文献紹介ゼミ

林 秀治

紹介する文献

- The Ups and Downs of Preposition Error Detection in ESL Writing
- Joel R. Tetreault, Martin Chodorow

概要

- ノンネイティブの英語学習者が書いた文の前置詞の誤りを検出する
- 大規模な学生のエッセイのセットで精度84%を達成したが再現率は19%

はじめに

- 近年、ノンネイティブが英語で文章を書く場面が増えている
- 前置詞の使用を伴うエラーがノンネイティブの文章で最も一般的なものの一つ

システムのモデル

- 34の一般的な英語の前置詞の正しい使用状況のモデルを構築するために最大エントロピー(ME)分類器を使用している
- 高校生用の教科書などを含むMetaMetrics Lexile corpusから抽出された700万の前置詞のコンテキストを使って学習
- コンテキストは語と品詞のタグからなる25の特徴で表されている

フィルタ

- 前処理フィルタ
  - スペルミスが含まれている前置詞コンテキストをスキップ
  - 分類器はミスのないテキストで学習しているし、最初にスペルチェッカーによる修正があると思われるため
- 後処理フィルタ
  - 特定のコンテキストで起きる偽陽性をなくす
  - ユーザの意図に依存する反意語をエラーとするケースなど

フィルタ

- 余分な使用のフィルタ
  - “some of people”のような複数の数量詞を持つ文や”can find fiends with with”のような前置詞が繰り返されている場合に対応する
  - このような余分な使用のエラーは全体の18%を占める
- 閾値処理
  - 偽陽性のあるケースをスキップする
  - ユーザの前置詞よりシステムの前置詞がわずかに上回る場合は誤りとしない

    My sister usually gets home around 3:00

    (around = 0.49, by = 0.51)

評価

- 二人の評価者により、8269の前置詞からなる注釈つきのコーパスを収集
- この研究では2種類のエラーに焦点を当てている
  - 不正な前置詞の使用
  - 余分な使用
- ベースラインとして25の特徴と組み合わせ機能を備えたモデルを追加した

評価

- ベースラインは精度79.8、再現率11.7%であった
- 次に語、タグ、語+タグ、3つすべての4つのモデルのコンビネーションモデルでもテストを行った
  - タグ、語＋タグ、3つすべて　の３つでは性能の向上が見られなかった
  - 語だけ再現率が1%向上した
- 700万のトレーニングコーパス中のスペルミスを持つ20万のコンテキストを削除し再度学習を行った

結果

- スペルミス削除後にタグと組み合わせたものが最も良い結果が得られた

モデル	精度	再現率
ベースライン	79.8	11.7
+語	79.8	12.8
+タグ(削除後)	82.1	14.1

- さらに、余分な使用のフィルタを+タグに使用したところ、精度が84%、再現率が19%に向上した

エラーの分析

- アノテーションしたセットから学習者がよく誤って使用した前置詞を調査した
  - in (21.4%)
  - to (20.8%)
  - of (16.6%)
- Top10の前置詞だけでエラーの93.8%を占めた

エラー詳細

間違えた前置詞	正しい前置詞	頻度(%)
to	null	9.5
of	null	7.3
in	at	7.1
to	for	4.6
in	null	3.2
of	for	3.1
in	on	3.1
of	in	2.9
at	in	2.7
for	to	2.5