用語バリエーションの認識と正規化

東京大学　岡崎　直観

なぜこの研究をやることになったのか？

- もともとは略語定義の抽出をやって
  2006; Okazaki, 2008)
  - 略語抽出の精度に関しては満足のいく
- MEDLINE全体から抽出されたPMAの
  - phorbol 12-myristate 13-acetate; phorbol myristate acetate; 4 beta-phorbol 12 beta-myristate 13 alpha-acetate; postmenstrual age; muscular atrophy; phorbolmyristate acetate; phorbol myristic acid; phorbol 12-myristate acetate; para-methoxyamphetamine; phorbol ester 12-myristate 13-acetate; premarket approval application; phorbol 13-myristate 12-acetate; phorbol myristyl acetate; premotor area; phenylmercuric acetate; 12-phorbol 13-myristate acetate; paramethoxyamphetamine; phorbol myristoyl acetate; stimulation with phorbol ester; 4beta-phorbol 12beta-myristate 13alpha-acetate; phorbol myristate 13-acetate; premenstrual asthma; post-menstrual age; S-phenylmercapturic acid; phorbol 12-myristate 12-acetate; phorbal myristate acetate; phorbol 12-myrisate 13-acetate

異表記をまとめられないのか？

用語バリエーション問題

- すぐに思いつく方法
  - Porterのステミング，編集距離，文字n-gram類似度
  - うまくいかない例がたくさん見つかる
    - Finland – inland
    - preservations – reservations
  - モデル化が単純すぎて，致命的な間違いをしてしまう
- 本研究でやりたいこと
  - どのような文字同士の差異なら用語バリエーションとして許容されるか？　(認識)
  - 用語バリエーションから標準的な形に変換するにはどうすればよいか？　（正規化）
  - フレーズ単位のどのような差異で用語バリエーションが発生するか？

得られたもの (1/2)

◆部分文字列の書き換え知識

collogen|collagen
colocate|collocate
colo-ileal|coloileal
colonisation|colonization
colonise|colonize
colour|color

→ 学習

綴り異表記辞書，活用辞書

順位	置換元	置換先	重み	適用例
1	uss	us	9.81	focussing, russa
2	aev	ev	9.56	mediaeval, neuronaevus,
3	aen	en	9.53	pseudothrombocytopaenia

得られたもの (2/2)

- フレーズ単位での用語バリエーション

```
* A B C | B A C                    (211)
  JJ JJ NN | JJ JJ NN              (161)
  NN NN NN | NN NN NN              (79)
  JJ NN NN | NN JJ NN              (52)
* A B C | A of B C                 (176)
  NN NN NN | NN IN NN NN           (29)
  JJ NN NN | NN IN NN NN           (20)
  VB JJ NN | NN IN JJ NN           (16)
* A B C D | A B and C D            (161)
  NNP NNP NNP NNP | NNP NNP CC NNP NNP
  NN NN NN NN | NN NN CC NN NN
  JJ NN NN NN | JJ NN CC NN NN
```