Conditional Random Fields を用いた略語抽出

岡崎　直観
東京大学大学院情報理工学系研究科

略語の問題

- 生物・医学分野において，略語は遺伝子・タンパク質などの重要なエンティティを表現
  - Swiss-Prot中のたんぱく質の34.6%は略語を持つ
  - 正式名称よりも略語の方がよく用いられる
    - *c-jun N-terminal kinase*: 3,047ヒット
    - *JNK*: 9,648ヒット
- 用語の多様性
  - *estrogen receptor: ER, ESR, OER, RE, …*
- 用語の曖昧性
  - *ER: estrogen receptor, endoplasmic reticulum, emergency room, …*

最終的な目標

RNA silencing processes are guided by small RNAs that are derived from double-stranded RNA. To probe for function of RNA silencing during infection of human cells by a DNA virus, we recorded the small RNA profile of cells infected by Epstein-Barr virus (EBV). We show that EBV expresses several microRNA (miRNA) genes. Given that miRNAs function in RNA silencing pathways either by targeting messenger RNAs for degradation or by repressing translation, we identified viral regulators of host and/or viral gene expression.

- グローバルな略語 (RNA, DNA) の曖昧性解消
  - 他文書内にて略語の定義をあらかじめ認識しておく
  - 略語の定義へのdisambiguationを行う（別タスク）
- ローカルな略語 (EBV, miRNA) の曖昧性解消
  - 文書内で略語の定義を認識しなければならない

略語抽出の従来研究

- ヒューリスティックやルール
  - 文字マッチングに基づく方法 (Yu et al., 02; Wren and Garner, 02; Schwartz and Hearst, 03)
  - 最長共通文字列一致 (LCS) に基づく方法 (Taghva and Gilbreth, 99; Chang and Schütze, 06)
  - 詳細なルールを記述 (Ao and Takagi, 05)
  - 定義候補のスコア付け (Adar, 04)
- 機械学習に基づくアプローチ
  - 複数のルールを機械学習で統合する (Nadeau and Turney, 05; Chang and Schütze, 06)
- 統計的解析に基づくアプローチ
  - 略語と定義の共起 (Hisamitsu and Niwa, 01; Liu and Friedman, 03; Okazaki and Ananiadou, 06)

従来研究: SaRAD (Adar, 2004)

- 複数の完全形候補をスコア付けする
  - +1: 短縮形の文字が完全形の単語の先頭に現れる毎
  - -1: 短縮形の前に余計な単語が付いている毎
  - +1: 短縮形が完全形の直後に来る場合
  - -1: 完全形の語数は短縮形の文字数を超える毎
- スコア計算例 (RNA)
  - -4: **determine the geographical** distribution of ribonucleic acid
  - 0: determine the **geographical distribution of ribonucleic acid**
  - 1: determine the geographical **distribution of ribonucleic acid**
  - 3: determine the geographical distribution of **ribonucleic acid**
- Medstractコーパスにおいて，精度95%，再現率85%

従来研究: 素性の統合 (Nadeau, 05)

- 略語定義に関する17素性をSVMや決定木で統合
  - 「略語中の文字が定義において語の先頭に一致する数」
  - 「略語中の文字が定義において大文字で記述される数」
  - 「定義と略語の間に存在する語数」
  - 「略語を構成する上で不要な語の数」
  - 「略語定義の先頭の語が前置詞，接続詞，限定詞か」
  - 「略語定義の末尾の語が前置詞，接続詞，限定詞か」
  - 「略語定義中の前置詞，接続詞，限定詞の数」
  - 「略語定義中の動詞の数」
- 126定義で学習
  - 精度92.5% 再現率88.1%（Medstractコーパス）

本研究の目標

- 従来の機械学習アプローチ
  - 素性が非常に少なく，従来のスコア付け関数と同等
  - 性能はルールベースの手法と同等以下
- 統計に基づくアプローチ
  - 性能は良いが，低頻度の略語定義が認識できない
  - ルールベースの手法と組み合わせる必要がある
- 目標
  - ルールベースを上回る略語認識システムを構築する
  - 略語定義を認識する「タガー」を構築する
    - 略語定義と略語定義ではない文字列の並びを識別
    - Conditional Random Fields (CRFs) でモデル化する

略語認識

- 入力テキスト中の略語を抽出する
  - 括弧表現を手がかりとする:　… ‘(‘ 短縮形 ‘)’ …
  - 短縮形を認識するルール (Schwartz and Hearst, 03)
  - 略語の候補
    - 多くても２語で構成される
    - 2～10文字で構成される
    - アルファベットの文字を最低でも一文字含む
    - 最初の文字は英数字
  - 候補の例

    o *We investigate the effect of thyroid transcription factor 1 (TTF-1).*

    x *The present system captures the machine sensitive words (registered in a dictionary) from the running utterances.*

略語定義認識

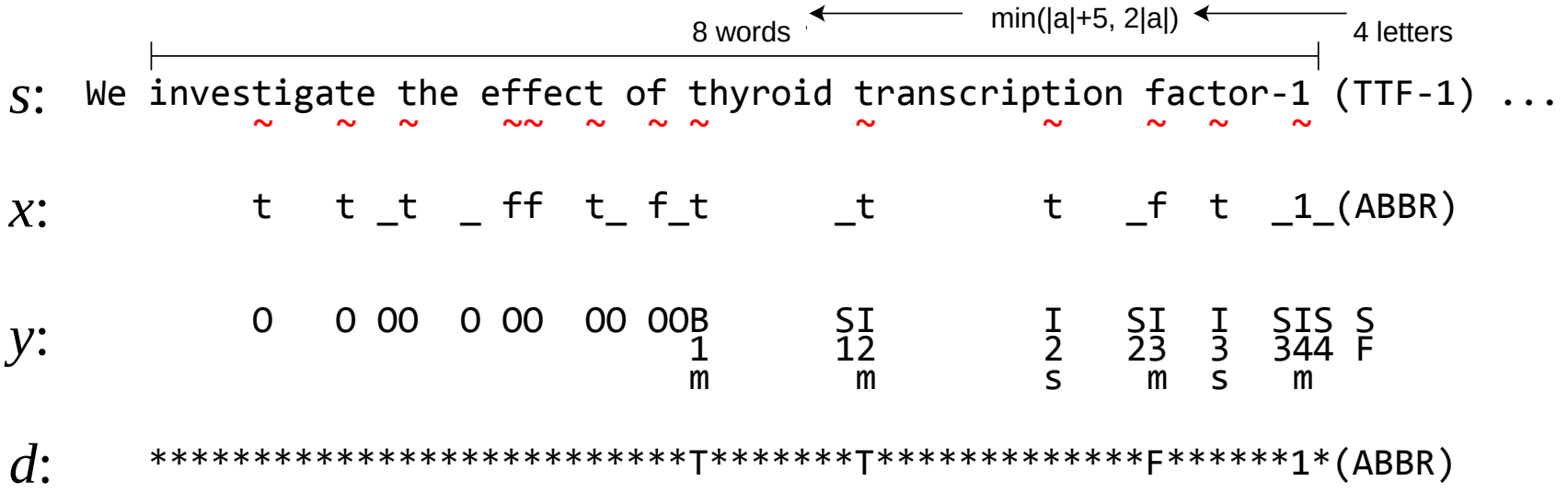
- 略語定義認識
  - 略語周辺の用語に定義が存在するか調べる
    - 存在する → 定義を抽出する
    - 存在しない → 定義を抽出しない
  - 定義として解釈できる表現は複数存在する
    - 複数の略語定義候補から最もふさわしいものを選ぶ

```
We investigate the effect of thyroid transcription factor-1 (TTF-1) ...

(1)                                          T    F  T   1
(2)                                  T            F  T   1
(3)                                  T        T    F      1
(4)                          T                    F  T   1
(5)                          T               T    F      1
(6)                          T       T            F      1
(7)                       F                  T       T   1
```

略語定義ラベリング

- 周辺の文字列 s に対して，略語文字割り当て系列 d
  - 系列 d に大量の '*' ラベルが含まれる
  - ラベル同士の依存関係（'T' → 'T'）にあまり意味がない
- 文字列 s から，略語定義に関連する部分を抜き出した記号列 x に対して，ラベル系列 y を求める



ラベルの集合

ラベル	内容	例
B*i*m	略語定義が略語 *i* 番目の文字にマッチして定義区間が開始	B1m
I*i*m	略語 *i* 番目の文字にマッチした（略語定義区間は継続）	I2m
I*i*s	略語 *i* 番目の文字にマッチした後に現れた文字をスキップ（略語定義区間は継続）	I2s
S*i*	略語 *i* 番目の文字にマッチした後にスペースが現れた（略語定義区間は継続）	S2
SF	略語自体（ラベル系列を終端）	SF
O	略語定義の外側	O

略語にマッチする特徴

略語定義の語数

略語定義の開始・終了位置

略語文字の並び順

Conditional Random Fields

- 入力記号列 $\boldsymbol{x}$ と出力ラベル系列 $\boldsymbol{y}$ の条件付き確率分布 $P(\boldsymbol{y}|\boldsymbol{x})$ をモデル化

$$P(\mathbf{y} \mid \mathbf{x}) = \frac{1}{\mathrm{Z}(\mathbf{x})} \exp\left\{\boldsymbol{\Lambda} \cdot \boldsymbol{F}(\mathbf{y}, \mathbf{x})\right\}$$

$$\mathrm{Z}(\mathbf{x}) = \sum_{\mathbf{y}} \exp\left\{\boldsymbol{\Lambda} \cdot \boldsymbol{F}(\mathbf{y}, \mathbf{x})\right\}$$

- 与えられた入力記号列 $\boldsymbol{x}$ に対する最適ラベル系列

$$\tilde{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}}\ P(\mathbf{y} \mid \mathbf{x})$$

CRFの学習

- N件のインスタンスから構成される学習データが与えられた時，対数尤度を最大にする

$$L_{\Lambda} = \sum_{i=1}^{N} \log P(\mathbf{y}^{(i)} \mid \mathbf{x}^{(i)})$$

- 対数尤度の偏微分（Gaussian事前分布を統合）が0になるときのパラメータをL-BFGSで求める

$$\frac{\partial L_{\Lambda}}{\partial \Lambda} = \sum_{i=1}^{N} \boldsymbol{F}(\mathbf{y}^{(i)} \mid \mathbf{x}^{(i)})$$

$$- \sum_{i=1}^{N} \sum_{\mathbf{y}^{(i)}} P(\mathbf{y}^{(i)} \mid \mathbf{x}^{(i)}) \boldsymbol{F}(\mathbf{y}^{(i)} \mid \mathbf{x}^{(i)}) - \frac{\Lambda}{\sigma^2} = 0$$

素性

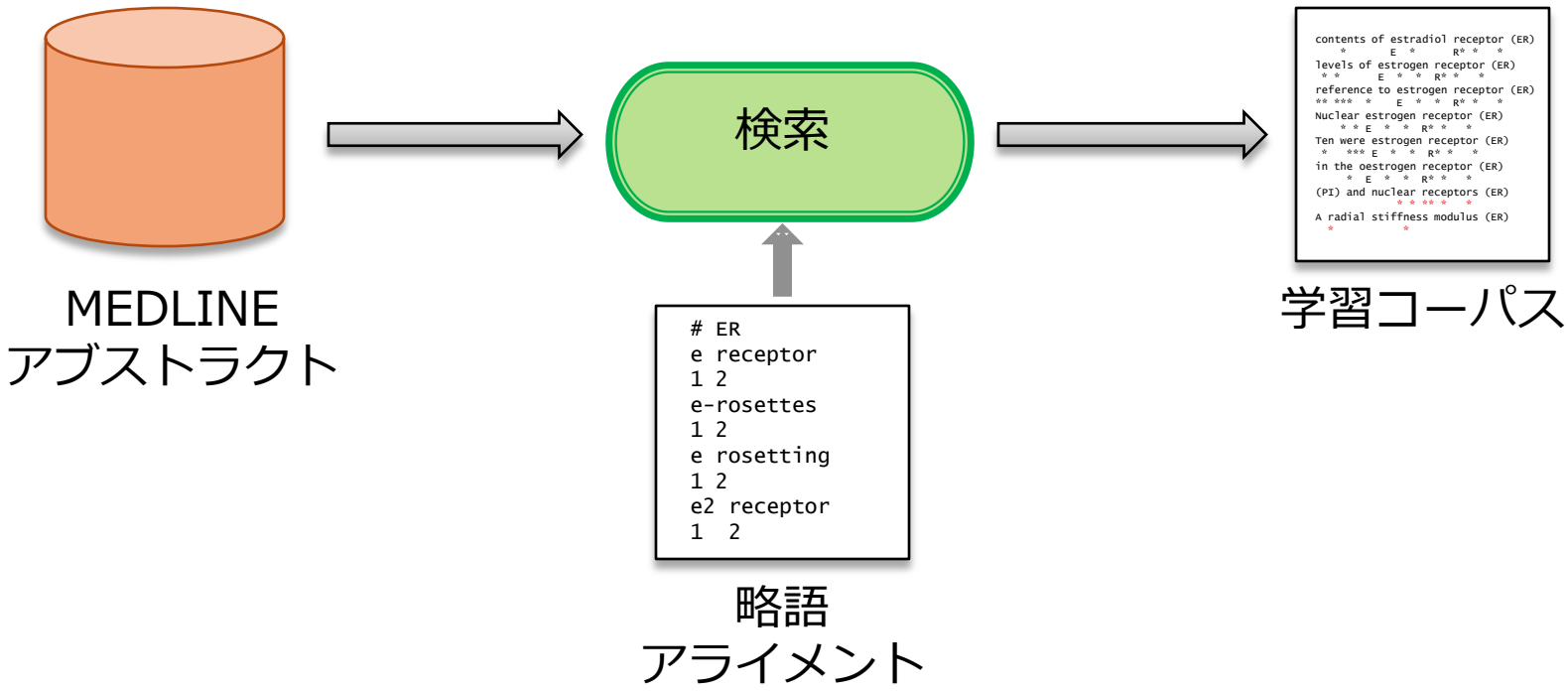
素性	オフセット	内容	例
文字	c_{-1}, c_0, c_{+1}	文字cを小文字に変換したもの	ch:0:'t'
大文字・小文字	c_{-1}, c_0, c_{+1}	文字cが大文字であるか (upper), 小文字であるか (lower)	upper:0, lower:+1
文字種別	c_{-1}, c_0, c_{+1}	文字cが英文字 (alpha), 数字 (digit), 記号 (symbol) であるか	alpha:0, alpha:+1
文字の相対位置	c_{-1}, c_0, c_{+1}	文字cがその語wにおいて先頭 (head), 末尾 (tail), 内部 (inner)	head:0, inner:+1
略語文字との一致	c_{-1}, c_0, c_{+1}	文字cが略語のp番目の文字と一致する (match=p)	match:0=1
単語	w_0	単語wを小文字に変換したもの	w:0=factor
品詞	w_{-1}, w_0, w_{+1}	単語wの品詞タグ	pos:0:NN
遷移素性	---	ラベルのbigram	O—B1m

ラベル系列に関する制約

- 今回設計したラベル系列には強い制約が存在
- 以下のラベル系列による定義候補は不適合
  - ラベル “Bm1” が系列中に2回以上出現する
  - ラベル “I2m” の後に “S3” が現れる
- ある位置 t におけるラベル割り当ては，過去 $(1, \ldots, t\text{-}1)$ に強く依存している
- 可能なラベル系列集合 $\mathbf{Y}^{(i)}$ を列挙して学習する
  - Forward-backwardやViterbiアルゴリズムは利用不可
  - ラベル系列数$|\mathbf{Y}^{(i)}|$が小さい場合は，対数尤度の偏微分を直接求めたとしても，計算量が大きくならない
  - 学習データにおける平均出力ラベル候補数$|\mathbf{Y}^{(i)}|$は5.56

学習データの半自動生成

- 5,480個の略語定義から317,349事例を自動生成
  - MEDLINEアブストラクトから “…… (略語)“ を含む文を抽出
  - 括弧表現の前が既知の定義なら正例，そうでないなら負例



作成した略語定義アライメント

- ERの261定義（=アライメント）の一部

```
# ER
e receptor
1 2
e-rosettes
1 2
e rosetting
1 2
e2 receptor
1  2
earliest recollections
1        2
early ratio
1     2
early reactions
1     2
early relapses
1     2
early repolarisation
1     2
```

```
early responders
1     2
early response
1     2
eating rate
1      2
echoreflectivity
1   2
effect ratio
1      2
ego resiliency
1   2
ejection rate
1        2
elastic recoil
1       2
elastic-resistive
1       2
electronic residue
1          2
```

```
estrogen receptor
1        2
ethionine-resistant
1         2
ethoxyresorufin
1     2
etoposide-resistant
1         2
evanescent resonance
1          2
evans' ratio
1      2
evaporation rate
1           2
event-related
1     2
everted repeat
1       2
exacerbation reactions
1            2
```

自動生成された学習セット

- ERに対して12,801事例が生成される

```
contents of estradiol receptor (ER)
    *       E  *      R* *   *
levels of estrogen receptor (ER)
 * *        E  *  *   R* *   *
reference to estrogen receptor (ER)
** ***  *    E  *  *   R* *   *
Nuclear estrogen receptor (ER)
    * * E  *  *  R* *   *
Ten were estrogen receptor (ER)
 *   *** E  *  *  R* *   *
in the oestrogen receptor (ER)
     *  E  *  *  R* *   *
(PI) and nuclear receptors (ER)
               * * ** *   *
A radial stiffness modulus (ER)
  *             *
```

同一の略語定義でも周辺環境が異なる

負例
＝略語の前が定義でない

学習インスタンスの例

```
# protein from endoplasmic reticulum ( ER )
#  ~  ~  ~ ~  ~~          ~~~       ~
% O POS:-1=IN POS:+1=IN POS:0=NN infix:0 match:0=2
% O POS:-1=IN POS:+1=IN POS:0=NN infix:0 match:0=1
% O space
% O POS:-1=NN POS:+1=JJ POS:0=IN infix:0 match:0=2
% O space
% B1m POS:-1=IN POS:+1=NN POS:0=JJ head:0 match:0=1
% S1 space
% I2m POS:-1=JJ POS:+1=-LRB- POS:0=NN head:0 match:0=2 match:1=1
% I2s POS:-1=JJ POS:+1=-LRB- POS:0=NN infix:0 match:0=1 match:-1=2
% S2 space
% SF shortform
```

略語周辺表現

属性（素性）で表現された入力系列

ラベル系列

可能なラベル系列

```
! O B1m S1 I2m S2 I2s S2 I2s I2s S2 SF
! O B1m S1 I1s S1 I1s S1 I2m I2s S2 SF
! B2m I2s S2 I2s S2 I1m S1 I1s I1s S1 SF
! O O O B2m S2 I1m S1 I1s I1s S1 SF
! O O O O O B1m S1 I2m I2s S2 SF
! B2m I2s S2 I2s S2 I2s S2 I2s I1m S1 SF
! O O O B2m S2 I2s S2 I2s I1m S1 SF
! O O O O O O O O O O SF
```

実験

- 30個の略語に対して，その定義5,480個を列挙し，アライメント情報を付与した
  - 学習データ: 24個の略語に対応する4,688個の定義
  - 評価データ: 6個の略語に対応する792個の定義
- CRFは独自に実装したもの (C/C++) を用いた
  - Intel Core2Duo 6600 (2.40GHz) CPU, 2GB MM
  - 学習時間
    - ラベル制約なしCRF (linear-chain): 約10.2時間 (295 iterations)で打ち切り, 約105～212 秒/iteration
    - ラベル制約ありCRF: 約1.8時間 (2027 iterations), 約2.2～4.3秒/iteration

評価

- 評価データ（89,101事例）における精度と再現率
  - 提案手法: 精度97.2%, 再現率97.9%
  - Linear-chain CRF: 正解なし (after 295 iterations)
  - SaRAD: 精度96.1%, 再現率95.8%
  - ※ CRFの正解はラベル全体の正解，SaRADは定義としての正解
- 提案手法は…
  - C reactive protein (PCR) のような並び順の異なる定義も抽出
  - 重みが高かった素性の例
    - “POS:+1:CC -> B1m“「次の語が等位接続詞ならB1m」,
    - “POS:0:NNS B1m”「現在の語が名詞ならばB1m」
  - 素性の95.6%が単語素性であった　→　改良の余地あり
    - “word:0:adenosin → O”, “word:0:5‘-adenosine -> B1m” みたいな素性が上位に来ても，評価データでは効果が薄い

今後の課題

- ベースラインよりも良い性能を示した
- さらなる精度向上が期待できる
  - CRFのグラフィカルモデルを工夫する
  - 素性とラベルの改良
    - 現状では略語の文字数に依存した学習モデルが生成される
  - 学習データが偏らないようにする
    - アノテートする略語のサンプリング方法を検討する
- 評価
  - 幅広い略語において評価する
  - 略語辞書構築のアプリケーションにおいて評価する