文献紹介ゼミ

林 秀治

紹介する文献

- Automated Error Detection in Digitized Cultural Heritage Documents
- Kata Ga´bor, Benoit Sagot
- EACL 2014. p56-61

概要

- 文書を電子化するパフォーマンスの最適化を目的とする
- 統計分類アルゴリズムと言語知識を使ったOCRの誤り検出と訂正を行う手法を提案する
- この論文では言語モジュールの結合と、誤り検出・訂正への影響を扱う

背景

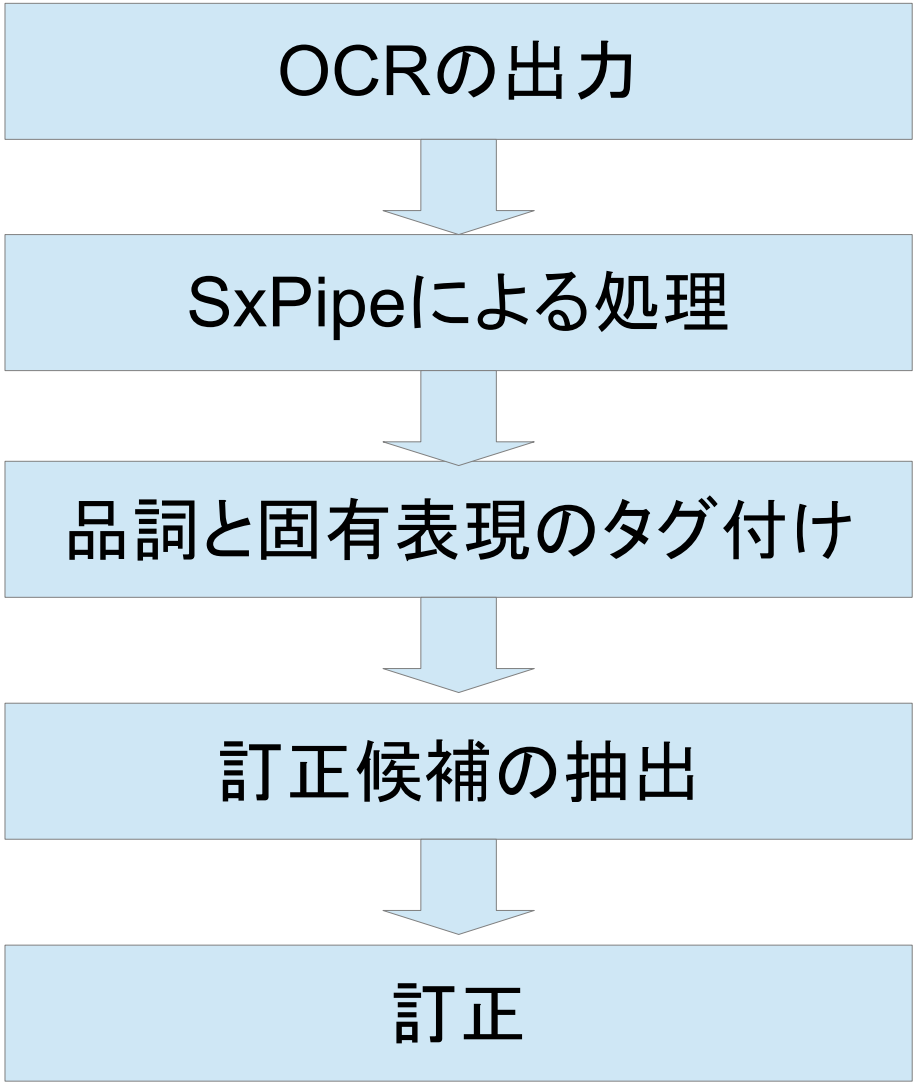
- 現在のState-of-the-artな光学文字認識(OCR)ソフトは精度90~99%を達成している
- この精度は、索引付けのためなら十分かもしれないが、実際に読むための文章ではより高い基準を満たす必要がある

- この論文では、OCRの精度向上のため、デジタル化された文章の自動後処理を扱う

背景2

- 今まで、OCRの後処理として言語処理が利用されたことはほとんどなかった
- この論文では特定の固有表現タグに対処するためにPOS Taggerを訓練した
- 筆者は固有表現が訂正タスクのコスト削減と特定の表現のエラー検出の高精度化に適していると主張している

システムの構造



システムの構造

- 文章のトークン化、文分割、固有表現の認識のためにSxPipeが使われている
- 固有表現タグがついたテキストをMElt-hを使って品詞タグをつける
- 訂正候補の抽出は、スペルチェックに適したnoisy channel modelをベースとしている

noisy channel model

- Noisy channel modelは文字列sを入力として与え、P(w|s)を最大にする語wを見つける
- ベイズの定理を使って以下のような式で表される

$$\mathrm{Argmax}(w)P(s|w)*P(w)$$

P(w)は整理されたコーパスから得られた言語モデルから与えられる

P(s|w)はエラーモデルから与えられる

エラーモデルはOCRを人手で修正したコーパスから与えられる

固有表現タグ

- 誤り検出としてまず、固有表現のようなイレギュラーを見つけるためにOCRの出力をSxPipeを使って分析する
- この研究では、固有表現を形態素または構文解析できない文字の連鎖と定義した
- 固有表現タグは品詞タグ付けの精度を向上させ、固有表現中に発生するOCRエラーの検出と訂正を可能にする

固有表現タグ付けの評価

- 評価は人手で行った
- 今回扱う固有表現は時間表現、住所、通貨単位、次元、化学式、法律ID
- 化学式は文献から抽出した文章を使って評価した
- 法律IDはPublications事務所の法律コーパスを使って評価した
- 残りはBNFコーパスを使って評価した

タグ付けの評価

- 実験は固有表現で人手でアノテーションされた332文の訓練データの異なる部分を使って実行された
- まず、ランダムに100文を抽出し、テストコーパスとした
- 残りの文をトレーニングデータとして50,100,150,232文で区切り使用した
- 評価のために100の固有表現でアノテーションされた文だけでなくFrench TreeBank(FTB)のテストセクションでも評価した

タグ付けの結果

文の数	FTB	NE
0	97.83	-
50	97.82	95.61
100	97.8	95.71
150	97.78	95.76
200	97.78	95.84
232	97.75	96.2