# 文献紹介ゼミ

林 秀治

# 紹介する文献

- Sentene-level Rewriting Detection
- Fan Zhang, Diane Litman
- Proceedings 9th Workshop on Innovative Use of NLP for Building Educational Applications (ACL Workshop) (2014)
- pp.149-154

# 概要

- 文章の校正のプロセスを理解するためには、どのような変更があったかを知る必要がある
- 従来の研究では主に単語やフレーズレベルであったが、文レベルでの修正の検出に焦点を当てる
- 自動的で文レベルでの校正による変更を検出する

# 関連研究

- 過去の研究(Faigley and witte,1981; Connor and Asenavage, 1994)では、変更を6種類に分類していた

  Addition：語または句の追加

  Deletion：語または句の削除

  Substitutions：語を同義語に置換

  Permutation：語または句の並び替え

  Distribution：1つの語を二つに分割

  Consolidation：2つの語を1つに連結

# Sentence-level revisions

- 1文1対応にするために4分類に変更

  Add

  Delete

  Modify

  Keep：修正なし

# Data and annotation

- コーパスは大学生の書いたshort paperの初稿と最終稿で構成された“Social Implications of Computing Technology”を使用
- 2つのトピック
  - Big Dataがオバマ大統領の選挙で果たした役割(C1) 11ペア
  - 知的財産について(C2)10ペア

# Revision change detection

- 対応付け、校正シーケンス生成、校正シーケンスのマージ3stepで行う
- 対応は必ずしも1対1ではなく1対多の場合もある

# 校正の自動検出

- 人手と同じ対応付け、編集シーケンス生成、編集シーケンスのマージの3stepで行う

# 対応付け

- Nelkenのアプローチをもとに3stepで行う
- データの準備
  - 最終稿に対応する新しい文がないとき、初稿と最終稿の文でペアを作成し、誤ったペアを作るため、別の文とのペアも作成する
- トレーニング
  - 作ったペアの類似度を計算し、ペアか否かを予測するロジスティック回帰分類器をトレーニング
- 対応付け
  - デカルト積を使って、最終稿と初稿の文のペアを構成し、分類器を使ってペアか否かを判断する

# 対応付け

- 文間の類似度の計算方法にLevenshtein distance(LD)(Levenshtein,1966)、Word Overlap(WO)、TF*IDFを使用し、それぞれの比較も行った

# 編集シーケンス生成

- ルールベースの方法で以下の手順で行う
- Step1:原文書の文iが校正後の文書の文jと対応するときStep2へ、そうでなければStep3へ
- Step2：2文が完全に一致するなら”Keep”、そうでなければ”Modify”とし、iとjを1増やしStep1へ
- Step3:iとjを比較し、jがiより大きければ”Delete”としiを+1、そうでなければ”Add”としjを+1しStep1へ

# シーケンスのマージ

- DistributionとConsolidataionはModify, Add, Deleteで表すことができる
  - Distribution：Modify-Delete-Delete-...
  - Consolidataion：Modify-Add-Add-...
- Modifyから始まるこれらは文をマージすることができればシーケンスグループにマージできる
  - 上記2パターンのいずれかがあった場合文をマージし、類似度を計算し分類器を使って分類できればマージする

# 評価実験

- 最終稿の文の数をN1、正しい対応ができた文の数をN2として、精度N2/N1を計算
- ベースラインとしてHashemiの手法を使用
- ４つのセットで実験：他の原稿で学習しC1とC2をテスト(Group1,2)、片方で学習し、片方をテスト(Group3,4)

| Group | LD | WO | TF*IDF | Baseline |
|---|---|---|---|---|
| 1 | 0.9811 | 0.9863 | **0.9931** | 0.9427 |
| 2 | 0.9649 | 0.9593 | **0.9667** | 0.9011 |
| 3 | **0.9727** | 0.9700 | **0.9727** | 0.9045 |
| 4 | 0.9860 | **0.9886** | 0.9798 | 0.9589 |