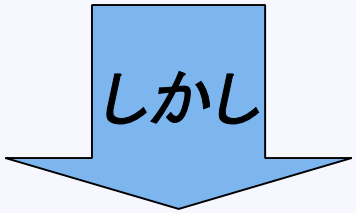
語彙獲得のための過分割未知語の検出

研究者
村脇有吾　黒橋禎夫
京都大学大学院情報学研究科

発表者　江口晃

はじめに

- 日本語の形態素解析は、形態素候補の列挙に辞書を用いる手法が主流

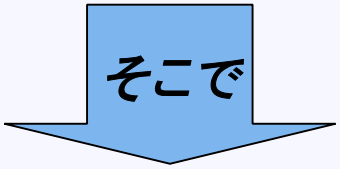


- 辞書にない形態素（未知語）の解析を誤りやすいという問題がある

手法の提案

- 未知語の解決策
  未知語の自動獲得による辞書の拡張

- 手法の提案
  未知語は、解析器が既知語の組に過分割してしまい、検出が難しい



過分割された未知語を検出する手法を提案する

未知語検出タスクの位置付け

- 検出
  各文の形態素解析結果から未知語の用例を検出
- 列挙
  各未知語用例に対して、辞書項目の候補を列挙
- 選択
  各未知語用例に対して、最適な辞書項目の候補を選択

検出タスクの性質

- 検出された未知語用例のみが獲得対象となるため、高い再現率が望まれる
- 選択が複数用例の比較により行われるため、誤検出が直ちに誤獲得には結びつかない。

ベースライン手法

- ベースライン手法
  形態素解析結果中の未定義語に着目する

  未定義語：未知語処理により列挙される形態素候補

- 過分割へのアイデア
  解析器が候補選択に使う連接の手がかりが品詞の連接であり、語彙の連接が考慮されないことがある。
  この語彙的な不整合から未知語用例を検出する。

未知語の過分割の例

- カタカナの過分割
  - カースト ⇒ カー ＋ スト
  - アブラハム ⇒ 油 ＋ ハム

- ひらがなの過分割
  - うざい ⇒ 卯 ＋ 剤
  - うざくて ⇒ 卯 ＋ 座 ＋ 区 ＋ 手
  - めんどかった ⇒ 面 ＋ 度 ＋ 買う
  - かぐや姫 ⇒ 家具 ＋ や ＋ 姫

提案手法のアイデア

- 表記ゆれの利用
  日本語では、1つの形態素が様々な表記を取りえる。この表記ゆれの利用を提案する。

- 表記ゆれの例
  「卯」 ⇒ 「卯」や「う」
  「剤」 ⇒ 「剤」や「ざい」
  「うざい」 ⇒ 「卯剤」、「卯ざい」、「う剤」
  しかし、こうした異表記はN-gramに出現しない。

代表表記

- 代表表記
  - 形態素の表記は、形態素解析器JUMANの辞書では、代表表記により吸収されている。
  - 表記ゆれの検出に代表表記を利用する

- 代表表記の例

  「卯」と「う」は代表表記「卯/う」に集約される

未知語箇所の検出

- 検出

$$L_{w_0, w_1} = C(w_0, r_1) / C(w_0', r_1)$$

w0に着目し、w1との連接を調べる。
閾値以上であれば、w0を検出箇所とする。
同様にw－１、w0の連接も検査する。

w－１、w0、w1：形態素列
w0’：w0の異表記
r1：w1の代表表記

N-gramの構築

- 検査対象表記の選定
  - 複数表記のない形態素や、カタカナ同士の表記ゆれを除外する。
  - 対象代表表記について、検査対象表記から異表記へのマッピングを作成する。

    例：対象代表表記「卯/う」、マッピング「う→卯」
- N-gramの構築
  - テキストを形態素解析する
  - 解析結果の形態素列について、w0の代表表記r1が検査対象なら、カウントCを更新する。

用例検出の実験設定

- タグ付け
  - 過分割の可能性のある短い形態素の連続のみを抽出し、それらに人手でタグ付けする。
  - 未知語用例の検出タスクにおいて検出が要求されるのはUだけで、EとOは参考として与えた。

タグの種類

U：未知語の語幹

例：「うざい」の「うざ」、「あきんど」

E：解析誤りのうち、現在の形態素解析器の文法・語彙で正しく解析しうるもの。

例：「何回かしか」の既知語の連続「か＋しか」は「か＋し＋か」と誤分解される。

O：その他の形態素解析の誤り

いずれにも該当しなければ何も付与しない

用例検出の実験設定

- テキスト
  ウェブコーパスから無造作に選ばれた8517文
  過分候補は535箇所抽出され、287個のタグが
  付与された

- N-gramの構築
  ウェブコーパス約1億ページ、頻度10で足切り

用例検出の実験結果

	U	E	O
ベースライン	57	1	69
提案手法	116	7	88
正解	159	18	110

用例検出の検出例

- 新たな検出例
  かもめ　⇒　鴨　＋　目
  すじこ巻き　⇒　する　＋事故　＋　巻く
- 検出されない未知語
  めも　⇒　目　＋　も
  しらす干　⇒　知る　＋　す（接尾辞）
- 前後の文脈によって変わる
  どれみちゃん　⇒　どれ　＋　味　＋　ちゃん
  どれみ！　⇒　どれ　＋　見る

未知語獲得の実験設定

- 対象テキスト

  TSUBAKIの検索結果上位1,000ページ

  クエリとして、「捕鯨問題」、「赤ちゃんポスト」、

  「JASRAC」を用いた。

- 正解判定

  獲得された未知語の精度を人手で評価する

  語幹と品詞の両者が正しい場合を正解とする

未知語獲得の実験結果

クエリ	ベースライン	提案手法
捕鯨問題	99.0% （204/206）	97.2% （210/216）
赤ちゃんポスト	98.9% （91/92）	95.9% （94/98）
JASRAC	96.7% （532/550）	96.1% （566/589）

未知語獲得の検出例

- 新たな検出例

  名詞：ぶろぐ、はてな、ドラえもん

  動詞：ぐぐる　イ形容詞：めんどくさい

- 誤り例

  名詞「とん」：「とんでもねえ」と組み合わせて獲得

  イ形容詞「めんどい」：候補選択を誤り名詞「めんどい」として誤獲得

おわりに

- 提案手法ではベースラインと比べて再現率を改善したが、検出が難しい用例も存在する。
- 今回はカタカナ過分割を対象外としたが、外国語の言語情報や、文脈的な手がかり等も組み合わせて、過分割問題を解決したい。