文献紹介ゼミ

林 秀治

紹介する文献

- Semi-automated typical error annotation for learner English essays: integrating frameworks
- Andrey Kutuzov and Elizaveta Kuzmenko
- Proceedings of the 4th workshop on NLP for Computer Assisted Language Learning at NODALIDA 2015
- pp.35-41

概要

- 3つのツールを使って誤りを検出・評価
  - Brat
  - Freeling
  - Aspell
- 約75%の誤りを自動で検出

背景

- 学習者コーパスは典型的な学習者の誤りテキストとして言語学習に貢献している
- しかし、学習者コーパスのアノテーションは、書き手の意図した内容を把握しづらいので困難である

- アノテーションを助けるための半自動の前処理

使用するツール

- Brat：Web上のテキストアノテーションフレームワーク。Web上で数人が同時にアノテーション可能。
- Freeling：いくつかの言語で使用できるオープンソースの言語解析器
- Aspell：オープンソースのスペル訂正システム

使用するコーパス

- REALECコーパス：ロシア人の学習者が書いた英語のコーパス
- Bratを使って、英語教師などの専門家によって人手でアノテーション（正解データ）

誤り検出手法

- 文章をFreelingでトークン化、文分割、形態素解析
- 全てのトークンと見出し語をAspellでチェック
- トークンも見出し語も既知の英単語でないならば、『Possible spelling error or typo』を付与
- Freelingによる形態素解析で、各形態素に可能性のあるすべての品詞タグを付与

品詞タグの付与

- 誤りによって実際の品詞の確率が低くなる

  例：play(s)のタグと確率

  He plays with his phone.

  VBZ(動詞、三人称単数)：0.663934

  NNS(名詞、複数)：0.336066

  He play with his phone.

  VB(動詞、原形)：0.565684

  NN(名詞、単数)：0.270777

  VBP(動詞、非三人称単数)：0.163539

- 下の例でも前後関係のためVBPが選ばれる

Mixing tools and the corpus

- 選ばれた品詞が低確率の場合誤りである可能性が高い
  - 『Possible grammar or morphology error』を付与

実験

- REALECコーパス(800の文書、213,694語)
- Aspellによって3,018のスペルミス、Freelingによって10,490の形態的なミスを検出
- 人手の結果と比率はほぼ一致

実験結果

- 完全一致で見た場合はほぼ一致しない
- 人手で誤りとした文で誤りを検出したかで見た場合F値0.57

	Precision	Recall	F-measure
Strict matches			
Overall	0.04	0.07	0.05
Aspell only	0.007	0.04	0.01
Freeling only	0.046	0.06	0.05
Sentence-wise matches			
Baseline	0.0973	0.169	0.123
Overall	0.4637	**0.7479**	0.57
Freeling only	**0.7643**	0.5383	**0.63**

Baseline：50%で誤りを付与

実験結果

- Aspellは辞書にない単語を誤りとして扱うため精度が下がるが再現率が上がる

	Precision	Recall	F-measure
Strict matches			
Overall	0.04	0.07	0.05
Aspell only	0.007	0.04	0.01
Freeling only	0.046	0.06	0.05
Sentence-wise matches			
Baseline	0.0973	0.169	0.123
Overall	0.4637	**0.7479**	0.57
Freeling only	**0.7643**	0.5383	**0.63**

Baseline：50%で誤りを付与