言語処理学会第２１回年次大会

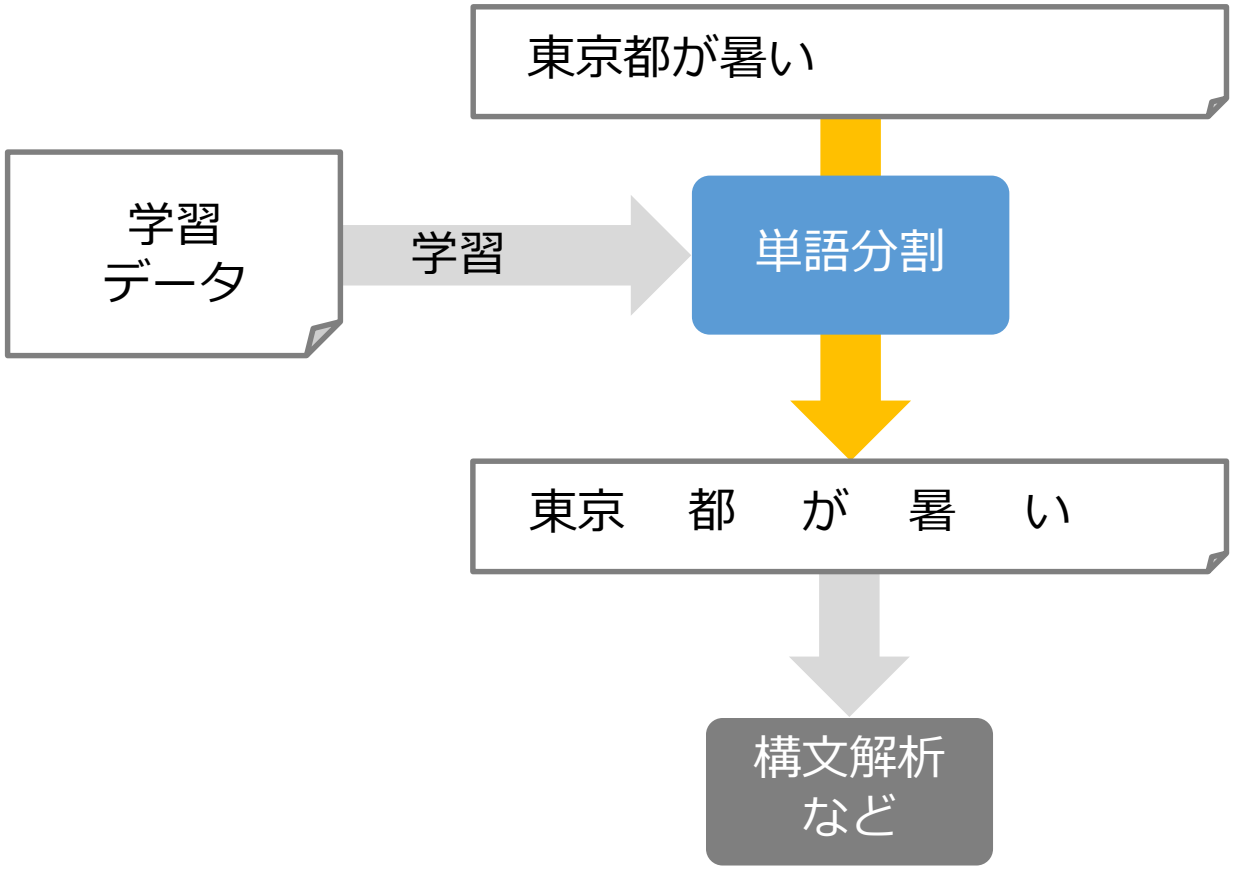
# 仮名漢字変換ログを用いた単語分割の精度向上


京都大学 大学院
修士二回　高橋 文彦
准教授　　森 信介

# 単語分割とは



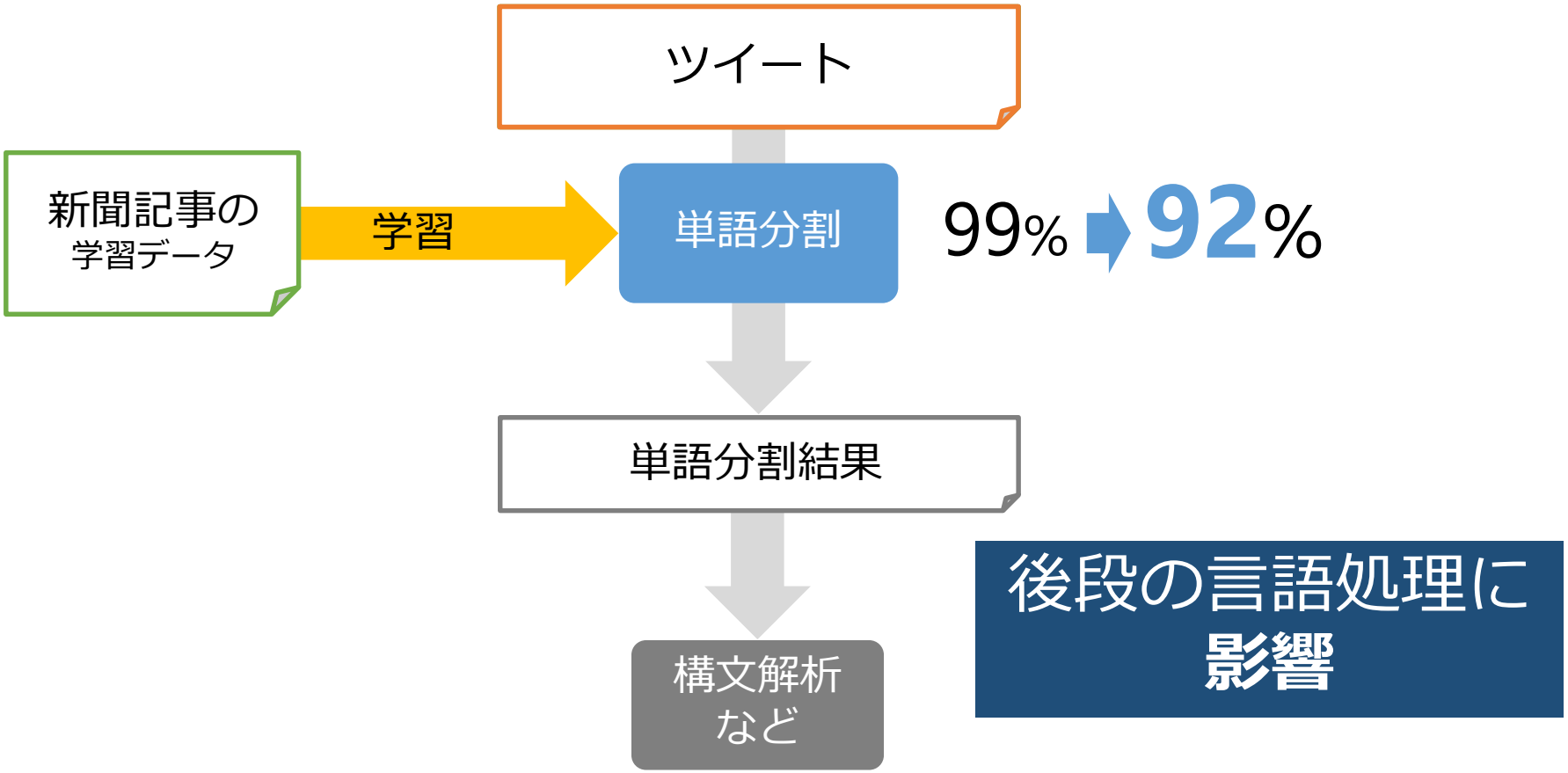
単語分割は様々な言語処理の根源的な処理



# 課題



学習データのないドメインに対して低い解析精度



# 本研究の概要



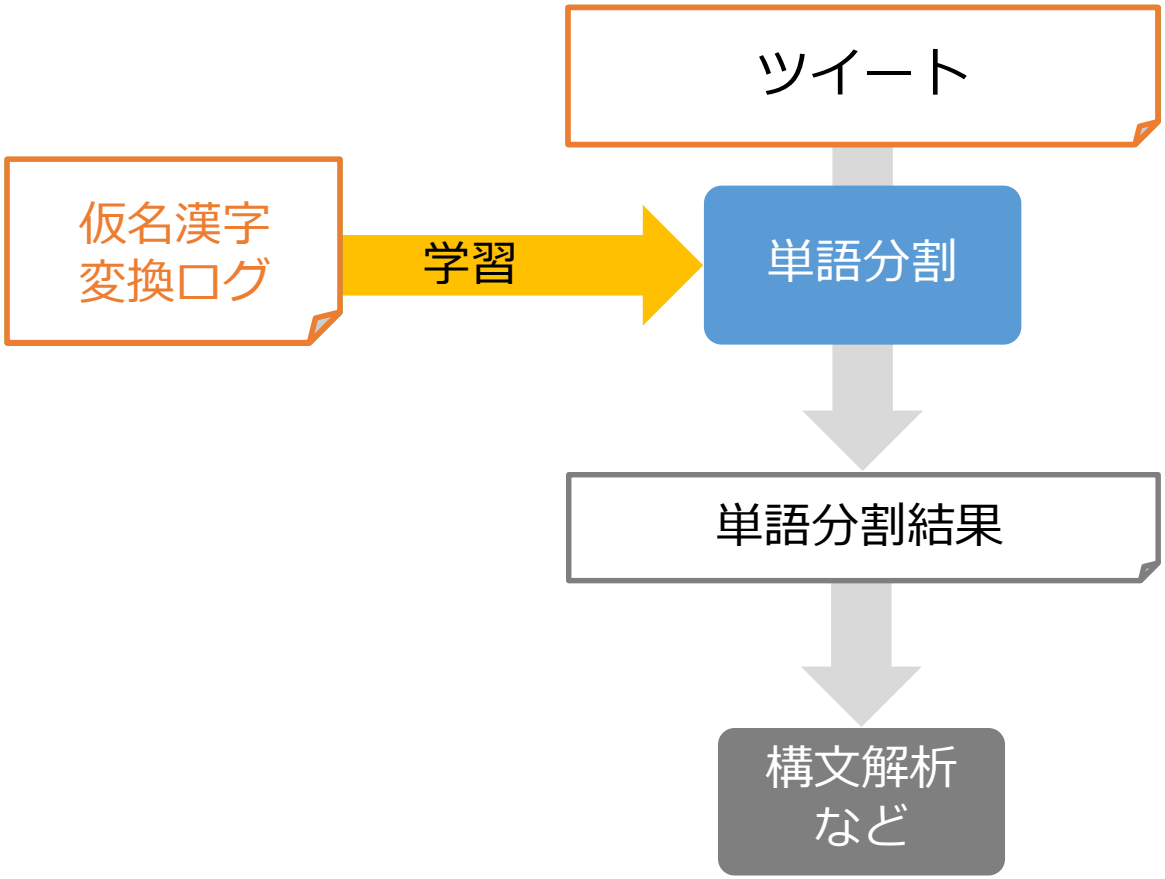

人が自然に作るデータなので**コストが低い**

➡ 単語分割の学習データに利用

# 課題



学習データのないドメインに対して低い解析精度



# 関連研究



- [森+, 1998] [Chen+, 2000]
  統計的仮名漢字変換
  語彙をコーパスの全部分文字列に拡張 [Mori+, 2006]
- [Tsuboi+, 2008] [Neubig+, 2011]
  部分的単語分割からの単語分割器の学習
- [Jiang+, 2013] **自然注釈の利用**

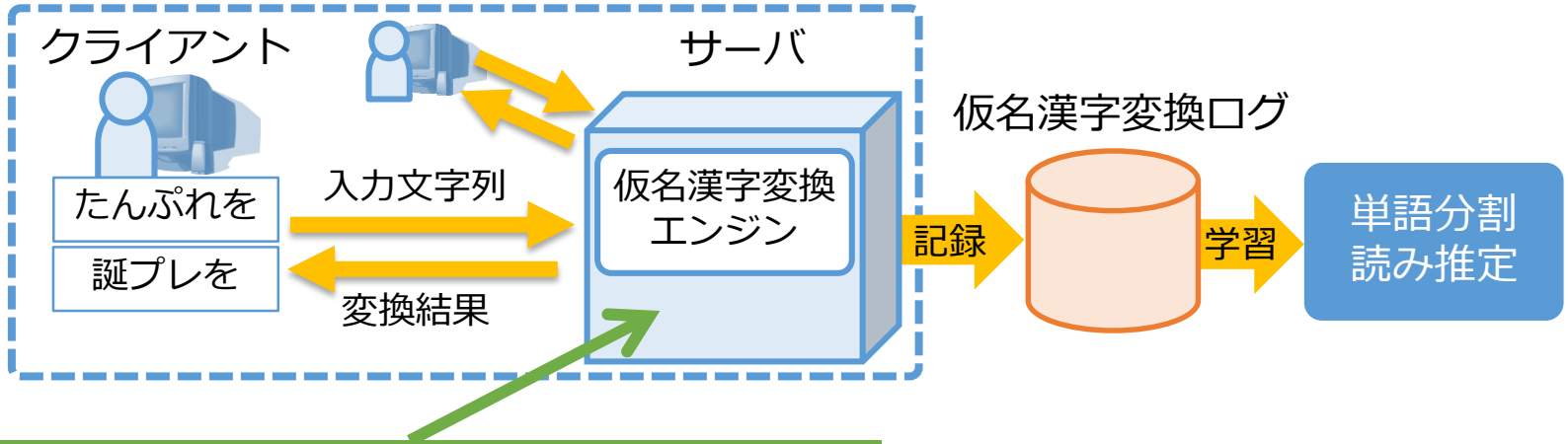
  HTMLタグを利用した単語境界情報の獲得

頑　健　な<a>形　態　素</a>解　析　を　行　う


頑　健　な | 形　態　素 | 解　析　を　行　う

# ログを収集する仮名漢字変換システム






## 擬似確率的タグ付与コーパス(PSTC)で学習

確率的単語分割 + 確率的読み付与

… の | 誕 ? プ - レ | あ - げ | る …

単語分割確率：$p = 0.8$

乱数 $r > p$：単語境界**なし**

乱数 $r < p$：単語境界**あり**

**擬似的に出現率を反映してアノテーション**

単語分割確率・読み確率：ロジスティック回帰により推定

実際には、ツイートのPSTCを作成

昨日 | 誕 | プレ | あげ | た

誕プレ | い | る | ？

「誕プレ」という未知語候補が
仮名漢字変換で提示

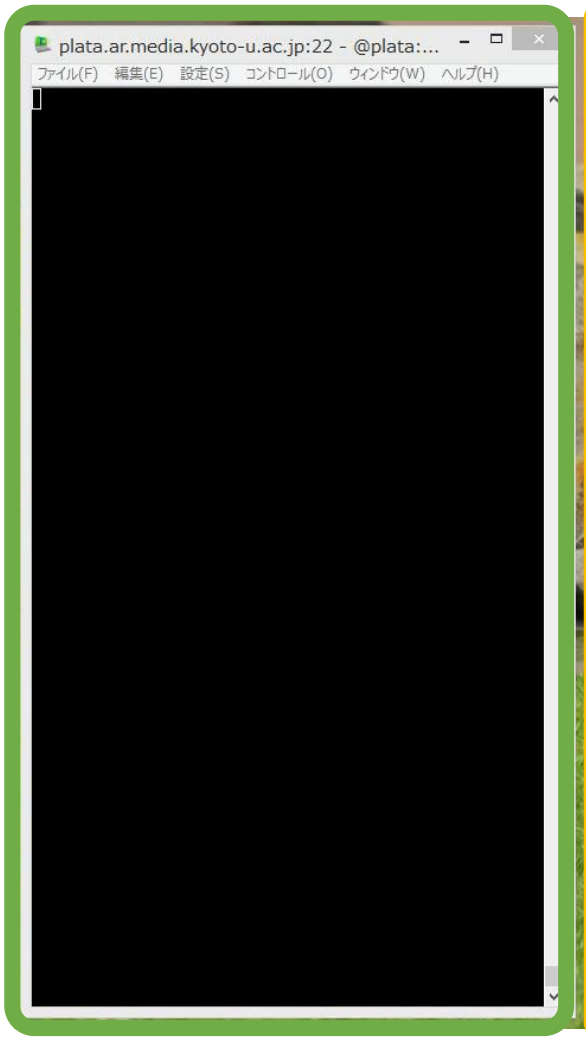
ツイートの
**文脈**と**未知語**がログに記録

# 仮名漢字変換システムのデモ

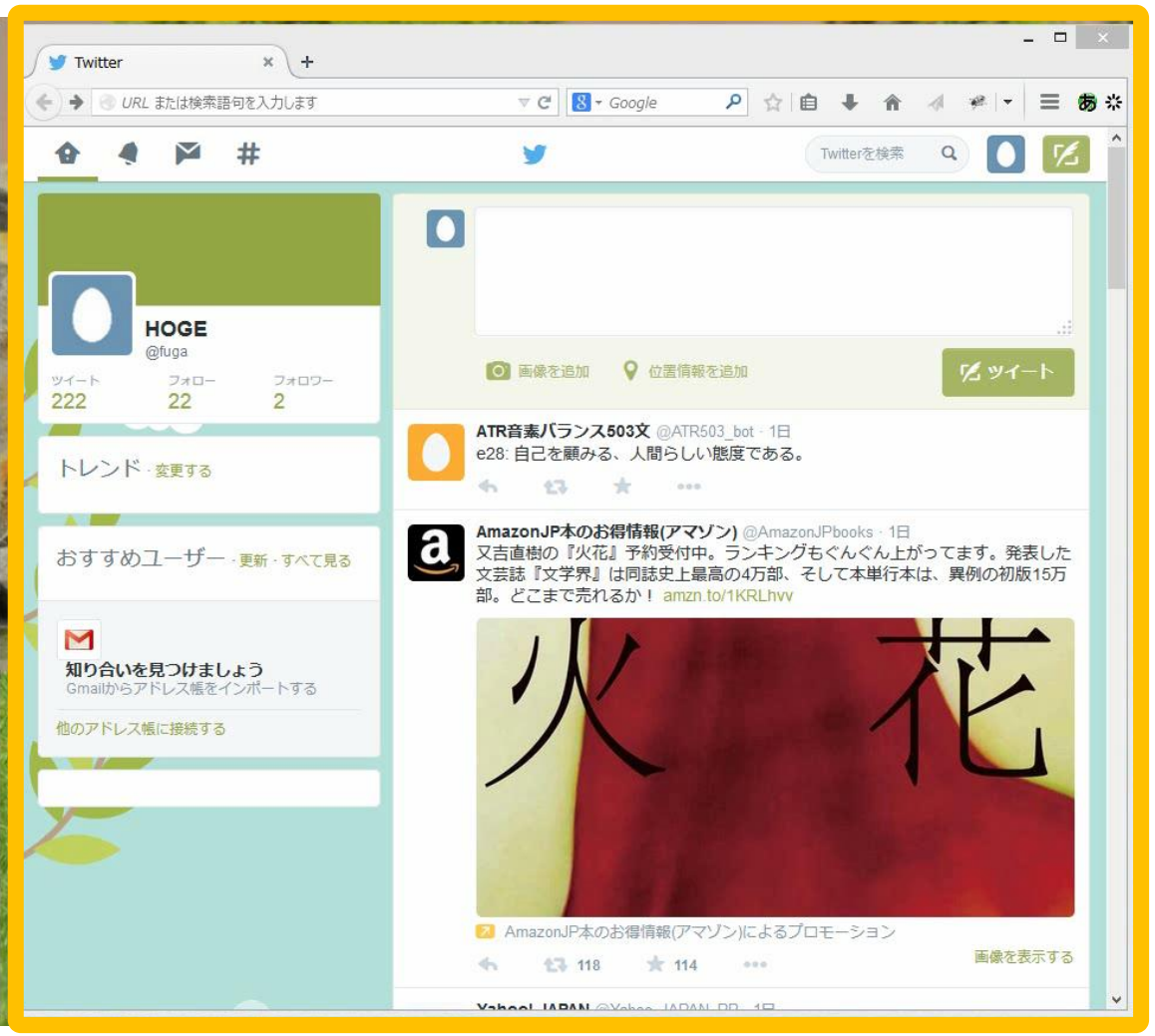


クライアントが逐次的にサーバにログを送信

**サーバ (コンソール)**




**クライアント (Twitterのページ)**

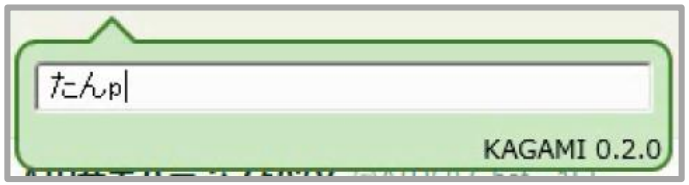


# 仮名漢字変換を使う工程



## 1．入力文字列の入力

文の読み情報




## 2．変換

**単語境界**と**表記**の選択


“誕プレに”を入力する過程


確定　ログの確定結果

誕プレ/たんぷれ | に/に

# 仮名漢字変換ログ



“それを誕プレにしよう”というツイートの
仮名漢字変換ログ

| 時間 | 入力文字列 | 確定結果 |
|---|---|---|
| 18:37:11.22 | それを | そ _ れ _ を |
| 18:37:12.60 | たんっぷれに | タンップレ/たんっぷれ \| に/に |
| 18:37:14.95 | たんぷれに | 誕プレ/たんぷれ \| に/に |
| 18:37:15.33 | あげよう | 上げ/あげ \| よう/よう |
| 18:37:19.83 | しよう | し _ よ _ う |


確定結果（変換ログの主要な情報）の特徴

- 完全な文でなく文の断片
- 誤った確定を含む
- 単語境界情報がない場合がある（変換しない場合）

**ノイズ**を含んだ、**文断片**のアノテーションデータ

# ログの利用方法



**AS-IS-log**：
確定結果をそのまま用いる

**CHUNK-log**：
入力の時間差が s 以下の場合、
確定結果をチャンキング　(s = 500[ms])

文断片の問題を回避

**MCONV-log**：
変換操作が n 回以下のログを除く( n = 2 )

ユーザーが編集した文は、
確定誤りを含む可能性が低く、
未知語を含む可能性が高い

| AS-IS-log |
| --- |
| そ _ れ _ に |
| タンッ/たんっ \| プレ/ぷれ \| に/に |
| 誕プレ/たんぷれ \| に/に |
| 上げ/あげ \| よう/よう |
| し _ よ _ う |

| CHUNK-log |
| --- |
| そ _ れ _ に \| タンッ/たんっ \| プレ/ぷれ \| に/に |
| 誕プレ/たんぷれ \| に/に \| 上げ/あげ \| よう/よう |
| し _ よ _ う |

| MCONV-log |
| --- |
| 誕プレ/たんぷれ \| に/に |
| 上げ/あげ \| よう/よう |

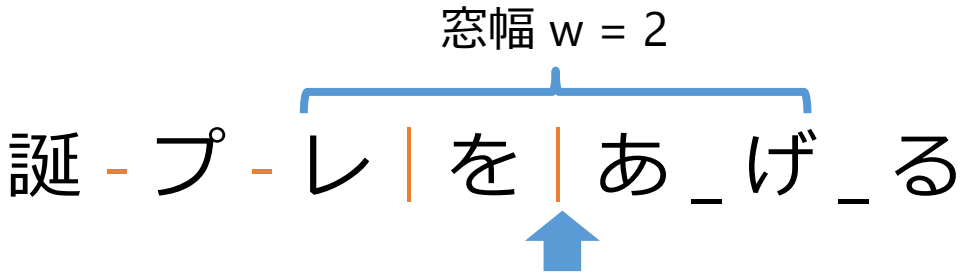
**アノテーションされていない部分のある学習データ**から推定器を学習

# 点予測による単語分割



[Neubig+, 2011]

他の推定情報を利用せず周辺文字情報のみを用いる推定手法




素性

| | | | |
|---|---|---|---|
| 文字3-gram | ： | -2/レ, を, あ | -1/を, あ, げ |
| 単語辞書 | ： | L(を) | R(あげ) |

➡ アノテーションされてない部分があっても学習可能

**窓幅w=3**

素性を**文字3-gram**, **文字種3-gram**, **単語辞書**で実験

# 評価実験



・ログの収集　ユーザー：５~１７人

収集期間：2014/04/24 – 2014/12/31(262日間)

人手アノテーション**テストコーパス**

| | ドメイン | 文数 | 単語数 |
|---|---|---|---|
| **TWI-test** | ツイート | 588 | 7,498 |

人手アノテーション**学習コーパス**

| | ドメイン | 文数 | 単語数 |
|---|---|---|---|
| **BCCWJ-train** | 一般分野 | 56,753 | 1,324,951 |

**ログ由来の学習コーパス**

| | エントリー数 | 単語境界 | 読み |
|---|---|---|---|
| **AS-IS-log** | 32,119 | 39,708 | 34,769 |
| **CHUNK-log** | 8,685 | 63,144 | 34,769 |
| **MCONV-log** | 4,610 | 10,852 | 8,959 |
| **CHUNK-MCONV-log** | 1,218 | 14,242 | 8,956 |

# 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 (F値)

| | 再現率 | 適合率 | F値 |
|---|---|---|---|
| **BCCWJ-train** | 90.31 | 94.05 | 92.14 |
| **+ AS-IS-log** | 90.33 | 93.77 | 92.02 |
| **+ CHUNK-log** | 91.04 | 94.29 | 92.64 |
| **+ MCONV-log** | 90.62 | 94.09 | 92.32 |
| **+ CHUNK-MCONV-log** | 91.40 | 94.45 | **92.90** |

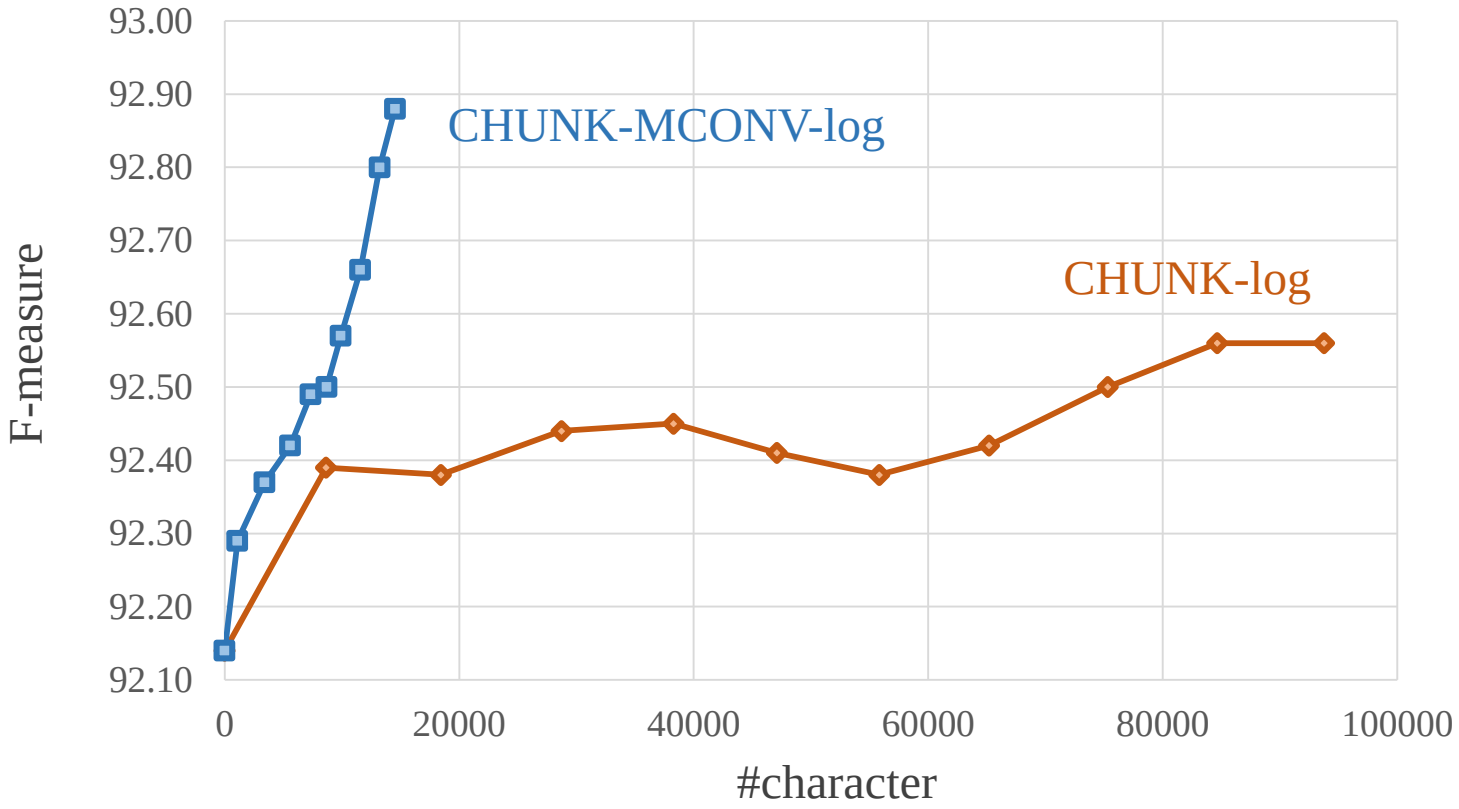
- **MCONV-log**によってノイズがフィルタリングされた
- **CHUNK-MCONV-log**で最大精度、統計的に優位な差 (p=0.01)
- 再現率 ＜ 適合率 ➡ 過分割

“艦これ” ➡ “艦” , “これ”

# ログの量と単語分割の精度



時系列順に並べたログを順次追加した、単語分割精度



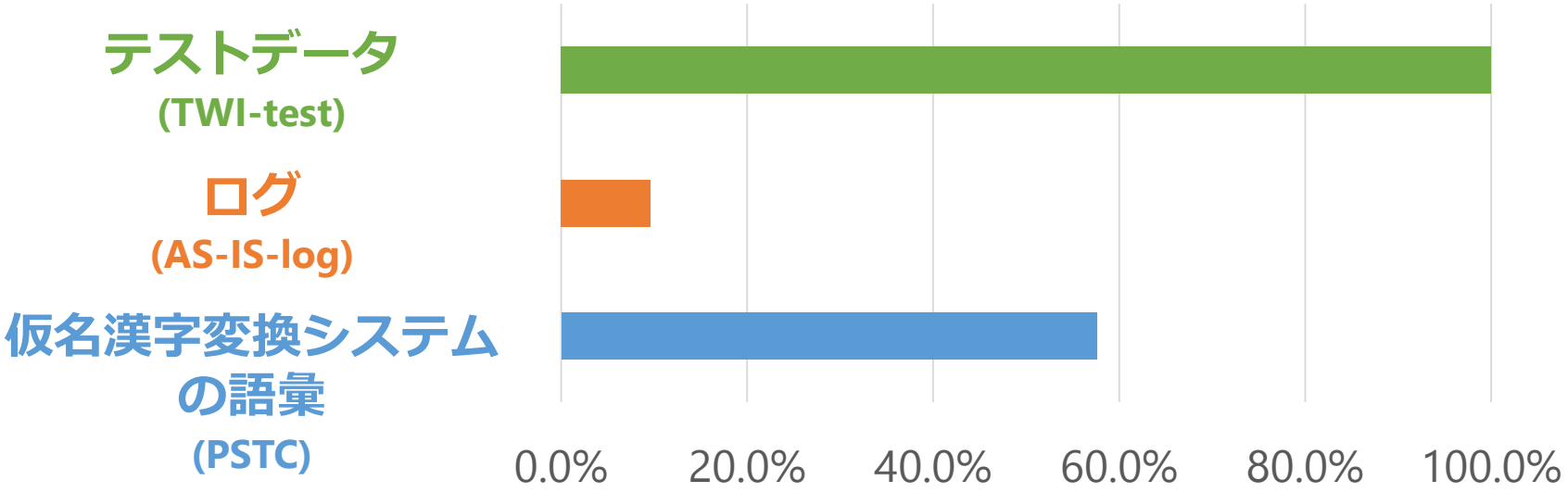

- **CHUNK-MCONV-log**は少ない量でより精度が向上
- 人手アノテーションの８００文字程度と同等の精度向上
- さらにログを集めることで精度向上が期待

# テストデータ中の未知語の分析



テストデータ 中の未知語の適合率

未知語：
BCCWJ, UniDicに含まれない単語
単語単位で扱う




仮名漢字変換システムの語彙中に、ログに未記録の未知語あり

➡ さらにログを集めることで、精度向上が期待

# まとめ

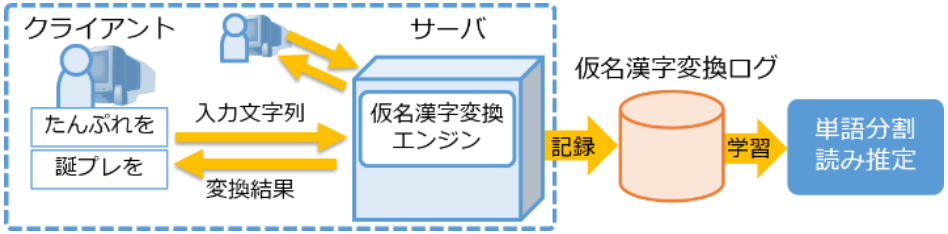


問題提起

## 単語分割の精度低下の問題

手法の提案

## 仮名漢字変換ログから有用な情報を獲得する方法を提案




結論

## ログによる精度向上を実現

- テストデータのドメインに合ったログで解析精度が向上
- ログを収集することでさらに精度向上が期待