用語バリエーションの認識と正規化

東京大学　岡崎　直観

- 同一概念を指す用語をまとめて扱いたい
  - *estrogen receptor, estrogen receptors, oestrogen receptor, oestrogen receptors, estrogen-receptor receptors of estrogen, …*
- 英語の用語のバリエーション (Jacquemin, 1999)
  - 語の変化 (paradigmatic variations)
    - 綴り変化や語尾変化
    - 略語: *estrogen receptor (ER)* ← これまでかなり取り組んできた (Okazaki, 2008)
    - 類義語: *carcinoma - cancer* ← 文脈や多義性の問題が顕著になる
  - 文法的なバリエーション (syntagmatic variations)
    - 語の挿入や並び替え: *receptor of estrogen – estrogen receptor*

単語の異表記

- 関連する研究
  - ステミング，lemmatization，スペル訂正，OCR誤り訂正，類似文字列マッチング
- 従来研究の問題点
  - ステミング: 語尾変化のみ
  - 編集距離，n-gram類似度: かなり粗い近似を行っている
    - *Finland – inland*
    - *preservations – reservations*
- 本研究の最終目標
  - 綴り変化や語尾変化の辞書から，単語異表記の識別・正規化モデルを構築する

- UMLS Specialist Lexicon
  - LRSPL（綴り変化）

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

  - LRAGR（語尾変化）

```
abundances|noun|count(thr_plur)|abundance
abundancies|noun|count(thr_plur)|abundancy
abundantly|adv|positive;periph|abundantly
abuse|noun|count(thr_sing)|abuse
abuse|noun|uncount(thr_sing)|abuse
abuses|noun|count(thr_plur)|abuse
abused|verb|past|abuse
```

  - LRNOM（名詞形導出）
  - LRWD（すべての語）

単語の異表記の識別

- 単語 s は t の異表記であるか？
  - 単語 s を t に変形する置換ルールのスコア（重み）の和で判別
    - s = ‘^anaemia$’, t = ‘^anemia$’ の場合:

[(’a’, ’’), (’na’, ’n’), (’ae’, ’e’), (’ana’, ’an’), (’nae’, ’ne’), (’aem’, ’em’)]

-1.13 +0.81 -2.21 -0.98 +0.86 +3.90

これらの和は 1.25 (> 0) なので，代表形であると判定

    - s = ‘^pain$’, t = ‘^pin$’ の場合:

[(’a’, ’’), (’pa’, ’p’), (’ai’, ’i’), (’^pa’, ’^pa’), (’pai’, ’pi’), (’ain’, ’in’)]

-1.13 -0.62 +0.32 0 0 0

これらの和は -1.43 (< 0) なので，代表形ではないと判定

学習で求める

  - ロジスティック回帰で識別モデルを構築・学習

$$P(1 \mid s,t) = \frac{1}{1+\exp\left(-\Lambda^T F(s,t)\right)}$$

単語 s が t の異表記である確率

重みベクトル

単語 s を t に変形する置換ルールに1を与える素性関数ベクトル

ロジスティック回帰

$$E_\Lambda = -\sum_{i=1}^{N} \log P(y^{(i)} \mid s^{(i)}, t^{(i)}) + \frac{|\Lambda|}{\sigma}$$

OW-LQN法 (Andrew, 07) で最小化

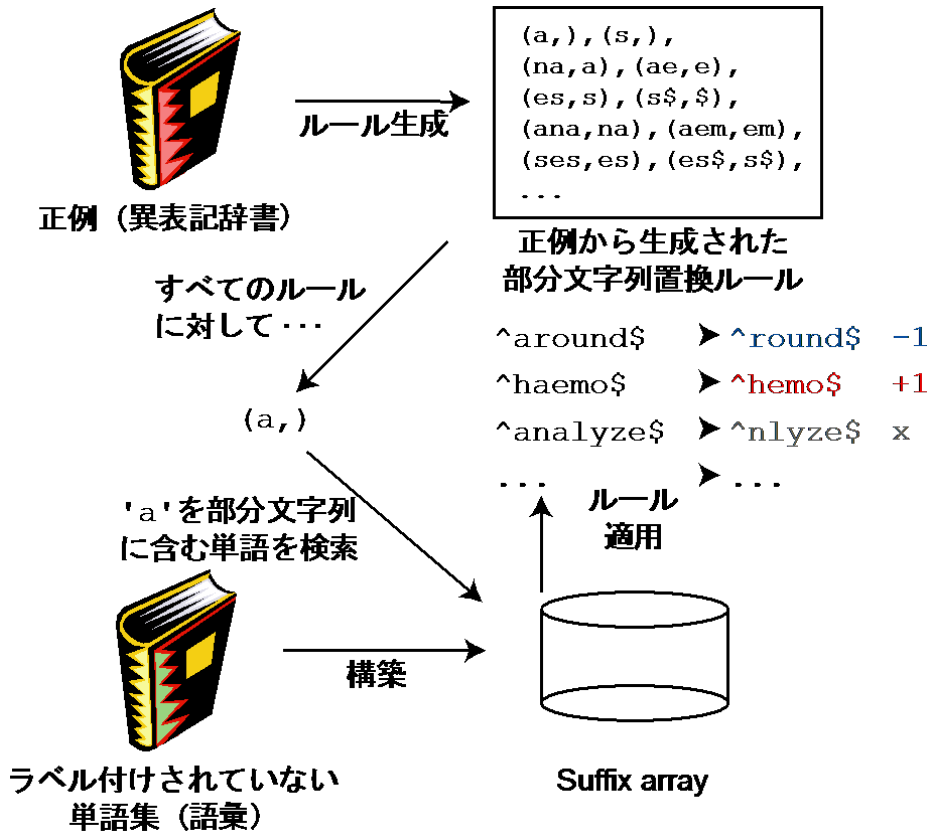
対数尤度

L_1正則化項（素性削減を行う）

MAP推定による学習

異表記辞書から学習データへの変換

- 異表記辞書の問題点
  - 負例が収録されていない
- 負例の生成方針
  - すべての語のペアを考慮
  - 正例になり得ない負例は刈る



- 綴り変化
  - 15,830 + 33,296

```
+1 ^analyze$ ^analyse$
-1 ^zingers$ ^singers$
-1 ^blaze$ ^blase$
+1 ^haemo$ ^hemo$
-1 ^around$ ^round$
-1 ^main$ ^min$
```

- 語尾変化
  - 113,215 + 124,747

```
+1 ^studied$ ^study$
+1 ^studies$ ^study$
+1 ^studying$ ^study$
-1 ^sturdy$ ^study$
+1 ^data$ ^datum$
-1 ^data$ ^date$
```

- 名詞導出
  - 12,988 + 85,928

異表記の識別から正規化（変形）へ

- 単語 s の代表形 t^* は何か？
  - $t^* = \underset{t \in \mathrm{gen}(s)}{\mathrm{argmax}}\, \underbrace{P(t \mid s)}_{\text{Reranker}}, \quad \underbrace{\mathrm{gen}(s)}_{\text{Candidate generator}} = \{t \mid P(1 \mid s,t) > P(0 \mid s,t)\}$

- Reranker
  - 正規化候補 t を順序付け
  - 最大エントロピー法でモデル化
  - 素性には，s を t に変形する置換ルール群，s と t のbigram, trigramを用いた

- Candidate generator
  - 可能なすべての t に判別式を適用するのは非効率
  - 正の重みが割り当てられた置換ルール群 D を s に適用する

```
s: ^anaemia$
    an    →   en    ~~^enaemia$~~
    an    →   in    ~~^inaemia$~~
    ana   →   an    ^anemia$
     na   →    n    ^anemia$
     nae  →    ne   ^anemia$
      an  →    e    ^anemia$
```

```
T = set()
U = set()
for i in range(len(s)):
    F = D.common_prefix_search(s[i:])
    for f in F:
        if f not in U:
            t = f.apply(s)
            if classify(s, t) == 1:
                T.add(t)
            U.add(f)
return T
```

異表記の識別と正規化の評価

綴り異表記，導出された名詞，活用変化形の識別性能

システム	Orthography			Derivation			Inflection		
	P	R	F1	P	R	F1	P	R	F1
Levenshtein distance	.329	.999	.488	.131	1.00	.232	.479	.988	.646
Normalized LD	.441	.847	.580	.133	.990	.235	.598	.770	.673
Dice coefficient (bigram)	.401	.918	.558	.137	.984	.240	.476	1.00	.645
LCSR	.322	1.00	.487	.156	.841	.263	.476	1.00	.645
PREFIX	.418	.927	.576	.140	.943	.244	.476	1.00	.645
Porter stemmer	.084	.079	.079	.197	.846	.320	.926	.839	.881
Morpha (Minnen, 2001)	.009	.007	.008	.012	.022	.016	.979	.836	.902
CST (Dalianis, 2006)	.119	.008	.016	.383	.682	.491	.821	.176	.290
Proposed method	.941	.898	.919	**.896**	.880	.888	**.985**	.986	**.984**
SVM training of features	.943	.890	.916	.894	**.886**	**.890**	.980	**.987**	.983
+ LD, NLD, DICE, LCSR, PREFIX	**.946**	**.906**	**.926**	.894	**.886**	**.890**	.980	**.987**	.983

綴り異表記，導出された名詞，活用変化形の正規化性能

システム	Orthography			Derivation			Inflection		
	P	R	F1	P	R	F1	P	R	F1
Morpha (Minnen, 2001)	.078	.012	.021	.233	.016	.029	.435	.682	.531
CST (Dalianis, 2006)	.135	.160	.146	.378	.732	.499	.367	.762	.495
Proposed method	**.859**	**.823**	**.841**	**.979**	**.981**	**.980**	**.973**	**.979**	**.976**

フレーズ単位でのバリエーション

- 語を超えたバリエーション
  - 語の追加／削除されやすいか
  - 語順／文法的構造の変更
- 用いるリソース
  - 自動獲得した略語の定義
  - 同じ略語に略される定義は，同一の概念を指す傾向にある
  - 例) PMAの定義

phorbol 12-myristate 13-acetate
phorbol myristate acetate
postmenstrual age
phorbolmyristate acetate
phorbol myristic acid
phorbol 12-myristate acetate
para-methoxyamphetamine
premarket approval application
phorbol 13-myristate 12-acetate

- 品詞パターン

```
* 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
```

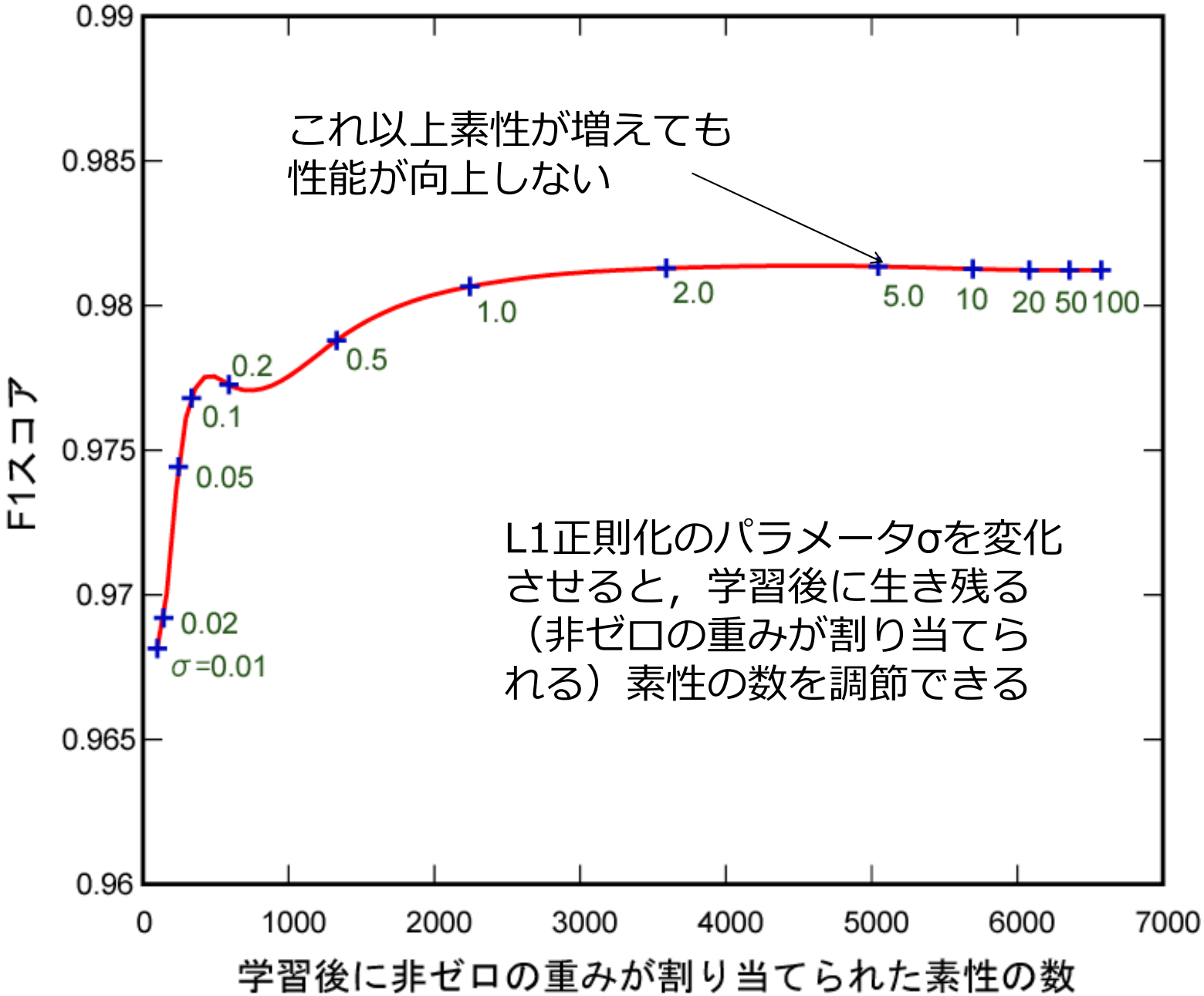
- 挿入／削除されやすい語

of	(13,863)
,	(12,786)
the	(10,864)
and	(6,693)
type	(4,404)
cell	(3,444)
protein	(3,325)
in	(3,321)
activate	(2,357)
for	(2,355)

まとめと今後の課題

- まとめ
  - 語の異表記の生成過程を部分文字列置換ルールで表現した
  - L1正則化ロジスティック回帰を用いて，異表記識別モデルを設計し，異表記辞書から識別モデルを学習する方法を提案した
  - 異表記識別モデルで得られた素性を利用し，与えられた文字列から正規化候補を列挙するアルゴリズムを提案した
  - 語の異表記の認識・正規化のタスクにおいて，ステミングや編集距離など，従来手法を大きく上回る性能を示した
  - フレーズ単位での用語バリエーションの出現パターンを分析した
- 今後の課題
  - 綴りの正規化やlemmatizer，stemmerなどのツールを整備する
  - フレーズ単位での用語バリエーションの認識精度を評価する

素性数とF1スコア（語尾変化）



これ以上素性が増えても
性能が向上しない

L1正則化のパラメータσを変化
させると，学習後に生き残る
（非ゼロの重みが割り当てら
れる）素性の数を調節できる

高性能な置換ルール例（綴り変化）

順位	置換元	置換先	重み	適用例
1	uss	us	9.81	focussing, russa, intussusception
2	aev	ev	9.56	mediaeval, neuronaevus, laevorotation
3	aen	en	9.53	hyaena, pseudothrombocytopaenia
4	iae$	ae$	9.44	gadoviae, kristiniae, coyleiae, cerevisiae
5	nii	ni	9.16	darwinii, avidinii, gibsonii, potronii
6	nne	ne	8.84	connexus, hartmannella, fennelliae
7	our	or	8.54	colour, neighbour, humour, rigour
8	aea	ea	8.31	paean, stomodaeal, gastraea, manichaean
9	aeu	eu	8.22	stomatodaeum, athenaeum, coprodaeum
10	ooll	ool	7.79	woollen, woolly, woolliness, coolly

- 学習結果を直接的に解釈できる
  - 最小置換ルールではなく，拡張置換ルールが上位を占めている