

# 商品名とカテゴリ名情報の統合による データ拡張手法と商品名抽出器への応用

情報学群 情報科学類  
自然言語処理グループ 山本研究室  
202011693 佐多亮明

# 研究背景

- SNSやブログ上にはユーザーによる無数の商品レビュー文
  - ↔ ほとんどの情報は非構造化データ
- レビュー文を用いた分析および活用
  - レビュー文と商品名をリンクさせたデータベース
    - ↔ 全ての商品名を網羅した辞書は存在しない
- コンピュータに人間のような文脈判断
  - 機械学習による商品名抽出器が必要

# 関連研究: 商品名抽出

商品カテゴリ情報に着目した自動収集教師データによる商品名抽出 [2012, 渡邊]

- 教師データ作成はコスト大 & 大量のデータが必要 → 自動的なデータ拡張が必要  
ラベル付けの例: 「私は、iPhone 13を持っています。」

- 商品名とカテゴリ名の文脈は類似していることが多い  
→ カテゴリ名にもラベルを付与して教師データを拡張

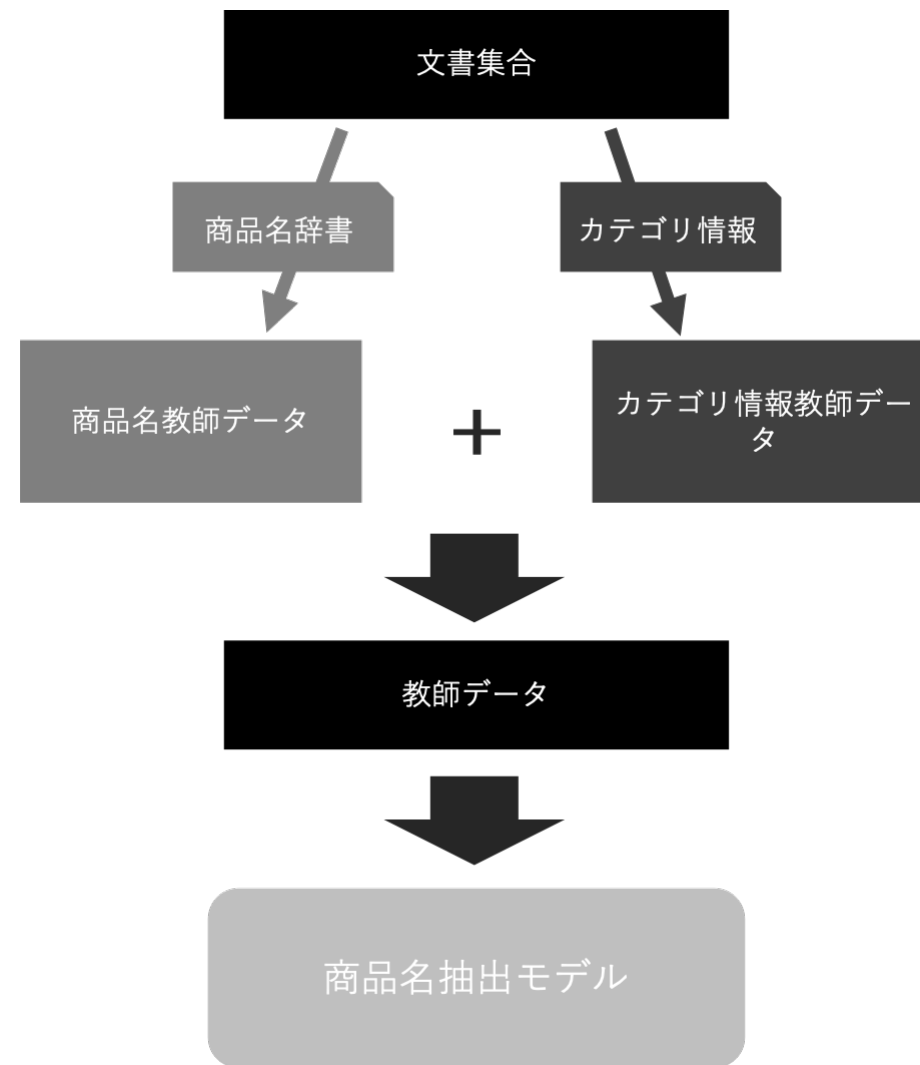
文章例:

- 商品名 … 先日、家電量販店でiPhone 13を購入した
- カテゴリ名 … 電車内ではスマホでネットサーフィンをしている

→ どちらもスマートフォンの話題でよく使われる単語で構成されている

# 関連研究: 商品名抽出

- 商品名教師データの作成
  - 文書集合から該当する商品名を含む文章を取得
  - 商品名にラベルを与える
- カテゴリ教師データの作成
  - 文書集合から該当するカテゴリ名を含む文章を取得
  - カテゴリ名にラベルを与える
- 商品名教師データとカテゴリ教師データの統合
  - 統合された教師データを抽出器に与えて学習



# 研究目的

- データ拡張手法の有効性の調査
  - 先行研究は曖昧な性能差のみ
  - どのような条件で有効かについての実験は無し
- 先行研究の性能改善
  - データセットの作り方を工夫し性能向上
  - 本研究ではBERTを使用
- 未知の商品名に対する評価
  - 先行研究は学習データに存在しない商品名での評価は無し
  - 実用上は未知の商品に対する推論の方が重要

# 提案手法 (1/2)

- データ拡張

- 先行研究をもとに実施

- 商品名教師データとカテゴリ名教師データを作成・統合

- カテゴリ名の置き換え

- BERTは商品名の文字列パターンも学習

- カテゴリ名を商品名として学習してしまう可能性

- カテゴリ名をランダムな商品名に変換

- 文章例:

- 変換前 … 先日、家電量販店でスマートフォンを購入した。
      - 変換後 … 先日、家電量販店でGoogle Pixel 7を購入した。

# 提案手法 (2/2)

- データ選択

- ECサイトで商品名を検索

- 上位5件までの検索結果にその商品名が含まれるかを調査

- 含まれている場合にのみ教師データに用いる

- そのカテゴリの商品名として使われている語句のみを教師データに用いる

- データ修正

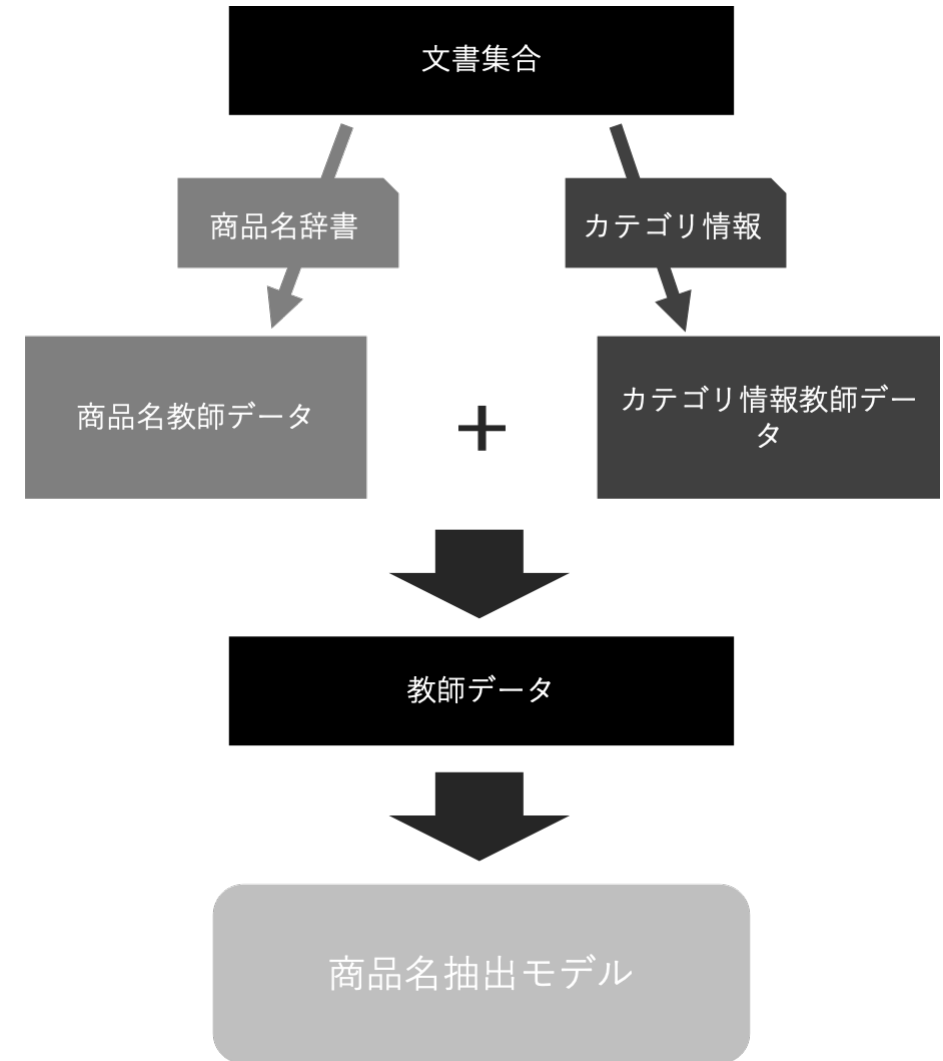
- データ選択でも排除できなかった商品名を手で削除

- 文書集合に含まれているブログなどの引用文を削除

→ 以上の手法をもとに実験を行う

# 実験設定

- 学習に用いる文書集合はツイートデータ
  - 教師データで用いる投稿時期は2011年, テストデータは2012年
- 抽出対象の商品カテゴリは「ゲームタイトル」
  - 1980年から2012年までに発売された商品名を使用
- 商品名教師データの数を可変して実験
  - データ数が与える影響を調査
- 評価指標はF1-score
  - 固有表現抽出では最も一般的
  - 0~1の値で表現され、大きいほど性能が高い





# 実験結果

- データ拡張

データ拡張なし → 0%  
データ拡張あり → 数十%



商品名教師データ数	カテゴリ名教師データ数	全データ数	f1-score (%)	
3019 ( 100% )	0 ( 0% )	3019	56.6%	
3019 ( <b>37.2%</b> )	5095 ( 62.8% )	8114	<b>64.0%</b>	→ ◎
8242 ( 100% )	0 ( 0% )	8242	77.4%	
8242 ( <b>62.3%</b> )	4978 ( 37.3% )	13220	<b>77.7%</b>	→ △
24187 ( 100% )	0 ( 0% )	24187	71.6%	
24187 ( <b>80.8%</b> )	5742 ( 19.2% )	29929	<b>71.4%</b>	→ ×

※ ()の数値は全データ数に対する割合

→ 商品名教師データの比率が大きくなると、データ拡張の効果が小さくなる

# 実験結果

- データ選択・データ修正

商品名教師データ (件)	前処理	f1-score (%)
24187	なし	27.1%
	データ選択	58.1%
	データ選択 + データ修正	71.6%

※ データ拡張はしていない

→ データ選択・データ修正ともに効果あり

# 実験結果

- 抽出例

データ拡張なし	データ拡張あり
でも明日買うかな...雨ふってるから引き取り面倒だし、ソールトリガー クリアしてないし	でも明日買うかな...雨ふってるから引き取り面倒だし、[(GAME) ソー ルトリガー] クリアしてないし
意外と [(GAME) エクストルーパー] ズが面白そうだぞ・・・	意外と [(GAME) エクストルーパーズ] が面白そうだぞ・・・
[(GAME) 那] 由 [(GAME) 多の軌跡] が7月26日って早くないか...?零Evo もあるし死ねる....	[(GAME) 那由多の軌跡] が7月26日って早くないか...?零Evoもあるし死 ねる....
[(GAME) 大神 絶景] 版が美しすぎるのでふて寝	[(GAME) 大神 絶景版] が美しすぎるのでふて寝

※ GAMEタグの中身が抽出部分

# 結論

- 抽出性能は商品名教師データの比率に依存
  - 全データ数に対して商品名教師データ数が小さい場合に性能向上が見られる
    - 商品名が取得しにくい分野ではデータ拡張が非常に有効である
- 教師データに用いる商品名の選定が非常に重要
  - データ選択・修正によって目的カテゴリのより適切な文脈が得られた
- 未知の商品名に対してもBERT商品名抽出器は有効

# 今後の課題

- 最新の文章に対する評価
  - 実験に用いたツイートは2011~2012年
- 複数カテゴリを同時に抽出するモデルの作成
  - 今回は「ゲームタイトル」のみに特化していた
- 性能に最も貢献する商品名教師データの比率の試算
  - 商品名教師データの数をより細かく変えて実験することで算出可能