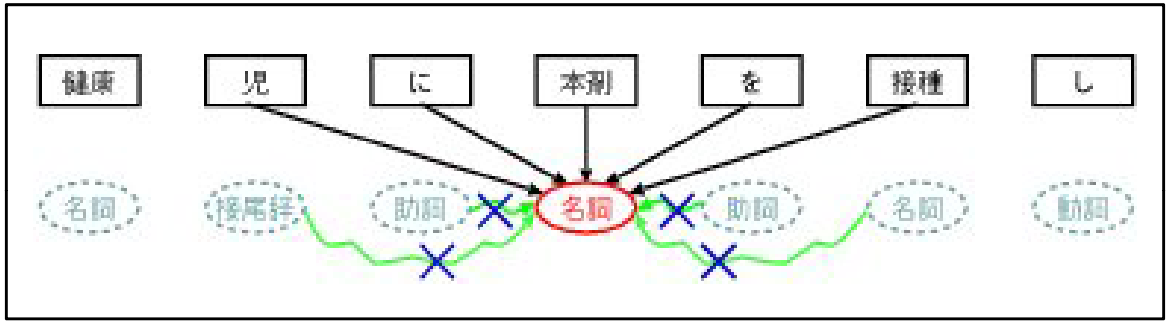
日本語の単語分割・品詞推定 あるいは KyTea の話

森 信介

京都大学



http://www.ar.media.kyoto-u.ac.jp/
Twitter: @KyTeaJP

単語分割・タグ推定

形態素解析関連の経験

- 全役割の経験者!! 私だけ?!
  - 国際学会発表 ACL, EMNLP, COLING, EuroSpeech
  - 設計 KyTea
  - 実装 クラスn-gramモデル
  - 基準の策定 @IBM TRL for ASR, TTS
  - アノテーション 5,000文/10営業日
  - 利用 音声認識, 仮名漢字変換, 文生成, 企業共同研究, …

不都合な真実 統計的単語分割編

1. 使ってみると精度が出ない (論文では約99%)
   - レシピ 95%, 将棋解説 91%
     - 学習データの分野と異なる
     - 均衡コーパス (BCCWJ) で学習しているのに

2. アノテーションの効果 ≫ 手法の改善の研究
   - レシピ 98%超, 将棋解説 98%超
     - 約3,000文のアノテーションなど (8時間 x 5日)

3. しかしアノテーションの効果は対数的
   - クロスドメインで有効な手法
   - 学習データの自動的な収集

趣旨への回答

- KyTea (きゅーてぃー): Kyoto Text Analyzer
  - 設計: 森 信介 (言語モデル屋, 言語資源の管理)
  - 実装: Prof. Neubig Graham
- 特徴・開発方針
  - 単純な解法
  - 10万文以上の学習コーパス、延べ30年以上の言語資源増強
  - 様々な (学習データのない) 分野での高い精度
  - 言語資源構築も含めた最適化 → 安価かつ迅速な分野適応
- 想定ユーザ (∋ 発注者😊)
  - 目的のテキストでの高い精度が必要な方
  - 音声認識・合成,仮名漢字変換
  - 検索・機械翻訳
  - 単語に対する何らかのラベルを推定したい方

単語分割・タグ推定

形態素解析関連の研究

- 分布分析による未知語収集 [COLING96]
- 形態素やクラスのn-gramによる形態素解析 [自然言語処理98]
- 発音とアクセントの推定 [Nagatno+ EuroSpeech05]
- 可変長クラスn-gramモデル [EuroSpeech05]
- 音声からの言語資源の獲得 [Kurata+ ICASSP07][Sasada+ IS08]
- 部分的アノテーションからのCRF学習 [Tsuboi+ COLING08]
- 複合語辞書の利用 [PACLING09]
- 点予測による単語分割・品詞推定 [Neubig+ LREC10][Neubig+ ACL11]
- 点予測による読み推定 [InterSpeech11]
- 仮名漢字変換ログからの言語資源の獲得 [Takahashi+ EMNLP15]
- 実世界参照による精度向上 [Kameko+ EMNLP15]
- 言語資源追加による精度向上の加法性 [LREC14][LRE16]

単語分割・タグ推定

形態素解析の利用

- 確率的タグ付与コーパスからの言語モデル構築 [ICSLP04] [Kurata+ ICASSP07][Sasada+ IS08][Takahashi+ EMNLP15]
- 仮名漢字変換 [情処論99][COLING06][Tokunaga+ WTIM11][Maeta+ WTIM12] [Okuno+ WTIM12][Takahashi+ EMNLP15]
- 音声認識 [Kurata+ ICASSP07][Sasada+ IS08][Neubig+ CSL12][Hirayama+ ASLP15]
- 機械翻訳 [Neubig+ EMNLP12][Sudoh+ JNLP14][Sudoh+ AMTA14]
- 言語理解 [CWC12][Maeta+ IWPT15]
- 計算機の思考の言語化 [Kameko+ CIG15]
- 動画からの手順書生成 [Ushiku+ IJCNLP17]
- 非言語データの言語による検索 [Ushiku+, SIGIR17]
- 理解に基づくレシピ検索 [Yamakata+ JSAI16]

基礎的言語処理を
自前のツールでカバー

1. 単語分割・品詞推定・読み推定 [LREC10, ACL11, IS11]

↓ 彼/名詞 と/助詞 清水/名詞 寺/名詞 に/助詞 行く/動詞

2. 用語認識／固有表現認識 [PACLING15, ACL16]

↓ 彼/名詞 と/助詞 清水/名詞 寺/名詞 に/助詞 行く/動詞

3. 係り受け解析 [IJCNLP11, NLP12, LREC14, IWPT15]

↓ 彼/名詞 と/助詞 清水/名詞 寺/名詞 に/助詞 行く/動詞

4. 共参照解析 or Wikification [IJCNLP13, LREC16]

彼/名詞 と/助詞 清水/名詞 寺/名詞 に/助詞 行く/動詞

= **バラク・オバマ** = **音羽山清水寺**

基礎的言語処理を自前のツールでカバー

1. 単語分割・品詞推定・読み推定: KyTea (きゅーてぃー)
   - **約100,000文** (BCCWJ, Twitter, レシピ, 将棋解説, …)

↓

   - 超短単位 (国立国語研究所+)

2. 用語認識／固有表現認識: PWNER (→ bi-LSTM?)

↓

   - 一般分野 10,000文, レシピ 4,000文, 将棋解説 3,000文

3. 係り受け解析: EDA

↓

   - 約55,000文

4. Wikification
   - 約3,000文

形態素解析

1. 単語分割
2. 品詞推定
3. 他のタグ推定 (読み, 原型, 正規表記, 意味タグ, ...)

INPUT: 吾輩は猫である。
OUTPUT:

我輩	ワガハイ	ワガハイ	我が輩	代名詞		
は	ワ	ハ	は	助詞-系助詞		
猫	ネコ	ネコ	猫	名詞-普通名詞-一般		
で	デ	ダ	助動詞	助動詞-ダ	連用形-一般	
ある	アル	アル	有る	動詞-非自立可能	五段- ラ行	終止形-一般
。	。			補助記号-句点		
EOS						

隠れセミマルコフモデルに基づく教師なし完全形態素解析
内海 慶, 塚原 裕史, 持橋 大地, NLP2015

形態素解析手法の歴史

- 人手設定の品詞接続コスト (JUMAN)

- 品詞 3-gram モデル *[Nagata COLING94]*
  - 未知語モデルあり
- 形態素 3-gram モデル [自然言語処理98]
  - 未知語モデルあり

いわゆる Joint Model

- CRF *[Kudo+ EMNLP04]* (MeCab)
  - 未知語モデルなし → 辞書を充実する方針

- 文字単位の点予測 [Neubig+ LREC10][Neubig+ ACL11][IS11] (KyTea)
  - 未知語モデル不要 → 学習外ドメインに強い
  - 必要最小限の曖昧性解消

形態素解析の用途

- 検索
- 文生成
  - 機械翻訳, 動画の説明, 計算機の思考の言語化 [Ushiku+, IJCNLP17] [Kameko+, CIG15]
    - 品詞・読み不要!?
  - 音声認識, 仮名漢字変換 [NL98]
    - 品詞不要!
- 後段の言語処理
  - 音声合成, 係り受け解析, …

現在の自然言語処理

- 機械学習によるアプローチ
  1. アノテーション基準の策定
  2. 言語資源の構築 (**人手によるアノテーション**)
  3. 分類器の学習・利用

- 一般分野での高い精度
  - 十分に大きなコーパスがあるから

- **応用分野では不十分な精度 (Twitter, レシピ, 医療)**
  - 「とにかく高い精度を実現しろ！」という声が

形態素解析の精度向上

- 手法の改良
  - 向上幅が小さい (例: 98.1% → 98.3%)
  - 国際学会に通らなければ実装者は報われない

- 言語資源追加 → 効果絶大
  - 95% → 98%!! (レシピの場合)
  - 不要なアノテーションによる無意味なコスト増の回避

    例) 言語モデル作成のための品詞付与
  - 副作用を避けつつ確実に問題となる誤解析を解消

    例) リリカルなのは, この先生きのこれるか

cookpad

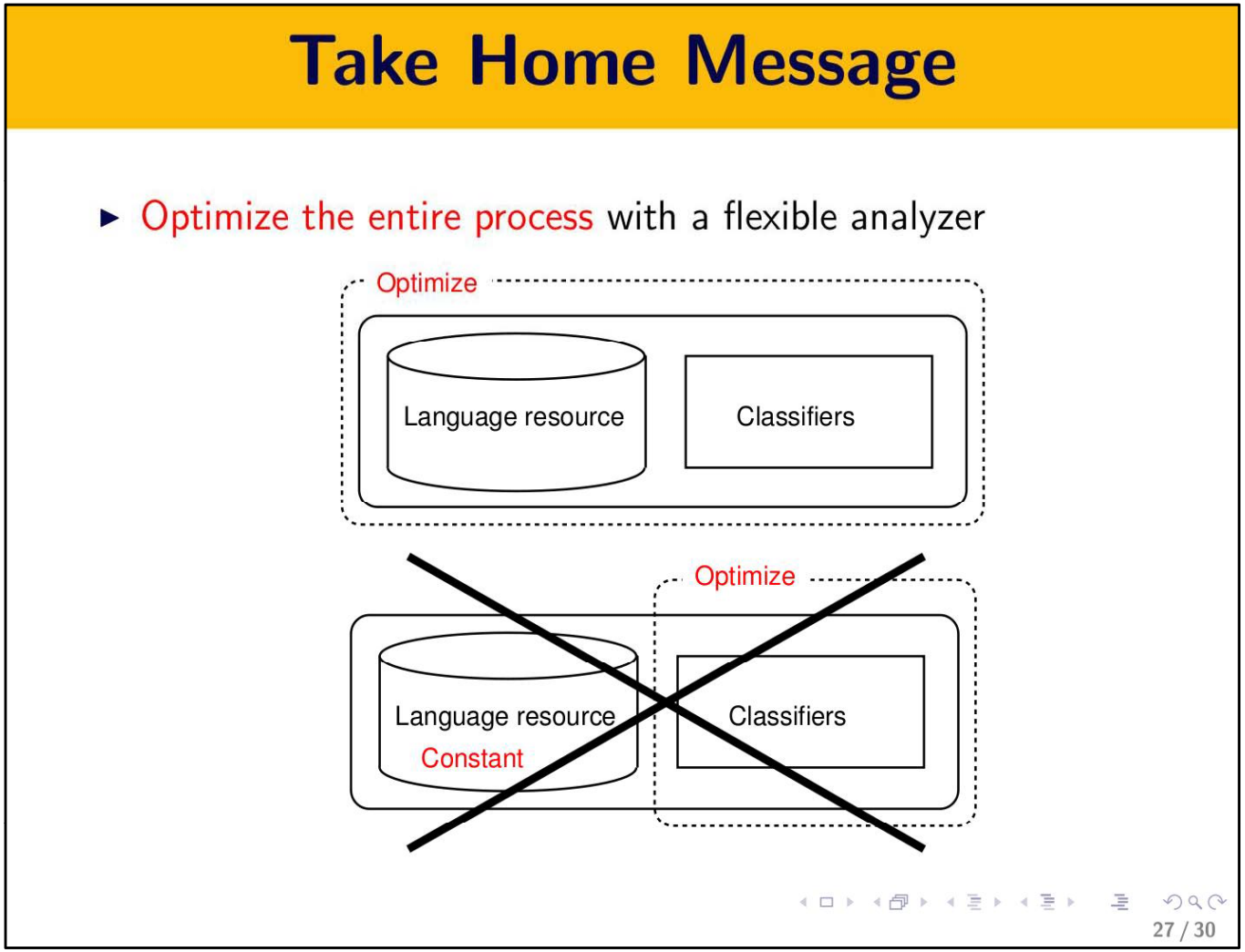
- これまでの蓄積が生きることも
  - 契約書の条文 99%

LegalForce

言語資源構築も含めて最適化

- [LREC2014] 座長にうけたスライド

- 機械学習の要件
  1. Simple
  2. Fast



辞書 v.s. コーパス [LREC14]

- 辞書追加の際に文脈があると有利か

- 単語分割, テスト: Y!Blog, 学習: BCCWJ-core
  1. BCCWJ-core only 95.54
  2. + 学習 Y!Blog にのみ出現する単語 (文脈なし) 96.75
  3. + 学習 Y!Blog にのみ出現する単語 (文脈あり) 97.15

- 文脈なしはありの75%の効果

- 文脈は捨てないように!!

この方/ほう が…
この方/かた は…

言語資源追加の効果は加法的 [LRE16]

Test domain	General	Recipe	Patent	Twitter
#sentences	3,680	724	500	542
Adaptation method	–	r-NE $n = 8$	KWIC 12 hours	KWIC 47 hours
No adapatation	99.01	95.56	96.15	85.90
Adaptation to				
recipe	99.01	96.28	96.56	85.99
patent	99.02	95.54	96.63	85.94
twitter	99.01	95.73	96.22	87.63
all	99.01	96.35	97.05	88.35

- r-NE: 分野特有の用語にアノテーション
- KWIC: Keyword In Context 形式の提示・修正 [COLING96]
- 各分野への適応が他の分野にプラスに作用 or 中立

配布モデルは1種類でOK

日本語処理の再設計

1. 正規化 NFKC
2. 文認識
3. 単語認識 (tokenization for En, Fr, etc.)
   - 各文字間に対する2値分類
4. 単語のラベル (品詞, 読み, etc.) 推定
   - 各単語に対する多値分類

5. 後段の NLP (係り受け解析 etc.), 検索, 言語モデル

KyTea 設計思想 ～機能～

- 様々な言語資源に対応 → 高い実用性
  - フルアノテーション
  - 部分的アノテーション (文の一部にのみアノテーションあり)
  - 単語と一部のタグ (例: Brexit//ぶれぐじっと)
  - 複合語 (両端だけが分割基準に合致, 例: 外国人参政権)
- 確率的単語分割・タグ付与 [ICSLP04]
- 配布モデルの学習コーパスを持っているのと同じ再学習

1. 配布モデルの素性頻度ファイルのダウンロード

   % wget http://www.ar.media.kyoto-u.ac.jp/tool/.../max-model.kff

2. 素性頻度ファイルに独自言語資源を追加して学習

   % train-kytea –feat max-model.kff –dict (独自辞書) –corpus (独自コーパス)

言語資源の形態

- フルアノテーション
  - 文全体（全ての単語）にアノテーション

吾輩/名詞 は/助詞 猫/名詞 で/助動詞 あ/動詞 る/語尾

- 部分的アノテーション
  - 文の一部にのみアノテーション (文脈情報あり)

吾輩は 猫/名詞 である

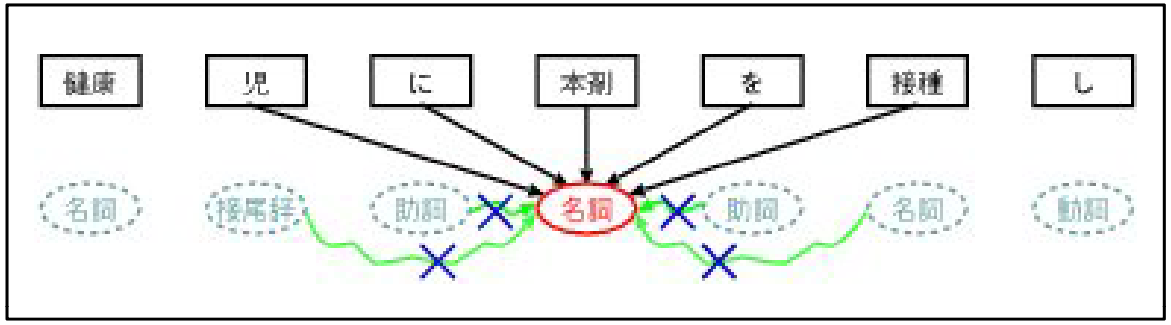
- 辞書
  - 表記とタグ (文脈情報なし)

猫, 猫/名詞, 猫//ねこ, 猫/名詞/ねこ

KyTea 設計思想 ～解法～

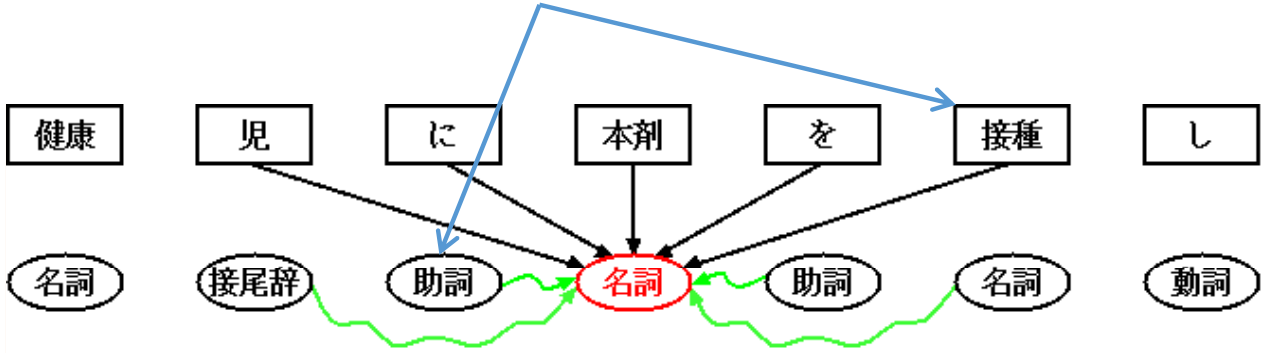
- 問題を最小単位に分割
  - 文字間の単語境界 (2値分類)
  - 単語の品詞・読み・原型・その他タグ (多値分類)

  例) 単語分割に品詞は不要
- 推定値を参照しない → 様々な言語資源に容易に対応可能
  - 部分的アノテーション, 不完全な辞書



KyTea 設計思想 ～解法～

- 問題を最小単位に分割
  - 文字間の単語境界 (2値分類)
  - 単語の品詞・読み・原型・その他タグ (多値分類)
- 様々な言語資源に対応 ← 推定値を参照しない
  - 系列予測



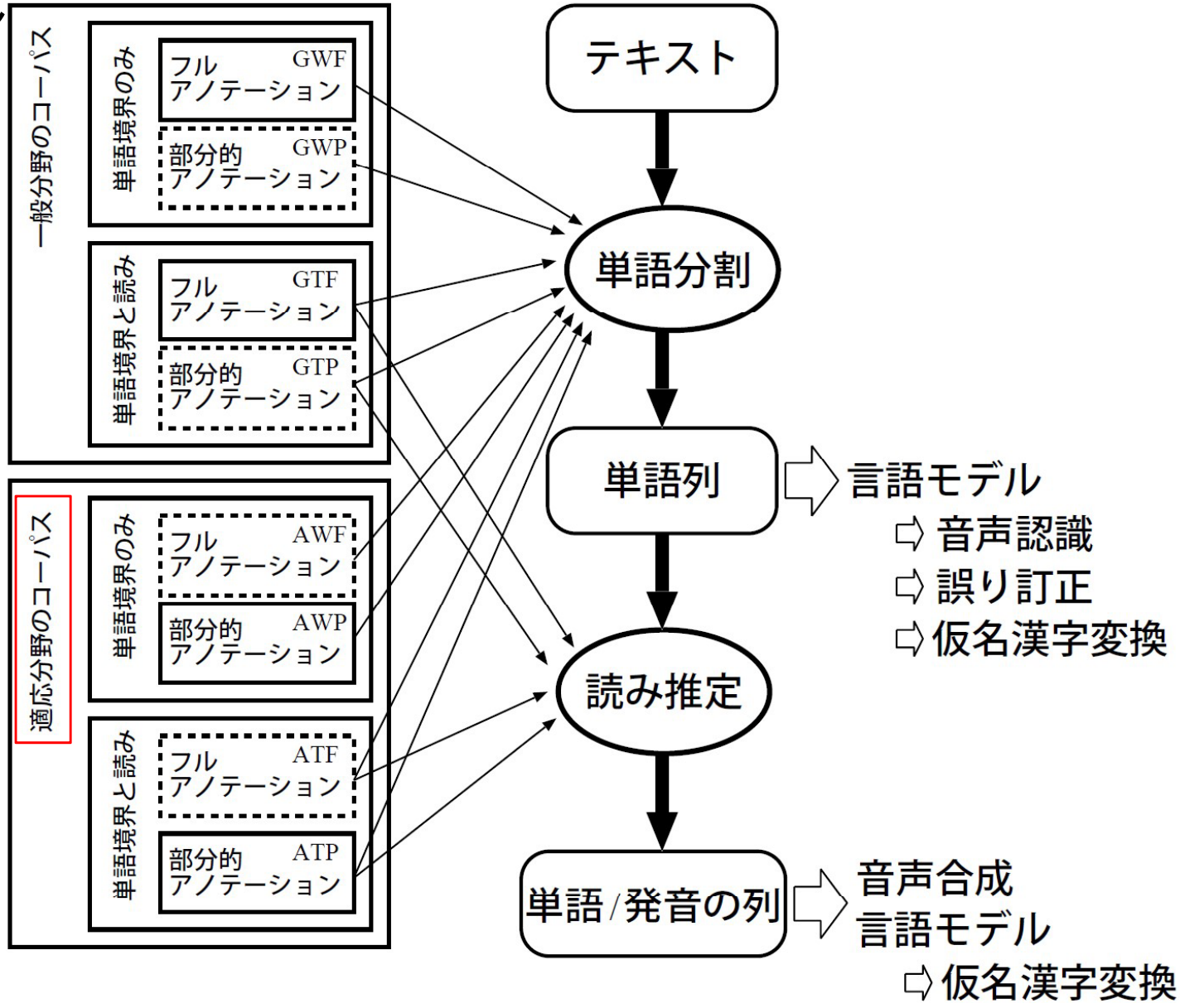
  - 点予測

必要なときに必要なだけの曖昧性解消

例) 言語モデルの構築

京都大学
森研究室

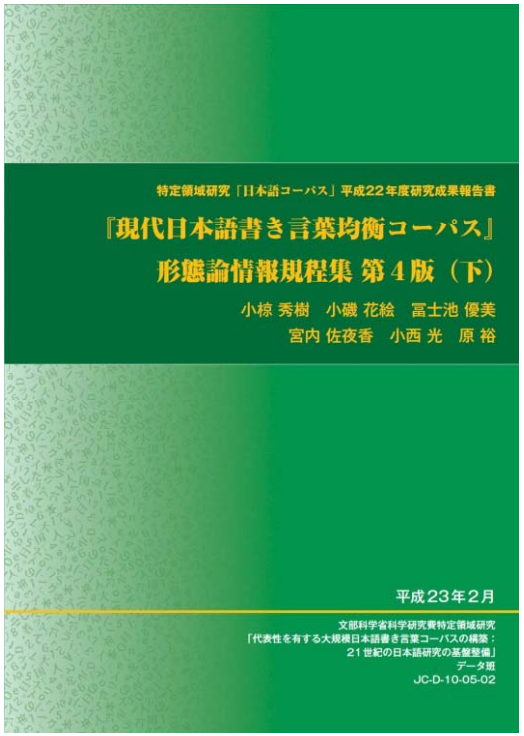
ユーザー



KyTea デフォルトモデル 超短単位 SSUW

- 解析器と基準は無関係!!

- 基本的には短単位 (国語研 KOTONOHAプロジェクト)
  - 十分に充実した基準書
  - 6万文のタグ付きコーパス



- 生成＆グラウンディングを意識した単位
  - 活用語を語幹と語尾に分割 (|V|x5 v.s. |V|+5)
    - 語幹: 内容に対応 (実世界・他言語)
    - 語尾: 純粋に文法的要素

ex.) 久しぶり | に | 会 | っ | た | 友達 | と | 飲み歩 | く

KyTea のこれから

- 継続的な言語資源拡充 (あるいは精度向上)
  - 国立国語研究所: UniDic, 話し言葉
  - 京都大学 (今後約20年間): 共同研究で様々な分野への適応
- 無意識のクラウドからの学習の実運用
  - 音声から [Sasada+ InterSpeech08]
  - 仮名漢字変換から [Takahashi+ EMNLP15]
- 実装の改良 (高速化, 少メモリ, etc.)
  - 単語分割: 96ビットから1ビットへの写像
    - 96ビット = 前後3文字 x16ビット (EUC, SJIS)
- スパコン上での超並列動作 (京大 メディアセンター)

今後も言語資源追加

- 加法的 → １つのモデルでOK!!
- 各所の日本語の処理のための言語資源を集約したい
  1. 様々な応用のために言語資源の構築
  2. 言語資源を譲ってもらう
  3. その言語資源を追加し学習
  4. モデルを配布
  5. 最初にもどる
- 共同研究があれば遠慮なくどうぞ
  - 予算があっても
  - なくても

FIN

点予測による形態素解析 Q&A

[Q] 人は全体をみて判断しているのでは？
　[A] 解ければいいです

[Q] ラベル間の依存を無視していいのか
　[A] 形態素解析・係り受けなどの問題分割は？
　[A] 逆に曖昧性解消を最小単位まで分割すれば？

[Q] では全体最適はだめか
　[A] 計算時間がかかると思うけどそれでもいい
　[A] 解ければなんでもいい
　[A] 要は部分的アノテーションで楽をしたい

部分的アノテーションの活用

- 部分的アノテーション

  頑　健　な | 形 - 態　素 | 解　析　を

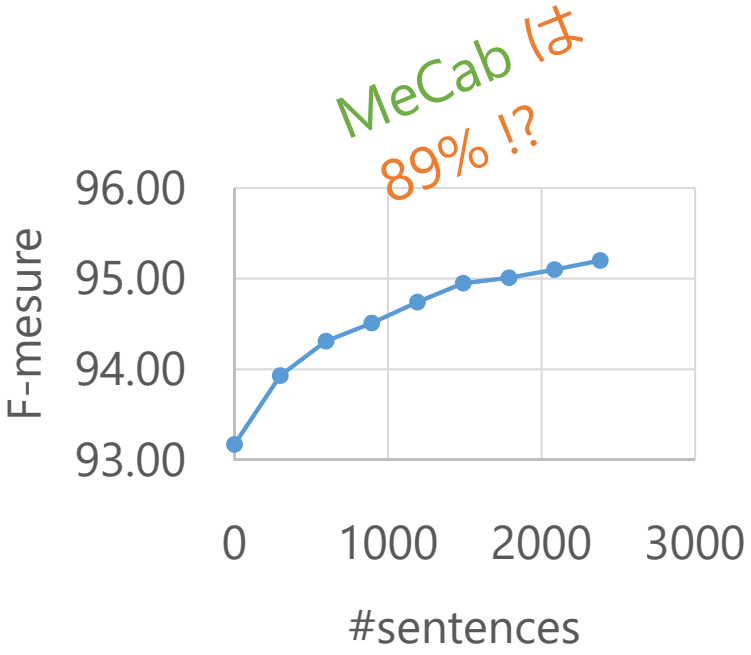
  あり　なし　不明

- アノテーション単位の最小化
  - 部分的アノテーションからの CRFs の学習
  - [Tsuboi+, COLING08] (フリー実装あり)

- 解析単位の最小化
  - Pointwise solution (SVM, LR)
  - [Neubig+, LREC10], [Neubig+, ACL11]

多様な実テキストに強い手法

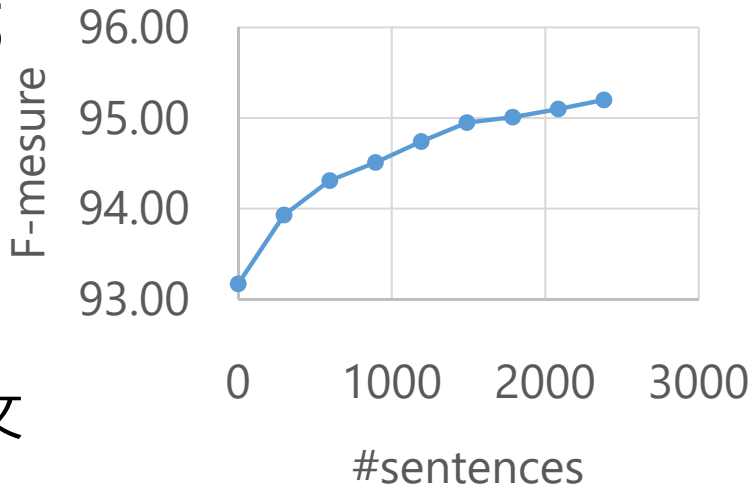
- テキスト解析器 KyTea http://www.phontron.com/kytea/
  - 未知語に対して頑健
  - Cf. JUMAN: 基本語彙に豊富な情報
  - Cf. MeCab: Web 企業での利用多数
- 分野適応が容易
  - Twitter の形態素解析で最高精度!?
    - JUMAN, MeCab に対してデフォルトで優位
    - 言語資源の追加でさらに精度向上



- 医薬・特許・レシピ・学内アナウンス・ガーデニング・将棋
  - 言語資源の追加は加法的な効果 [LRE16]

それでも適応は必要

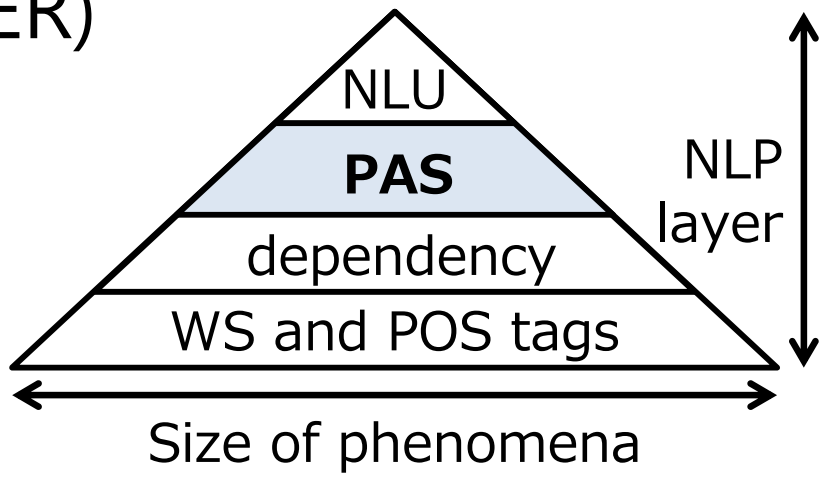
- BCCWJ-core (新聞・雑誌・書籍・白書・Y!知恵・Y!QA)
- Twitter: 単語分割F値 0.93 → 0.95
  - フルアノテーション: 約2,500文
- レシピ: 単語分割F値 0.95 → 0.98
  - 部分的アノテーション: 9/10 x 1,303文
- 将棋の解説: 単語分割F値 0.92 → 0.98
  - 辞書 (7,209指し手表現, 532棋士名)
  - フルアノテーション: 3,299文, 24,966単語
  - 部分的アノテーション: 3,000文 (1 s-NE for each sent.)



言語処理のデザイン

- 前段を仮定しない
  - 対象分野での精度低下は不可避
  - 前段のためのアノテーションを回避

- 単語列から直接解析 (品詞は無しか自動推定)
  - 単語列 ⇒ 固有表現認識 (NER)
  - 単語列 ⇒ 係り受け (Dep)
  - 単語列 ⇒ 述語項構造 (PAS)



- 部分的単語分割から直接学習

不都合な真実

- 結局アノテーションの追加は効果絶大
  - 特に精度が低い初期段階 (新たな課題・分野)
  - カバー率が重要
  - 未知語に集中的にアノテーション
    - 部分的アノテーションが便利

- しかしアノテーションの効果は対数的
  - 指数的な増量が必要
  - どこかの段階で手法改善の研究が重要に
    - 様々な形態の言語資源からのモデル学習

辞書・コーパス追加

- ひたすら手動で加える
  - NEologD *[@overlast; Twitter]*
    - 継続的努力
    - 単語分割基準を満たさない複合語 → 単語分割精度は向上しない
- 自動獲得
  1. 生コーパスから「単語/品詞」[COLING96]
  2. 自動音声と関連テキストから「単語/読み」[Kurata+ ICASSP07] [Sasada+ IS08]
  3. 仮名漢字変換ログから「単語/読み」[Takahashi+ EMNLP15]

“ついで”に形態素解析

- 形態素解析自体が目的ではない
- NERの高精度化のついでに [LREC14]
  - NERのアノテーションのついでに単語境界も付与

各	/ホット	ドッグ/F	に	チリ	、	チーズ	、		
(each)	(hot	dog)	(*cmi*)	(chili)	,	cheese	,		
オニオン	を	ふりかけ	る						
onion	(*cmd*)	(sprinkle)	(*infl.*)						
/ホット	ドッグ/F	を	アルミ	ホイル	で	覆	う		
(hot	dog)	(*cmd*)	(aluminum	foil)	(by)	(cover)	(*infl.*)		

各|ホ−ッ−ト|ド−ッ−グ|に␣チ␣リ␣、…,
|ホ−ッ−ト|ド−ッ−グ|を␣ア␣ル␣ミ…,

“|”: boundary, “−”: not boundary, “␣”: no information

生コーパスからの単語・品詞の獲得

[COLING96]

1. KWIC (Keyword in context) の計算
   - 「楽し」の例

げながらのやりとりが	楽し	い。いったい今、世界
画をほうふつとさせて	楽し	い。これほどさまにな
とに変えていくのは、	楽し	い。経済のしくみを大
っている方がはるかに	楽し	い。建具卸売業を経て
食事をする時はとても	楽し	い。今、特に子供が1
ルだ。遊覧船の復活は	楽し	い。水の上のにぎわい
を店に支払う。踊って	楽し	いだけでなく、脚と足
もかかわらず、食事を	楽し	いものと感じていなか
とに、決まった筆者の	楽し	いコラムが登場します
。私は戦争のために、	楽し	い温かい家庭生活を味
市民にとって水族館は	楽し	い教室であり、憩いの
の大好きなお祭りだ。	楽し	い催しがあれば、数多
酒を飲んでもネアカな	楽し	い酒だし、趣味も豊か
日本には、まだ走って	楽し	い道路がない。サイク
した。山登りを趣味に	楽し	い老後を過ごしている
。炭鉱が消えてから、	楽し	い話題が少なかった。
分の陳列の場に戻る。	楽し	かった2匹は、冬の氷
相好が崩れ走っていて	楽し	かったことの1つに、
新し、消費者が気軽に	楽し	く、新型車の情報を得
年会を、ゆったりと、	楽し	くやるコツを知ってい

生コーパスからの単語・品詞の獲得

[COLING96]

2. 左右の文字分布 ***D*** に変換

– 「楽し」の例

頻度	確率	文字		文字	頻度	確率
13	6.8%	、	楽し	い	16	8.3%
6	3.1%	。		か	2	1.0%
2	1.0%	う		く	3	1.6%
13	6.8%	が		げ	4	2.1%
10	5.2%	て		さ	8	4.2%
8	4.2%	で		そ	10	5.2%
4	2.1%	と		て	1	0.5%
1	0.5%	ど		ま	7	3.6%
2	1.0%	な		み	43	22.4%
14	7.3%	に		む	38	19.8%
19	9.9%	の		め	16	8.3%
4	2.1%	は		も	4	2.1%
7	3.6%	も		ん	40	20.8%
2	1.0%	ら				
2	1.0%	り				
1	0.5%	ろ				
82	42.7%	を				
1	0.5%	折				
1	0.5%	変			（頻度：192）	

生コーパスからの単語・品詞の獲得

[COLING96]

3. 各品詞の左右の文字分布 ***D***(品詞) の線形和に分解

min $F(\boldsymbol{p})$, where

$$F(\boldsymbol{p}) = \left| \boldsymbol{D}(\alpha) - \sum_k p_k \boldsymbol{D}(pos_k) \right|^2$$

subject to $\sum p_k = 1$

– ***D***(楽し) = 0.1 x ***D*** (動詞) + 0.9 x ***D*** (形容詞)

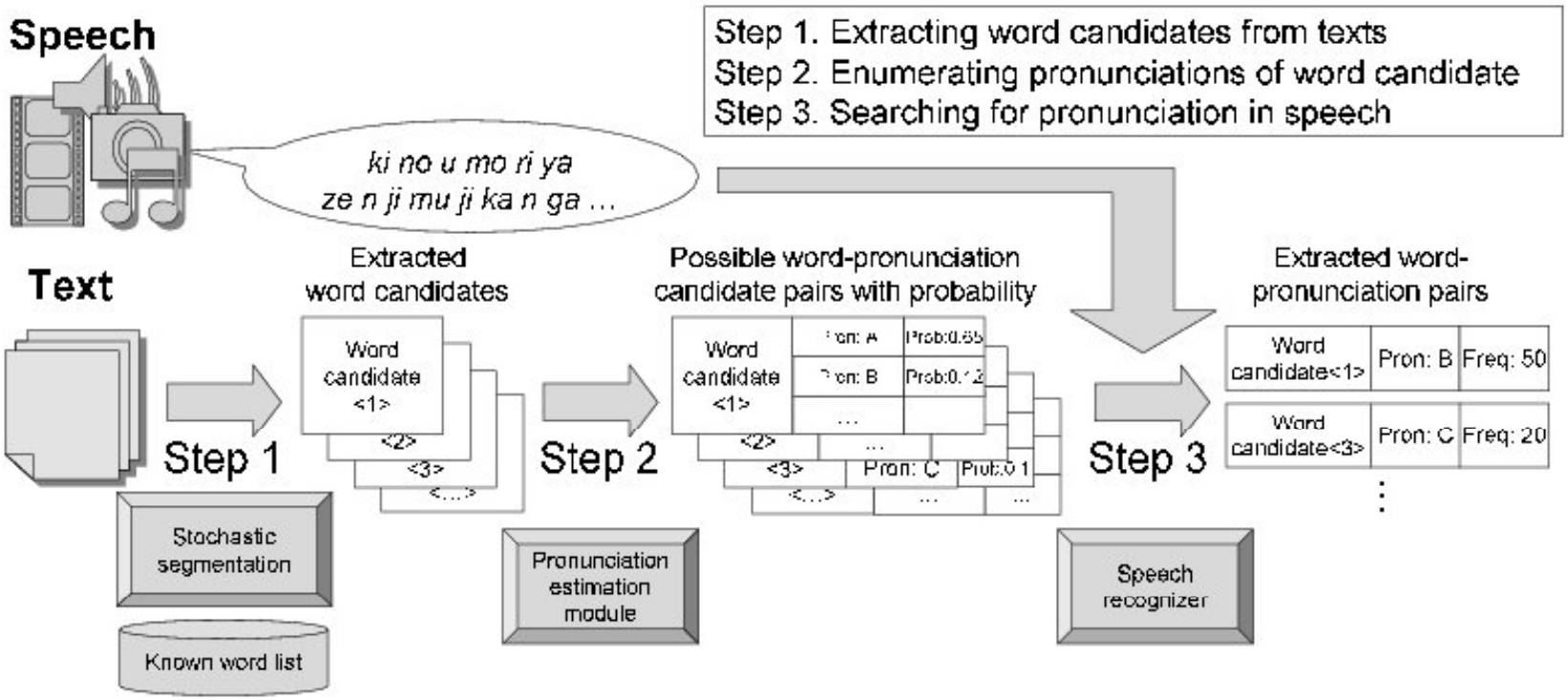
生コーパスからの単語・品詞の獲得
[COLING96]

- EDRコーパスで学習
- 日経サイエンスの形態素解析実験
  - 95.9% → 98.2%

- UniDic のカバレージが高くてここまでの差は出ないが
- 現在も未知語収集に利用
  - 最適化は LR
  - 抽出された未知語候補を人手で修正 → 部分的アノテーション

音声・テキストから「単語/読み」
[Kurata+ ICASSP07] [Sasada+ IS08]

1. 確率的単語分割
2. 確率的読み推定
3. 言語モデル作成 & 音声認識



音声・テキストから「単語/読み」
[Kurata+ ICASSP07] [Sasada+ IS08]

1. 擬似確率的タグ付与コーパス [Mori+, ICSLP04]

・確率的単語分割 (Cf. KyTea LR 版)

	横	｜	ア	–	リ	｜	い	｜	っ	｜	た
単語分割確率：	0.89		0.01		0.99		0.97		0.99		

・確率的読み推定 (Cf. KyTea LR 版)

	横アリ/**[よこあり, おうあり]**	
読み確率：	0.67	0.33

➡ 擬似的に出現率を反映して未知語がアノテーションされる

擬似確率的単語分割コーパスの例

昨日｜横｜アリ｜い｜っ｜た

横アリ｜って｜近｜い｜？

「横アリ/よこあり」という未知語候補が仮名漢字変換の変換候補に提示

音声・テキストから「単語/読み」
[Kurata+ ICASSP07] [Sasada+ IS08]

- LVCSR [Kurata ICASSP07]
  - 放送大学の講義
  - テキストの追加
  - CER: 26.1% → 9.7%

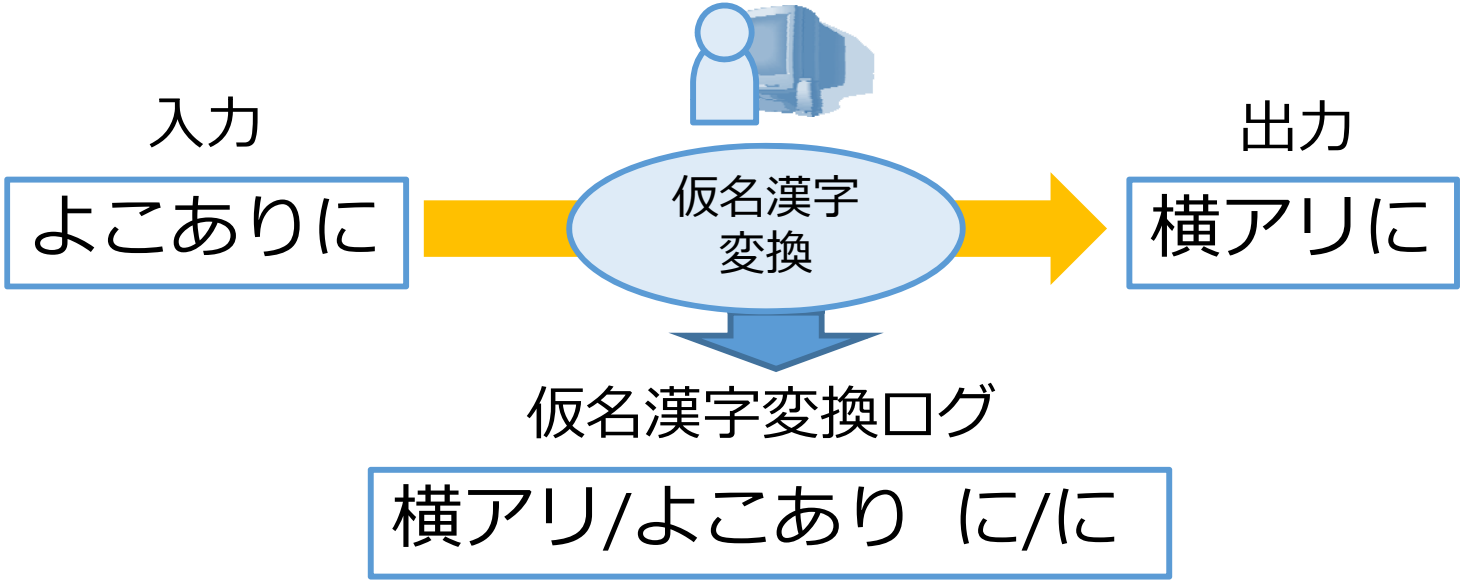
CER: Character Error Rate

- 読み推定 [Sasada+ IS08]
  - ベースライン: 日経新聞 + 英語表現辞典 (52,955文)で学習
  - 30分 x 34日のテレビニュース + 新聞等 3.7M 文
  - F値: 99.29 → 99.36

仮名漢字変換ログからの学習
[Takahashi+ EMNLP15]

- 仮名漢字変換ログ

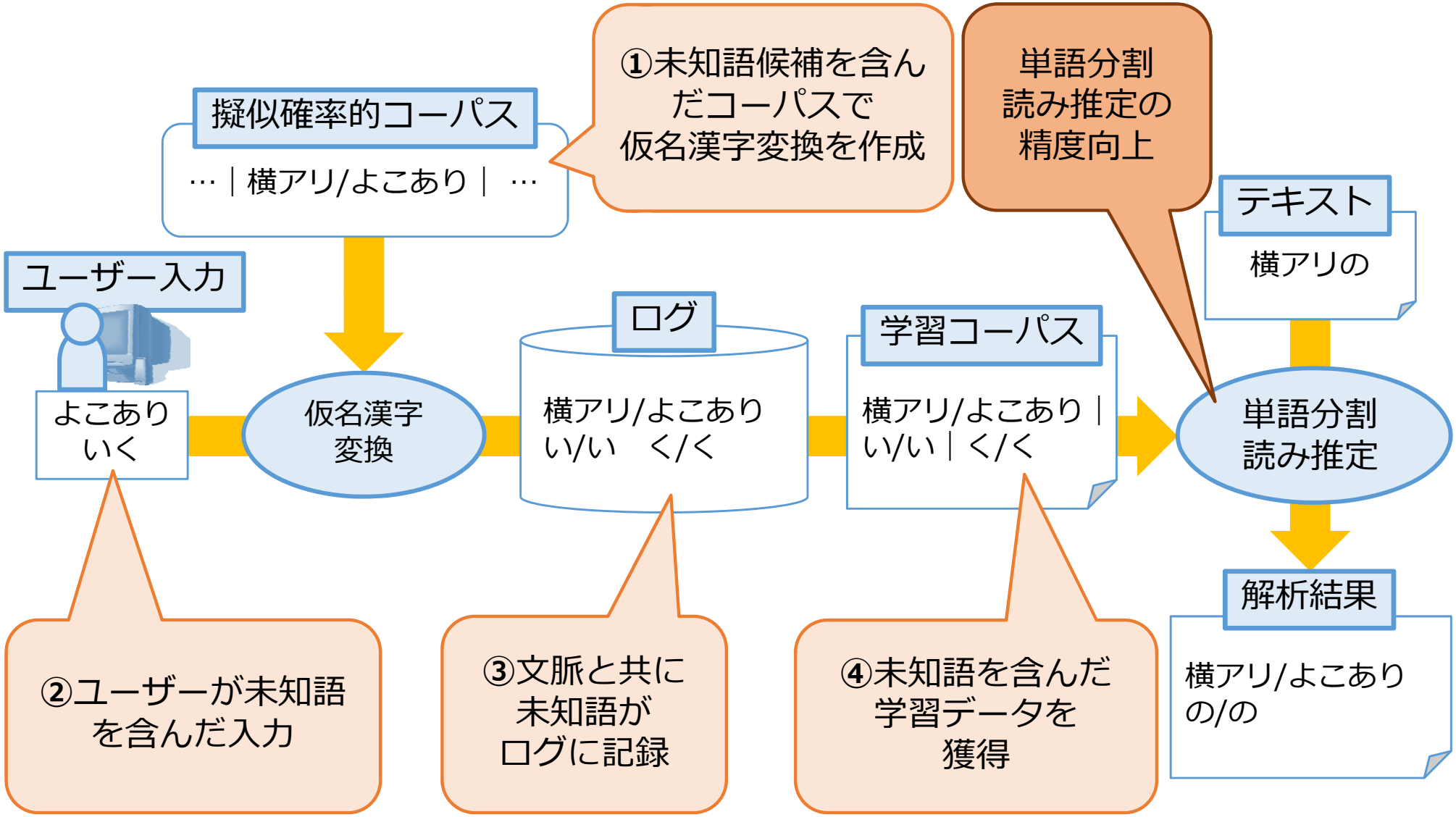
  仮名漢字変換の際の履歴情報



単語境界と読み情報が付与された文の断片

➡ 単語分割・読み推定の学習データに利用

仮名漢字変換ログからの学習
[Takahashi+ EMNLP15]

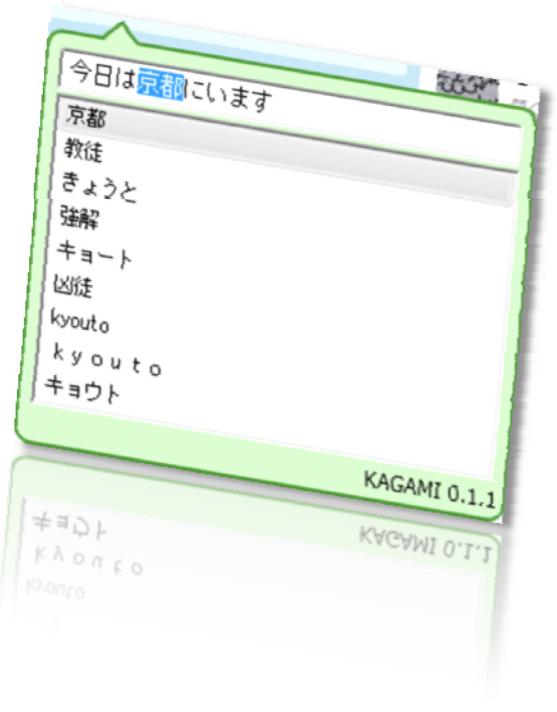


ログを収集するインプットメソッド
- KAGAMI -

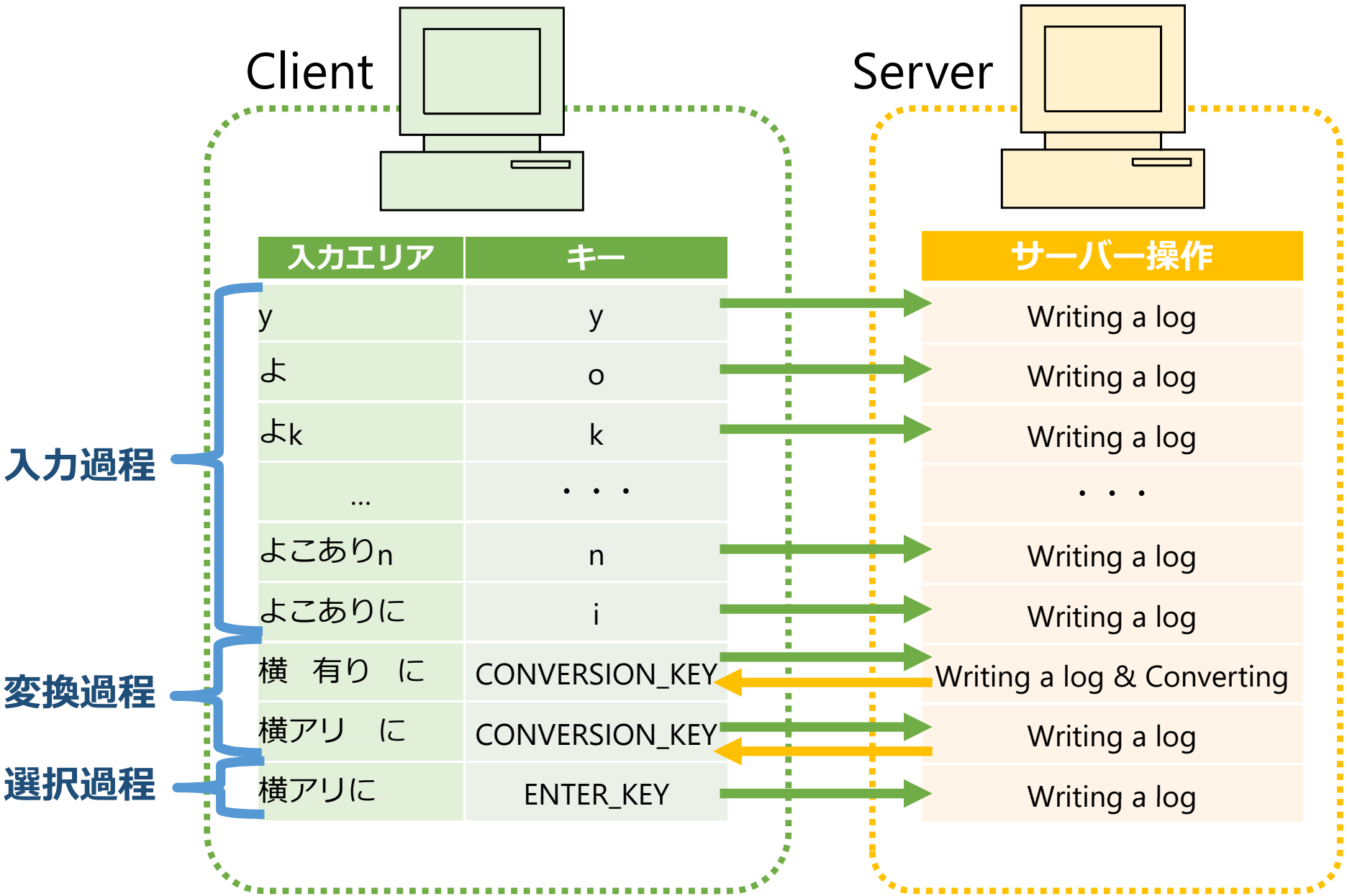
ログをサーバーに送信するインプットメソッドを開発

- Firefox アドオンとして提供
- ホームページにて配布中
- サーバーサイド仮名漢字変換

  逐次的にサーバーにログを送信



インプットメソッドの動作



仮名漢字変換ログの利用

AS-IS-log：
変換結果をそのまま部分的アノテーションコーパスとして用いる

AS-IS-log の例
横–ア–リ | に
く–ら–っ | ベ–ル
比–べ | る
と
も␣の␣の
安–め | か | と

CHUNK-log：
確定時間と次の入力の開始時間との差が s 以下の場合、変換結果をチャンキング (s = 500[ms])

細分化を回避

CHUNK-log の例
横–ア–リ | に | く–ら–っ | ベ–ル
比–べ | る | と | も␣の␣の
安–め | か | と

ALIGN-log：
ツイートに対して変換結果をアライメント

ALIGN-log の例
横–ア–リ | に | 比–べ | る | と | 安–め | か | と

細分化、確定誤りを回避

しかし、推敲の過程やつぶやかれなかったツイートのログが除外される

実験設定

一般分野テキストとツイート（未知語を多く含んだテキスト）に対して
単語分割・読み推定

・ログの収集

ユーザー：５人
収集期間：2014/04/24 – 2014/10/21

・コーパス

学習データ

記号	文数	単語数	文字数
BCCWJ-train	56,753	1,324,951	1,911,660
AS-IS-log	22,523	-	65,250
CHUNK-log	6,572	-	65,250
ALIGN-log	1,850	-	52,387

ログ由来の学習コーパス

テストデータ

記号	文数	単語数	文字数
TWI-test	2,976	37,010	58,316
BCCWJ-test	6,025	148,929	212,261

一般分野テキスト

単語分割の評価

単語分割の評価：**単語**アライメントを取り再現率・適合率・調和平均

ツイート（TWI-test）の単語分割精度

	再現率	適合率	F 値
BCCWJ-train	89.80	94.17	91.93
BCCWJ-train + AS-IS-log	90.17	94.02	92.05
BCCWJ-train + CHUNK-log	90.61	94.34	92.44
BCCWJ-train + ALIGN-log	90.12	94.23	92.13

一般分野テキスト（BCCWJ-test）の単語分割精度

	再現率	適合率	F 値
BCCWJ-train	99.01	98.97	98.99
BCCWJ-train + AS-IS-log	98.96	98.89	98.93
BCCWJ-train + CHUNK-log	99.05	98.88	98.97
BCCWJ-train + ALIGN-log	98.99	98.93	98.96

ツイート解析は困難

一般分野 99%
ツイート 92%
WERで8倍

過分割が原因

再現率＜適合率

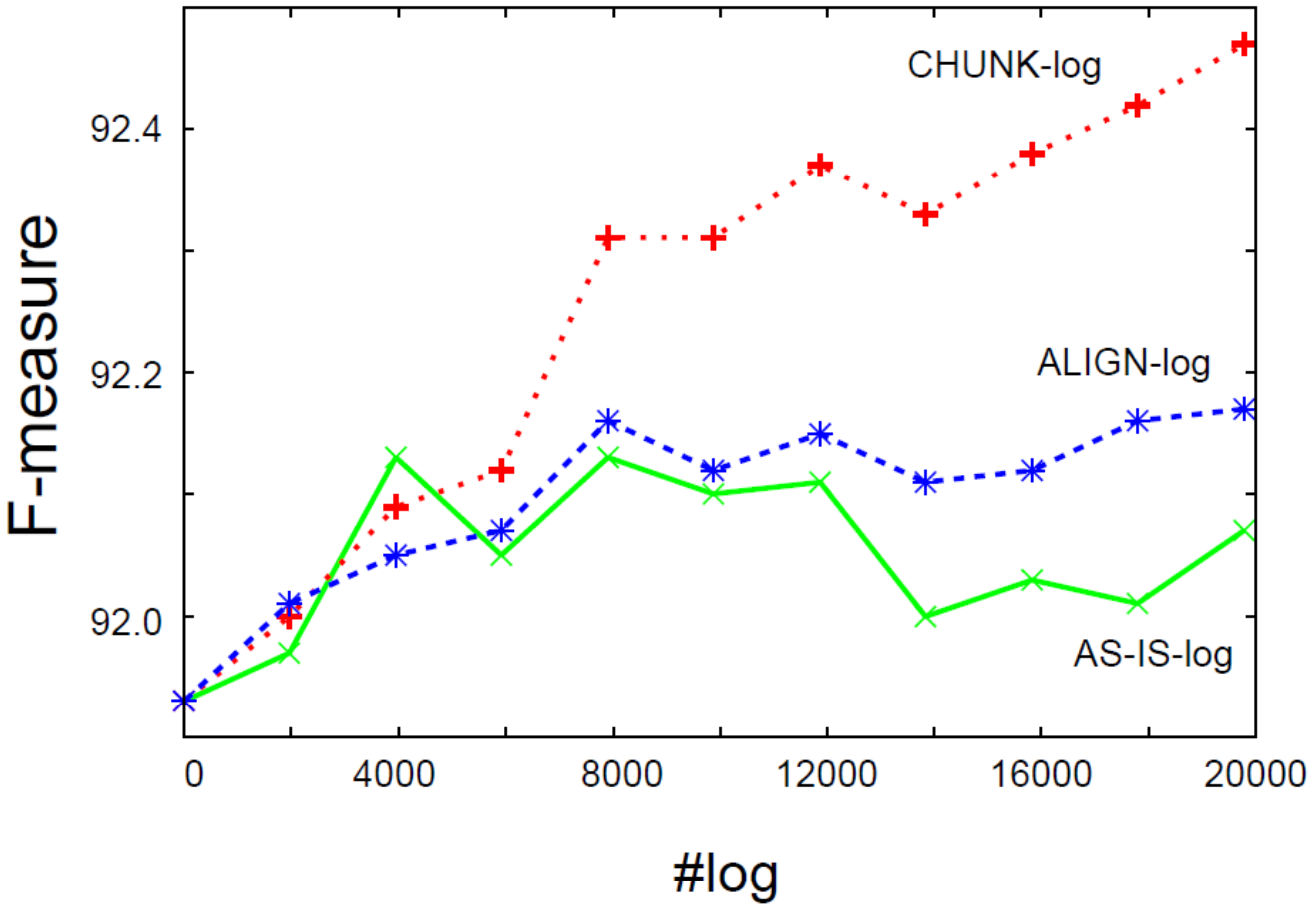
…｜横｜アリ｜…

ログの有効性

ログを追加して、精度が向上
特にCHUNK-logで有意差

ログの量と単語分割の精度

時系列順に並べたログを追加していきその単語分割精度



CHUNK-log

ログ追加でより精度向上

ALIGN-log

8000件以降から精度が微増

AS-IS-log

ノイズが多く含まれる

読み推定の評価

読み推定の評価：読みの**文字**アライメントを取り再現率・適合率・調和平均

ツイート(TWI-test)の読み推定精度

表 8　ツイートの読み推定精度

	再現率	適合率	F 値
BCCWJ-train	94.32	96.69	95.49
BCCWJ-train + AS-IS-log	94.32	96.72	95.50
BCCWJ-train + CHUNK-log	94.34	96.75	95.53
BCCWJ-train + ALIGN-log	94.34	96.73	95.52

ツイート解析は困難

一般分野 99.4%
ツイート 95.5%

一般分野テキスト（BCCWJ-test）の読み推定精度

表 9　一般分野テキストの読み推定精度

	再現率	適合率	F 値
BCCWJ-train	99.36	99.38	99.37
BCCWJ-train + AS-IS-log	99.34	99.36	99.35
BCCWJ-train + CHUNK-log	99.36	99.38	99.37
BCCWJ-train + ALIGN-log	99.37	99.38	99.38

ログの有効性

ログを追加して、
精度が向上

辞書・コーパス追加

- 人手修正は強力
- 自動獲得の効果は限定的
- 長期間運用してどうなるかは不明
  - 有志でフリーの仮名漢字変換サーバーを実装中
  - 音声認識もやってみたい