文献紹介

自然言語処理研究室
後藤　大明

出典

日本語のガ格に対する副助詞「は」の使用の推定
横野光, 稲邑哲也, 研究報告自然言語処理（NL）, No.6, pp.1-7, 2012

概要:本論文では日本語のガ格要素に対して,副助詞「は」を用いて表現するかどうかを推定するモデル を提案する.「は」は主題を表すために用いられることが多く,提案モデルはそれに着目し前文脈において 対象要素がどのような使われ方をしているかを考慮する.新聞記事を対象とした実験により,文中要素の みに着目した既存手法よりも提案モデルの方が良い性能を示すことが明らかになった.

キーワード:「は」と「が」, 自然言語生成,教育応用

背景

- 書き手が自分の意図を正しく伝えるためわかりやすい文章を作成することが必要
- 内容が正確であっても表現方法によっては読み手に誤った解釈を促す
- 自然な文章の書き方は統語規則と異なり、母語話者の直感に頼るため、規則として表現することは困難

背景

- 文章の不自然さに関わる文法項目
  - 副助詞「は」に注目
- 「は」の使用例
  - 格助詞「が」「を」の代わり
  - 格助詞「に」「から」に接続
- 「が」と「は」の使い分けの基準は曖昧
  - 「は」の習得は日本語学習者にとって大きな壁

目的

- 日本語学習者が自信の作成した文章に対して自動で添削を受けることは困難
- 副助詞「は」の使用モデルを提案
- 述語のガ格に対象を限定した使用モデルを提案

関連研究

- 助詞の誤り訂正に関する研究
  - 誤り傾向を考慮した格助詞の訂正法[笠原, 2012]
  - 中国語母語話者の日本語作文の助詞誤りに対する助詞誤り訂正手法[今村, 2012]
- 統計手法に基づいた誤り訂正では非母語話者によって作成されたテキストと修正されたテキストが必要
  - (e.g Konan-JIEM, Leaner Corpus)

関連研究

- 助詞の誤り訂正に関する研究
  - 同文中の要素に着目したモテル[三浦, 2012]

- 本研究では同文中の要素に加え前文脈を考慮したモデルを提案

「は」の用法

- 「は」の用法
  - 主題、対比
- 主語が主題になりやすい文
  - 属性を述べる文
- 談話において「は」を使う要素
  - 主題ではないが既出である場合
  - 前文脈に出現している要素と関係がある場合

「は」の用法

- 従属節
  - 従属度によって主題をとるか決まる
  - 従属度が高い「～ながら」「～つつ」
    - 主節に対する独立度が低いので節内で主語を取れない
  - 従属度中「～ので」「～など」
    - 主語を持てるが主題としては扱えない
  - 従属度低「～から」等位節
    - ガ格を主題として表すことができる

提案手法

- 「は」と「が」の使い分けを二値分類問題としてモデル化
  - 推定の対象となる要素が文章中でどのように用いられたか
  - ほかの要素がどのように用いられているか

- 前提条件
  - 推定対象の要素が文中に出現する
  - テキスト中の共参照関係は既知
  - 述語の格要素は既知

推定対象要素についての素性

- 推定対象要素が与えられたテキストにおいて初出であるか否か
- 代名詞か否か
- 推定対象要素から最も近い対象述語までの間に他の要素をガ格とする述語があるか否か
- 推定対象要素から最も遠い対象述語までの間に推定対象要素をガ格とする述語があるか否か

対象述語についての素性

- 述語の原型
- 述語の品詞
- 述語に後接する機能表現
- 述語のカテゴリ

- カテゴリは竹内らの動詞項構造シソーラスの大分類1を利用
- 複数の述語が存在する場合、対象述語がどの節にあるかで区別

推定対象要素を含む文中のほかの要素についての素性

·推定対象要素以外の述語の格となる要素について

·副助詞「は」を用いて表現されているか否か

·省略されているか否か

·その要素が属している節と述語の何格かによって区別

推定対象要素と共参照関係にある要素についての素性

- 推定対象要素を含む文からm文前までを前文脈とする

- 副助詞「は」を用いて表現されているか否か

- その要素と推定対象要素との表層の違い(完全に一致, 主辞のみ一致, 代名詞, 省略, その他)

- その要素が属している節と述語の何格かによって区別

- その他の素性

- 推定対象要素を含む分の文頭に接続詞がある場合の表層形

評価実験

- NAISTテキストコーパスに表層格の情報がアノテーションされたもの
- 分類モデルにはSVMを利用

	majority	比較手法	提案手法
正解率	0.518	0.734	**0.760**

- majorityは事例に対して訓練データにおいて多かった方を割り当てたもの

実験結果

·成功例

- 遺伝性の重い神経疾患である脊髄性筋委縮症の原因と思われる遺伝子を日本とカナダの国際共同研究グループが発見した。
  ＳＭＡの診断や治療に結び付く成果で、十三日付の米国の論文誌「セル」に発表した。
  SMA は 常染色体劣性の遺伝性疾患で、脊髄中の神経細胞が障害を受けて筋肉が委縮し、手足のまひや呼吸障害が起きる。

·失敗例

- 高成長を続けるサービス業の中堅企業は、北陸、関東、九州などに多く、**近畿** が 低調なことが、ニッセイ基礎研究所のまとめたリポート「一九九五年度主要産業の展望」でわかった。
- 同教育庁は「拒否の理由があいまい。受け入れの余裕があれば拒否できない」として、調査のうえ強制命令を出す見通しだ。
  南アでは五年前に集団地域法が廃止されるまでは、法的にも白人と黒人の居住地域が分離され、学校も人種別だった。
  同法廃止後も、概して人種ごとに収入が違うことから人種別の居住地域は続いており、白人地域の学校の方が施設が整備され、**教育水準** は 高い。

追加実験

- 「は」と「が」の使い分けを実際に人が判断したデータを用いて評価
- 作業者３人が　格助詞、は、省略で分類
  - ３人が同じ判定 strict
  - ２人以上が同じ判定 majority
- 誰も選ばなかったものをシステムが選ばなかったとき negative

追加実験結果

- 人手でアノテーションした結果を正解 human
- 元テキストを正解 original

設定	human	original
strict	0.723 (94/130)	0.808 (105/130)
majority	0.658 (123/187)	0.840 (157/187)
negative	0.866 (142/164)	0.884 (145/164)

- 提案モデルが新聞記事に現れる「は」と「が」の使い分けを学習しているため普通の文章に対して結果が悪くなった

おわりに

- 文中のガ格要素に対して「が」を用いるか「は」を用いるか推定するモデルを提案した
- 提案モデルは前文脈において要素がどのような使われ方をしているか考慮し、文中要素の情報のみを利用したモデルよりも良い性能であることがわかった
- 主題の「は」のみに限定しているため要素間の類似度や構造間の類似度を考慮し、対比を推定する必要がある