Fairy Devices

日本語部分形態素アノテーションコーパスの構築

2017.5.15

@IPSJ-SIGNL231 (大阪大学中之島センター)

林部 祐太 (フェアリーデバイセズ)

形態素辞書更新による「正解析→誤解析」

ついつい物欲しそうになってしまう

✓ /物/欲し/そうに/ (JUMAN 7.01)

✗ /物欲/し/そうに/ (JUMAN++ 1.02)

鱗片の外側には細かい伏せた毛がある。

✓ /伏せ/た/毛/ (mecab-ipadic)

✗ /伏せ/た毛/ (mecab-ipadic-NEologd 2017.03.20)

本発表の概要

- 目的：辞書の更新等に伴う『正解析→誤解析』を検出したい
- 方法：**検出用**部分アノテーションコーパスを**(効率よく)**作成

都市には14路線ある|地下?鉄|網が整備されている。

✓ /地下/鉄/網/ /地下鉄/網/

✗ /地下/鉄網/

|藤堂?ユリカ|、藤原みやびの歌唱を担当。

✓ /藤堂/ユリカ/ /藤堂ユリカ/

✗ /藤/堂/ユリカ/

既存の形態素アノテーションコーパス

- フルアノテーションコーパス
  - 京都大学テキストコーパス
  - 京都大学ウェブ文書リードコーパス (KWDLC)
  - 現代日本語書き言葉均衡コーパス (BCCWJ)
  - EDR, GDA, KNBC, CSJ, …
- 部分アノテーションコーパス
  - ドメイン適応のための内製コーパス
  - 機能表現用例データベース 「MSUT1」, 「はごろも」
  - ウェブサイト「部分アノテーションの共有」

コーパス構築の2つのアプローチ

Wikipediaに対して，形態素解析器が誤りやすい箇所を中心に部分アノテーション

A) 自然アノテーションを用いた半自動アノテーション
B) 文字列検索ツールを用いた手動アノテーション

A)　自然アノテーションを用いた半自動アノテーション

自然アノテーションの利用 [Jiang+13]

球状星団 [編集]

数十万という恒星が密集している星団。年齢100億年以上の星々からなり、銀河形成の初期段階でできたものとされている。
我々の銀河系の場合、全天に分布しているが、銀河中心のあるいて座の方向に多く見られる。

両端を単語境界と仮定

銀河中心のある|**いて座**|の方向に多く見られる

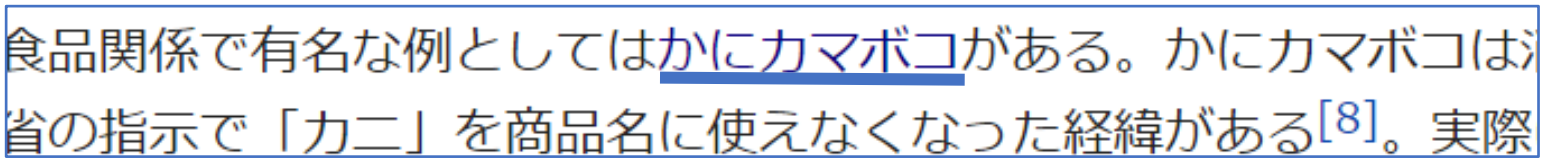
確認・修正の必要性

- 分割基準の違い
  - |スレンダー|な形状で軽快さを演出されている
  - |波|打ったような形をしていることもある
- 用法による分割の違い
  - |みずほダイレクト|に申し込むと…
  - より|ダイレクト|に楽曲に反映…
- リンクミス
  - 不対電子の|スピンか|ら生まれている
  - アンコール放送分は除いて|い|る

A)　自然アノテーションを用いた半自動アノテーション

手順

1. Wikipediaからリンクを含む文を抽出(約350万文)



2. JUMAN++やMeCab+UniDicで形態素解析

3. 自然アノテーションに違反する形態素分割を目で確認

…/例/と/して/はか/に/カマボコ/…

✗　✓

B) 手動アノテーション

Wikipediaのリンクの大半は名詞

→半自動アノテーションでは機能語等のアノテーションが不足

→手動でも部分アノテーション

文字列検索ツール

かい

"かい" (7765件, 1.024秒)

Search | toggleMA | マッチ表層: All | マッチ品詞: All | submit

ID	記事	テキスト	マッチ表層	マッチ品詞	Type
4439838	2040514	愛と平和と理解を信じる**かい**？	かい	助詞	OK
4622254	1893900	日本軍が**いくらかいる**だけである。	いくらか--いる	副詞--動詞	OK
5385074	2259490	経験を積んだため**かい**くらか落ち着きがあり、余裕のある態度…	か--いくらか	助詞--副詞	OK
5453045	2025217	聞いてわからんの**かい**。	かい	助詞	OK
5739444	167642	**遥かい**にしえ、天界には神と神に従う天使たちがいた。	遥か--いにしえ	副詞--名詞	OK
944791	2681440	**あかい**らかは、ワタナベエンターテインメント所属のお笑いコ…	あかい	名詞	NG_SEG
1217282	1576401	この芝居にはスコットランドやスコットランド人に対するから…	かい	助詞	NG_SEG
3395645	85994	休日に運転されている特急**かい**おう５号が行き違い待ちで当駅…	か--いおう	助詞--動詞	NG_SEG

マッチ表層: All | マッチ品詞: All

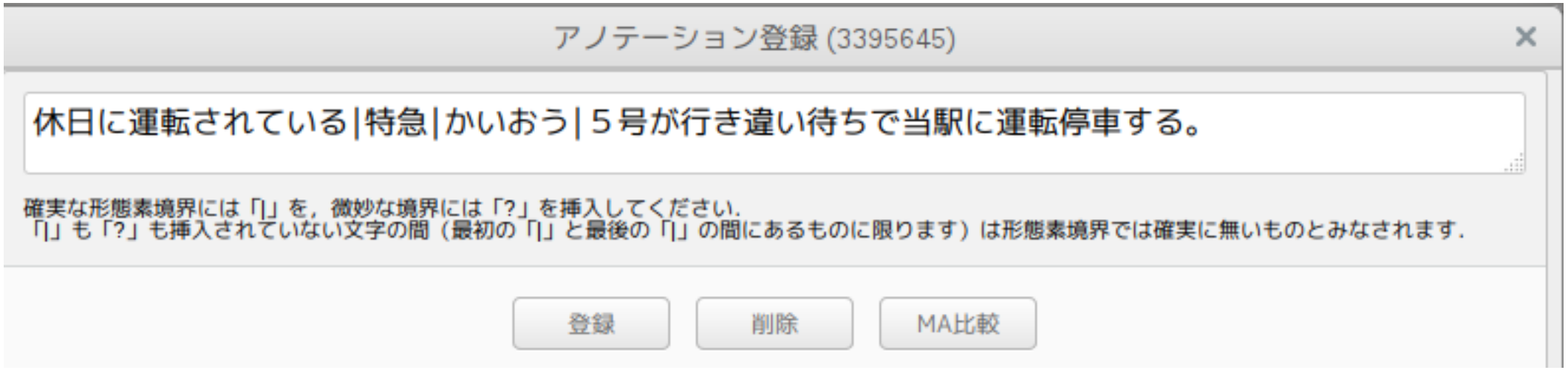
- か--いまひとつ 1
- か--いる 38
- か--いれる 1
- か--いろいろ 2
- か--いろいろな 1
- かい 87

マッチ品詞: All | submit

- ✓ All 7765
- UNK--接尾辞 3
- 副詞--副詞 2
- 副詞--動詞 14
- 副詞--名詞 3
- 副詞--形容詞 8
- 助詞 44

この芝居にはスコットランドやスコットランド人に対するから**かい**が見受けられる。

この/指示詞/連体詞形態指示詞 芝居/名詞/サ変名詞 に/助詞/格助詞 は/助詞/副助詞 スコットランド/名詞/地名 や/助詞/接続助詞 スコットランド/名詞/地名 人/名詞/普通名詞 に/助詞/格助詞 対する/動詞/ から/助詞/接続助詞 かい/助詞/終助詞 が/助詞/格助詞 見受け/動詞/ られる/接尾辞/動詞性接尾辞 。/特殊/句点


アノテーション登録 (3395645)
休日に運転されている|特急|かいおう|５号が行き違い待ちで当駅に運転停車する。
確実な形態素境界には「|」を，微妙な境界には「?」を挿入してください．
「|」も「?」も挿入されていない文字の間（最初の「|」と最後の「|」の間にあるものに限ります）は形態素境界では確実に無いものとみなされます．
登録
削除
MA比較

出来たコーパス: 約2,000文

A) 自然アノテーションを用いた半自動アノテーション
JUMAN++を使って約750文
UniDicを使って約600文

B) 文字列検索ツールを用いた手動アノテーション
約650文

コーパスの例

…改組する形で|片山?潜|ら|と労働組合期成会を結成した

弟に陸景、陸玄、陸機、陸雲、|陸?耽|らがいる

…都市には１４路線ある|地下?鉄|網が整備されている

１９５４年に|京都?大学|生理?学|教室入室

だが彼|こそ|は高度な科学力とすぐれた肉体で…

…学徒兵が外出|がてら|に主人公の家で記念写真を撮る

…|藤堂?ユリカ|、藤原みやびの歌唱を担当

|村川?梨衣|にとって初のソロＤＶＤである

今後の課題

1. コーパスのさらなる規模の拡大
   - 現時点で約2,000文
   - 機能表現辞書「つつじ」を参考に機能表現のアノテーションを追加したい

2. 方言や口語的な表現を含む文に対するアノテーション
   - 星空文庫の小説などにアノテーション

3. 品詞などの形態素情報の付与
   - 品詞や語彙素など

まとめ

- 辞書の更新に伴う『正解析→誤解析』を避けたい

- 検出用コーパスを2つのアプローチで作った
  半自動と手動

- コーパス・ツールを公開中
  https://github.com/FairyDevicesRD