 Welcome
  - 多分野対応 (全学問領域 or 個別企業)
  - 学習機能
  - 予測
  - UI / クライアント
  - ⋮
- ログの活用 (ユーザー Welcome!) [高橋+ 2015]
  - Cf. KAGAMI for Firefox (Twitter)

    `https://plata.ar.media.kyoto-u.ac.jp/takahasi/kagami/`

References

- Chen, Z. and Lee, K.-F.: A New Statistical Approach To Chinese Pinyin Input, in *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 241–247 (2000)
- Maeta, H. and Mori, S.: Statistical Input Method based on a Phrase Class n-gram Model, in *Workshop on Advances in Text Input Methods* (2012)
- Mori, S. and Takuma, D.: Word N-gram Probability Estimation From A Japanese Raw Corpus, in *Proceedings of the Eighth International Conference on Speech and Language Processing* (2004)

- Mori, S., Takuma, D., and Kurata, G.: Phoneme-to-Text Transcription System with an Infinite Vocabulary, in *Proceedings of the 21st International Conference on Computational Linguistics* (2006)
- Mori, S. and Neubig, G.: A Pointwise Approach to Pronunciation Estimation for a TTS Front-end, in *Proceedings of the InterSpeech2011*, pp. 2181–2184, Florence, Italy (2011)
- Neubig, G. and Mori, S.: Word-based Partial Annotation for Efficient Corpus Construction, in *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (2010)
- Neubig, G., Nakata, Y., and Mori, S.: Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 529–533 (2011)

- Sasada, T., Mori, S., and Kawahara, T.: Extracting Word-Pronunciation Pairs from Comparable Set of Text and Speech, in *Proceedings of the InterSpeech2008*, pp. 1821–1824 (2008)
- 高橋 文彦, 森 信介■仮名漢字変換ログを用いた単語分割の精度向上, 言語処理学会第 21 回年次大会発表論文集 (2015)
- 森 信介, 土屋 雅稔, 山地 治, 長尾 真■確率的モデルによる仮名漢字変換, 情報処理学会論文誌, Vol. 40, No. 7, pp. 2946–2953 (1999)
- 森 信介■無限語彙の仮名漢字変換, 情報処理学会論文誌, Vol. 48, pp. 3532–3540 (2007)
- 森 信介, 宅間 大介, 倉田 岳人■確率的単語分割コーパスからの単語 N-gram 確率の計算, 情報処理学会論文誌, Vol. 48, No. 2, pp. 892–899 (2007)

- 森 信介, 前田 浩邦■利用過程で得られる言語情報を活用する音声言語処理システム, NLP 若手の会第 4 回シンポジウム (2009)
- 森 信介, 小田 裕樹■擬似確率的単語分割コーパスによる言語モデルの改良, 自然言語処理, Vol. 16, No. 5, pp. 7–21 (2009)
- 森 信介, Neubig, G. ■仮名漢字変換ログの活用による言語処理精度の自動向上, 言語処理学会年次大会 (2010)
- 森 信介, 笹田 鉄郎, Graham, N. ■確率的タグ付与コーパスからの言語モデル構築, 自然言語処理, Vol. 18, No. 2 (2011)