
形の似ている漢字の誤用訂正

長岡技術科学大学学生 小川 耀一郎
長岡技術科学大学 山本 和英

研究背景

◆ 日本語学習者の作文誤りを自動訂正

- 日本語教師の負担を軽減
- 学習者自身の自己学習を支援

◆ 関連研究

- 助詞・格助詞の誤り訂正が多く研究されている

例) 兄は手紙 → 兄の手紙

研究目的

◆ 形の似ている漢字の誤用

- 因る(困る)、生治(生活)など
- 日本語学習者にとって間違えやすい
- 限定的な誤りであるが、文全体の校正には重要となる

私は**白**転車に乗る



私は自転車に乗る

このように訂正したい

形の似ている漢字

- ◆ 小学生対象の書き取り調査やウェブの漢字辞書等を参考
- ◆ 608組（1942字）の類似セットを作成

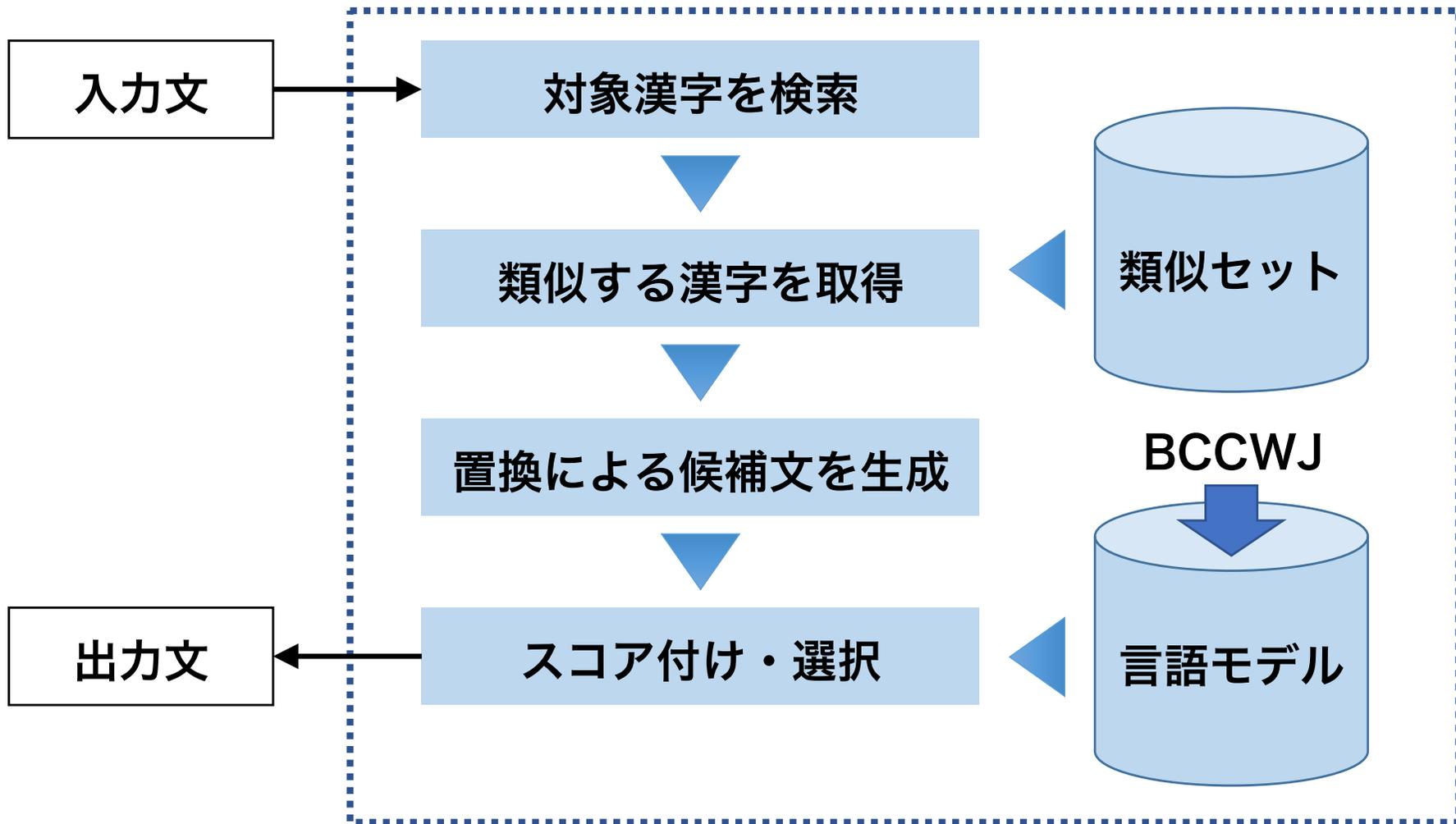
類似しているが異なる点

- 1画多いまたは少ない
- 辺の長さ
- 部首
- 旁

白	自	
午	牛	
複	復	腹
使	便	

など

システム構成



言語モデル：文の自然さを確率で表すモデル

言語モデル

◆ 言語モデルの構築

例) 明日友達と会う予定だ

単語分割

明日 / 友達 / と / 会う / 予定 / だ

単語3-gram

- コーパスから事前に単語N-gramをカウントしておく

◆ 言語モデル確率の計算

友達 / と / ?

会う : 30回 → 後に続く確率が高い

合う : 5回

- 文中のそれぞれの単語N-gramで確率を計算する

訂正の流れ

私は白転車に乗る

対象漢字を検索

私は白転車に乗る

類似する漢字を取得

[白, 自]

置換による候補文生成

- ・私は白転車に乗る
- ・私は自転車に乗る

スコア付け・選択

- ・私は白転車に乗る -20
- ・私は自転車に乗る -10

私は自転車に乗る

複数箇所の訂正

◆ 先頭の対象漢字から順に繰り返す

白転車を便った

候補文選択①

白転車を便った：-24
自転車を便った：-15

候補文選択②

自転車を便った：-15
自転車を使った：-10

自転車を使った

未知語フィルタ

◆ 生成した候補文を形態素解析

- 置換した漢字が”未知語”と解析される場合は誤用の可能性が高い
- “未知語”となる候補文を除外することで誤選択を防ぐ

候補文

1 /分間/打ち/続け/た

1 /分/間/打ち/続け/た



未知語

未知語フィルタ

1 分間打ち続けた

~~1 分間打ち続けた~~

実験と結果

◆ 学習者コーパス

- 実際の学習者作文の誤用文
- 47箇所の誤用を含む44文

◆ 擬似誤りコーパス

- 自動生成した誤用文
- 11,355箇所の誤用を含む10,677文

$$\text{再現率} = \frac{\text{正しい訂正箇所}}{\text{誤用箇所}}$$

$$\text{適合率} = \frac{\text{正しい訂正箇所}}{\text{システムの訂正箇所}}$$

$$F\text{値} = \frac{2 \times \text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}}$$

テストデータ	再現率	適合率	F値
学習者コーパス	94%	88%	91%
擬似誤りコーパス	93.9%	88.3%	91.0%

正しく訂正できなかった例

入力文：その島は珍しい

出力文：その鳥は珍しい

正解文：その島は珍しい

- 文の意味や文脈情報の活用が必要

入力文：ともだちのかぞくに合います

出力文：ともだちのかぞくに合います

正解文：ともだちのかぞくに会います

- ともだち/の/か/ぞ/くに/合い/ます
- 平仮名の形態素解析技術向上が必要

まとめ

- ◆ 形の似ている漢字の誤用を自動で訂正するシステムを開発した
- ◆ 多くの誤用の事例に対して高い性能での訂正が可能
- ◆ 手書き文字認識技術への応用も可能
- ◆ このシステムは下記URLで公開している
<http://www.jnlp.org/SNOW/S14>