文献紹介ゼミ

林 秀治

紹介する文献

- Detection of Grammatical Errors Involving Prepositions
- Martin Chodorow, Joel R. Tetreault, Na-Rae Han
- ACL-SIGSEM Workshop on Prepositions 2007. p25-30

概要

- 前置詞はEnglish as a Second Language(ESL) Learnerの誤りの原因の大部分を占める
- この種の誤り検出のアプリケーションを開発することは、学習者に貴重な学習資源を提供する

- ルールベースのフィルタと組み合わせた最大エントロピー分類器を使い、再現率30%、精度80%を達成した

背景

- 前置詞は、非ネイティブがマスターする文法の中で最も難しいとされるもののひとつ
- 日本でも村田らが日本人が書く英語の誤りの18%が前置詞であると報告

- 間違った前置詞や、前置詞を伴う誤りの検出を目指す

The Selection Model

前置詞の誤りには以下のようなケースがある

- 間違った前置詞の使用
  - They arrived to the town
- 使用できない文脈での使用
  - They came to inside
- 必要な文脈での欠落
  - He is fond this book

The Selection Model

前置詞の誤りには以下のようなケースがある

- 間違った前置詞の使用
  - They arrived to the town
- 使用できない文脈での使用
  - They came to inside
- 必要な文脈での欠落
  - He is fond this book

The Selection Model

- テキストの特徴に基づいて、34種の前置詞を推定する最大エントロピー(ME)モデルを使用
- MEモデルはMetaMetrics corpusとSan Jose Mercury Newsの新聞記事から抽出された700万の前置詞の周辺文脈で訓練
- モデルは25種の文の形態で訓練された

Some features used in ME Model

Feature	Description	No. of values with freq $\geq$ 10
BGL	Bigram to left; includes preceding word and POS	23,620
BGR	Bigram to right; includes following word and POS	20,495
FH	Headword of the following phrase	19,718
FP	Following phrase	40,778
PHR_pre	Preceding phrase type	2
PN	Preceding noun	18,329
PNMod	Adjective modifying preceding noun	3,267
PNP	Preceding noun phrase	29,334
PPrep	Preceding preposition	60
PV	Preceding verb	5,221
PVP	Preceding verb phrase	23,436
PVtag	POS tag of the preceding verb	24
PVword	Lemma of the preceding verb	5,221
PW	Lemma of the preceding word	2,437
TGL	Trigram to left; includes two preceding words and POS	44,446
TGR	Trigram to right; includes two following words and POS	54,906

Evaluation on Grammatical Text

- モデルを訓練に使用されなかった18,157の前置詞の周辺文脈を使ってテストした

- 各文脈に対し、モデルが34の前置詞の確率を推測し、最も確率が高いとされた前置詞と実際に使われている前置詞を比較する
- 比較の結果、どちらも同じ前置詞だったものは69%で、カッパ値は0.64

結果の改善

- 最も確率の高い前置詞と2番目のものとの差が数%しかないものがいくつかあった
- 例えば、以下の文では、_の位置にorとinの両方が使用できる
  - They thanked him for his consideration _ this matter
- 1番目と2番目の確率の差が小さい場合は検出を行わない

結果の改善

- 差が60%以上のときのみ検出を行うようにすると50%は検出が行われなかった

- しかし、処理された50%では、結果が一致したものが90%でカッパ値は0.88であった

Evaluation on ESL Essays

- 中国、日本、ロシア人が書いた文章からランダムに2000の前置詞の周辺文脈を抽出

- この文章の誤りを先ほどのモデルを使って検出したところ様々な理由でうまくいかなかった

スペルミス

- 学習者の文章は多くのスペルミスを含んでいるが、訓練に使った文章ではスペルミスがなかった

- 前置詞に隣接する位置か、隣接する句の頭にスペルミスの単語があった場合、検出を行わない

カンマの有無

- “I disagree because from my point of view there is no evidence”のような文があったとき、”because”のあとにカンマがあるかないかで結果が変わる
- カンマがない場合、”because”のあとに”of”を選択した
- カンマの誤りを検出した場合検出を行わない

反意語と受益者格

- fromとtoなど逆の意味をもつ前置詞が選ばれた場合、誤りとしなかった

- for + person/organizationのような形からなる受益者格は学習が困難であったので対象外とした

タグ付けの結果

- 先ほどの2000の前置詞を二人の評価者がチェックし、その結果と分類器の結果を比較

	Rater 1 vs. Rater 2	**Classifier vs. Rater 1**	**Classifier vs. Rater 2**
Agreement	0.926	0.942	0.934
Kappa	0.599	0.365	0.291
Precision	N/A	0.778	0.677
Recall	N/A	0.259	0.205

ルールベースフィルタの追加

- 複数の数量詞の前置詞の削除
  - Some of people
- 繰り返される前置詞の削除
  - people can find friends with with the same interests

- これらのフィルタを追加することで精度が0.778から0.796に、再現率が0.259から0.304に向上