地方議会会議録における オノマトペの自動抽出手法の提案

◯木村泰知(小樽商大)　渋木英潔(横浜国大)、
内田ゆず(北海学園大)　乙武北斗(福岡大)、
高丸圭一(宇都宮共和大)　森辰則(横浜国立大)

オノマトペとは

- ✓擬音語および擬態語
  - 例. 擬音語「どんどん(叩く)」
  - 例. 擬態語「のびのび」「どんどん(進める)」
- ✓音, 雰囲気, 程度, 様子を効果的に伝える手段
  - 例. 雨の音「しとしと」「ぽつぽつ」「ざーざー」
- ✓日本語の話しことばで多用される

特に、3文字以下の誤抽出が多い

文字長	正抽出数	誤抽出数	正抽出率
2	15	604	2.4 %
3	580	1,549	27.2 %
4	999	356	73.7 %
6	61	0	100.0 %

2文字

「ぼっ」まつぼっくり…
「ふっ」いきいきふっつ高齢者プラン
「ごん」ごんは、いたずらを後悔し…

3文字

「えへん」せえへん人が出てくる…
「ぼやっ」何ぼやったんですか。
「ちゃっ」全部押さえられちゃったらば

提案手法

形態素解析器と構文解析器を用いて
2つの規則のみで抽出する

形態素解析によって1 つの形態素に分割されており，
以下の品詞に解析されているオノマトペ候補
(A-1) 4 文字以上の「副詞」「名詞」
(A-2) 3 文字以下の「副詞」

規則A

文節の先頭にあり，以下の3 条件のいずれかにあてはまるオノマトペ候補
(B-1) 4 文字以上である
(B-2) 3 文字で最後が「り」である
(B-3) 3 文字以下で直後に「と」「っと」「に」のいずれかが続く

規則B

実験方法

✓対象とする会議録

1. 2010年度、402 自治体、約1319万文(約3億語)

✓実験データ

オノマトペ	12,261発言
非オノマトペ	6,437発言

正解データの作成方法

1. 形態素解析によってオノマトペを抽出
2. 正解ラベルの作成 (177語を人手で確認)

✓比較実験

- ベースライン… JUMAN辞書にオノマトペを登録
- 提案手法　　… 規則Aと規則Bの適用

実験結果（提案手法）

提案手法の精度　88.4%

正解ラベル	総数	抽出	非抽出
オノマトペ	12,261	11,484	777
非オノマトペ	6,437	1,390	5,047
合計	18,698	12,874	5,824

ベースライン 67.5% から 20.9ポイント向上

実験結果

- ✓ ベースラインから 20.9 ポイント向上
  - 予備抽出で67.5%
  - 提案手法で88.4%

- ✓ 4文字以上の精度 … 変化なし
  - 例「ぴったり」「ごちゃごちゃ」
  - 予備抽出で91.12%
  - 提案手法で91.09%

- ✓ 3文字以下の精度 … 45.9ポイント向上
  - 例「にやり」「じん」
  - 予備抽出で39.5%
  - 提案手法で85.4%

考察　誤抽出・未抽出(1/7)

形態素解析，同音異義語，その他の観点から要因を明らかにする

未抽出　6.3%(777 例)

正解ラベル	総数	抽出	非抽出
オノマトペ	12,261	11,484	777
非オノマトペ	6,437	1,390	5,047
合計	18,698	12,874	5,824

誤抽出　21.6%(1,390 例)

考察　誤抽出・未抽出(2/7)

形態素解析　同音異義語，その他の観点から要因を明らかにする

誤抽出の多いオノマトペ

たった	ごくごく	おいおい	しとり	さらさら	くすり	さんさん
かったん	かんから	とことこ	なんなん	たっ	ぽっ	とくとく

方言文法「～しとります」「～言わへんかったんで～」「～せないかんから～」

未抽出の多いオノマトペ

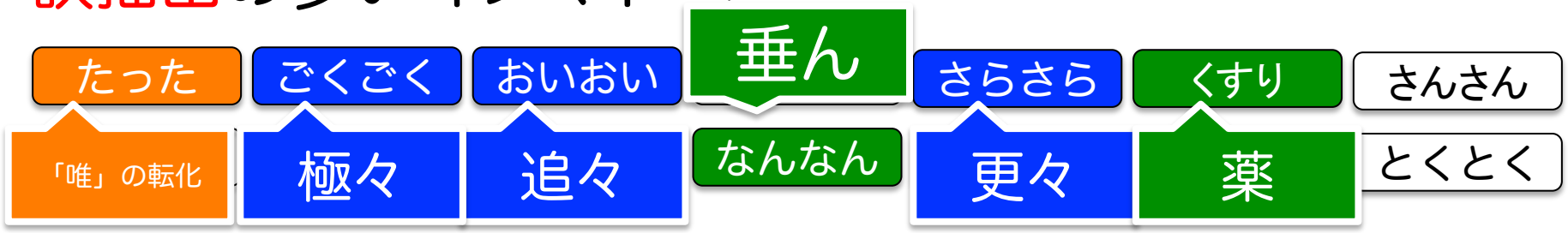
ばさっと	➡	ば[助詞]	さっと[副詞]	
ごちゃ	➡	ご[接頭詞]	ちゃ[名詞]	
ずらずらと	➡	ずら[動詞]	ず [助動詞]	ら[名詞]	と[助詞]

形態素解析の精度向上が必要

考察　誤抽出・未抽出(3/7)

形態素解析, 同音異義語　その他の観点から要因を明らかにする

誤抽出の多いオノマトペ



「唯」転化、表層的に一致する副詞 、ひらがな表記

形態素解析の精度向上が必要

考察　誤抽出・未抽出(4/7)

形態素解析，同音異義語　その他の観点から要因を明らかにする

誤抽出の多いオノマトペ

たった	ごくごく	おいおい	しとり	さらさら	くすり	さんさん
かったん	かんから	とことこ	なんなん	たっ	ぽっ	とくとく

制度・施設の名称として利用される

語の意味･構文解析の係り先情報等を利用する必要がある

考察　誤抽出・未抽出(5/7)

形態素解析，同音異義語　その他の観点から要因を明らかにする

✓「きらら」
- 語義「明るくまぶしく輝き続けているさま」
- 一般の辞書に掲載される 意味は「雲母」の別称
- 名詞として解析されるため未検出

語の意味・構文解析の係り先情報等を利用する必要がある

考察　誤抽出・未抽出(6/7)

形態素解析，同音異義語，その他の観点から要因を明らかにする

- ✓ 3 文字以下のオノマトペ
  - 促音や長音を伴い、連続して出現することがある
- ✓ 連続して出現するオノマトペ
  - 例.「ぽん」「ぽんぽん」「ぽんぽんぽん」
    - ■ 「ぽん」と「ぽんぽん」はオノマトペ辞典に掲載あり
    - ■ 本手法による抽出結果
      - – 結果「ぽんぽん」→「ぽん[副詞] ＋ぽん[副 詞]」
      - – 評価「ぽんぽん」の未抽出
      - – 課題　検出には成功しているため，精度評価の検討の余地あり

考察　誤抽出・未抽出(7/7)

形態素解析，同音異義語，その他の観点から要因を明らかにする

- ✓ 2文字のオノマトペ
  - 規則(B-3) により誤抽出
    - ■「ぽっ」「まちぽっと」(組織名称)
    - ■「たっ」「たっとい」(尊い)
      - 「たっ」は87,534 回出現するが、オノマトペの可能性が低い

文字長の短い抽出規則について，事例の分析が必要

まとめ

- ✓ 地方議会会議録からオノマトペを抽出する手法の検討
  - 形態素解析と構文解析の利用
  - ２つのルール
- ✓ 実験結果
  - 精度 88.1%
  - 対象データ　177 語（12,261 例）のオノマトペを含む文
- ✓ 有効性
  - 短いオノマトペにおいて誤抽出が大幅に削減された
  - 3文字以下のオノマトペの抽出精度 45.9 ポイント向上
- ✓ 今後の課題
  - 形態素解析において口語表現や方言文法への対応が必要
  - 同音異義語に対応するために，構文解析の係り先情報の利用する必要あり