# 修　　士　　論　　文

# 素性にモーラとシラブルを用いた略語の自動推定

## Automatically Figuring Out Abbreviations Using the Numbers of Morae and Syllables as Features


指導教員　　**近山 隆 教授**

**東京大学大学院**
**工学系研究科**
**電気系工学専攻**

氏 名　　**37-086538 和田 健太**


提 出 日　　**平成22年2月9日**

# 概要

自然言語処理では略語を含む同義語の推定が非常に重要な意味を持っている．たとえば検索システムにおいて同義語をあらかじめ特定していれば，検索クエリに対する同義語で検索した結果もユーザに返すことができる．また，文章要約システムでは略語が分かっていれば原文よりもより短い要約文を生成することが可能である．

一般に同義語全般の推定はその使用法による解析など，かなり限定した手法でしか推定することができないのに対し，同義語の一種である略語は省略される前の言葉 (原語) にかなりの情報が含まれており，同義語全般の推定に対し比較的容易に推定することが可能である．

そこで本研究ではモーラとシラブルという音声学上の概念を導入し，CRF により確率的に原語から略語を自動推定する手法を提案する．従来の手法では原語と略語の表層上の文字が異なる略語は推定することが出来無かったが，本手法ではそれを改善した．さらに，推定した略語を絞り込み，尤もらしい略語を上位へと上げる手法を複数提案した．先行研究では括弧表現に限定した略語の抽出などが行われてきたが，本研究ではより広範な文書群を検索対象とできる共起尺度による絞り込みを行った．この絞り込みにより，1 位に推定した略語の $f$ 値として 0.6 を越える結果を得ることができた．

# 目 次

# 図 目 次

# 表 目 次

# 第1章　序論

## 1.1　背景

我々人間は異なる単語を用いて同じ意味を持つ文を生成することができる．例えば「会議に出る」と「会議に出席する」という二つの文は，表層上の文字と発音は異なるものの，それぞれの文が意味するものは同じである．これは文に限らず単語にも同じ事が言え，表記・発音は異なるが同じ意味を持つ語の事を同義語と呼ぶ．同義語としては「本」に対して「書物」,「病気」に対して「やまい」,「東京大学」に対して「東大」などが例として挙げられるが，このうち「東京大学」↔「東大」は同義語の一種である略語である．略語は同義語の中でも頻出するものであり，かつある程度の法則性を持って生成されている語であると考えられる．

自然言語処理では略語を含め同義語を獲得することが重要な意味を持っている．たとえば，検索システムにおいて検索キーワードとして用いるクエリが他の同義語を持つと分かっていれば，その同義語でも検索を行うことができる．実際に Google では 100 以上の言語で同義語の検索結果も提示している．また，一つの文章にて何度も用いられる比較的長い名詞というのは省略される事が多いが，文章要約システムにおいては略語があらかじめ分かっていればそれを用いて要約を行うことが出来る．一般に，同義語全般を推定することはその使用法を解析するなどの手法を取るしかなく容易には実現することができないのに対し，略語は原語の情報を一部用いることができるため，比較的容易に推定することが可能である．

ところで略語というのは人間が生成するものであるため，人間の感覚というものが重要になってくる．[1] では音韻論の立場から略語がどのように生成されているかを考察し，音声学上の概念であるモーラが略語と深い関係にあることを示した．

## 1.2　提案手法

本研究ではモーラとシラブルを条件付確率場 (Conditional Random Fields; CRF) の素性として用い，同義語の一種である略語を自動推定する手法を提案する．本研究で対象とする略語は日本語略語であり，具体的には Table 1.1 に挙げるような略語を考える．以下では省略される前の語を「原語」，省略された語を「略語」と呼ぶことにする．

さらに，推定した略語のうちより尤もらしいものを推定上位に上げるための絞り込み手法についていくつか提案し，どういった手法で絞り込むのが最も良いのかを検討する．

Table 1.1: 本研究で扱う略語の例

| 原語 | 略語 | 略語の特徴 |
|---|---|---|
| 文部科学省 | 文科省 | 「文部」「科学」「省」の頭文字を取った頭字語と呼ばれる略語 |
| 最高裁判所 | 最高裁 | 「最高」の文字を全て取った略語 |
| マイクロコンピュータ | マイコン | カタカナ略語 |
| 航空自衛隊 | 空自 | 頭文字の「航」ではなく「空」を取ったタイプの略語 |
| 原動機付自転車 | 原付 | 「自転車」が完全に抜け落ちた略語 |
| アメリカ | 米 | 原語には使われていない文字が使われている略語 |

この研究によって原語 · 略語対を獲得することにより，自然言語処理において避けられないことばの曖昧性の解決に繋がるほか，前述した検索エンジンへの応用 (特に社内文書検索) や文章要約システムへの応用が可能となる．

## 1.3 提案手法の特徴

略語を獲得する手段として，従来では主に

1. 括弧表現から略語を抽出する
2. 原語から略語を推定する

の 2 つが研究されてきた．

1 番目は新聞記事などで「東京大学 (東大)」のように略語を表記することが多い点に着目した手法であり，表層上の文字を頼りに原語と略語の対を抽出している．しかしこの手法では原語と略語の頭文字が一致しているかというような条件により略語抽出を行っているために「大阪大学 (阪大)」のような頭字語が一致しない略語を抽出できないという欠点を持っている．また，括弧表現が存在していない新語については抽出を行うことができない．

2 番目の手法では，与えられた原語から略語を推定する手法であり，本研究と同じ立場を取っている．のちに説明するこの研究では，原語を語基と呼ばれる要素に区切り，語基から不要な文字 (主に長音や撥音) を削除することにより尤もらしい略語を生成している．しかしながら，やはりこの手法でも表層上の文字を頼りに略語を生成しているため,「アメリカ合衆国」↔「米国」のような略語に使われている文字が原語に含まれていない場合は正しい略語を生成することができない．

本研究で提案する手法は，括弧表現や表層上の文字に制限されない．括弧表現に依存せずに略語を推定することができるため，比較的新しい原語についても略語を推定することが可能である．さらに，学習用データを用いることによって「アメリカ」と「米」が略語要素として対応関係にあることを知識として持ち，新たに原語要素として「アメリカ」が与えられた時に「米」を略語候補として生成し，尤もらしい略語を推定することが可能である．

## 1.4 本論文の構成

以下，本論文の構成は以下のようになっている．

**2 章** 関連研究
原語 · 略語対を獲得する手法，略語を推定する手法，さらにモーラとシラブルという音声学上のタームについて述べる．

**3 章** CRF による略語推定
本研究の主旨である略語の推定手法とその実験結果を述べる．

**4 章** 推定略語の絞り込み
推定した略語のうち，正解略語をより上位に上げるための手法として，略語長を用いた場合と検索エンジンを用いた場合の 2 種類について述べる．

**5 章** 結論
結論及び今後の課題について述べる．

# 第2章　関連研究

## 2.1　原語 · 略語対の獲得

原語と略語が一つの文書に含まれている場合，それを抽出し原語 · 略語対を獲得するのは非常に賢明な手法であると言える．先行研究としては [2–7] が挙げられる．本節では原語 · 略語対を獲得する 2 種類の先行研究 [2,3,7] について述べる．

### 2.1.1　コーパスからの略語抽出

酒井らは原語と略語の対をコーパスから抽出する手法を試みた [2,3]．この手法は以下のような手順による．

#### 略語である可能性のある名詞の判別

ある名詞の略語である可能性のある名詞を表層上の情報を用いて略語か否か判別する．「ある名詞の略語である可能性のある名詞」を以後「略語可能性名詞」と定義する．名詞 $A$ を構成する文字をすべて含み，かつその出現順序が等しい名詞 $B$ の集合を $P(A)$ とする (すなわち $B \in P(A)$)．ここで，略語可能性名詞 $A$ は以下の条件を満たすものであると定義される．

**条件 1:** 名詞 $B \in P(A)$ には名詞 $A$ が含まれていない．

**条件 2:** 名詞 $A$ と名詞 $B \in P(A)$ の先頭の文字が同一である．

**条件 3:** 名詞 $A$ に対して $|P(A)| \leq 3$．ただし，$|P(A)|$ は $P(A)$ の要素数．

例えば名詞 $A$「地銀株」は名詞 $B \in P(A)$「地方銀行株」に対して全ての条件を満たしている．

この条件を満たす名詞 $A$ は略語可能性名詞であり，それに対応する名詞 $B \in P(A)$ を原型名詞と定義する．しかし，これらの条件のみでは原型名詞が「新潟バイパス」，略語可能性名詞が「新バイパス」というような明らかに間違った対応関係も抽出してしまう．そこで，名詞の共起情報から重みを計算して正しい対応関係を判定する．

**原型名詞に対する略語の判別**

原型名詞と略語可能性名詞の対応関係をベクトル空間法を応用した手法にて判定する．意味素性辞書の知識を用いることにより，名詞に対して意味素性を割り当てることができる．例えば名詞「美術館」は「407 博物館， 367 公共機関」のように意味素性が割り当てられる．ここでは以下のような記号を定義する．

$S(t)$**:** 名詞 $t$ を含む文の集合

$R(t)$**:** 文の集合 $S(t)$ に含まれる名詞の集合

$M(R(t))$**:** 名詞集合 $R(t)$ の各名詞に対して意味素性を割り当てた場合，割り当てられた意味素性の集合

以下の手順で略語を判定する．

**Step1:** 略語可能性名詞 $A$ を含む文集合 $S(A)$ に含まれている名詞集合 $R(A)$ を抽出する．なお，$R(A)$ には $A$ も含まれる．

**Step2:** 名詞集合 $R(A)$ の各名詞の意味素性を意味素性辞書から得る．

**Step3:** 名詞 $A$ に対して $R(A)$ の各名詞に割り当てられた意味素性の重みを要素としたベクトル $V_A$ を生成する．$R(A)$ の各名詞に意味素性を割り当てた結果，意味素性集合 $M(R(A))$ が生成される．そして各要素を $m \in M(R(A))$ としてベクトルの要素である各意味素性 $m$ の重み $W(m)$ を計算する．

$$W(m \in M(R(A))) = \frac{mf(m, A) \times rank(m)}{2} \times \log \frac{N}{2 \times sf(m)}$$

ただし

$mf(m, A)$**:** 文集合 $S(A)$ に含まれている各名詞に意味素性を割り当てた場合，意味素性 $m$ が割り当てられた総数

$rank(m)$**:** 意味素性 $m$ の上位概念の数．意味素性辞書には各意味素性の上位概念が割り当てられている．

$sf(m)$**:** 対象としているコーパスにおいて，意味素性 $m$ が割り当てられている総数

$N$**:** 対象としているコーパス中の全ての文の総数

**Step4:** Step1 ∼ 3 をある略語可能性名詞 $A$ に施し，名詞 $B \in P(A)$ に対して $R(B \in P(A))$ の各名詞に割り当てられた意味素性の重みを要素としたベクトル $V_B$ を生成する．

**Step5:** 生成された意味素性を要素とした 2 つのベクトル $V_A$, $V_B$ の余弦 $\cos(\theta)$ を計算する．余弦 $\cos(\theta)$ が設定したしきい値以上であった場合，Step6 に進む．

**Step6:** 以下の 2 つの制限を満たした原型名詞と略語可能性名詞を原型名詞に対する略語であると認定する.

**制限 1:** 対象コーパスにおいて名詞 $A$ および名詞 $B \in P(A)$ が 2 回以上出現する.

**制限 2:** 名詞 $A$ のベクトルと名詞 $B \in P(A)$ のベクトルで，同一の意味素性が 4 つ以上含まれている.

この手法により酒井らは適合度 72.0%，再現率 15.4%で略語を抽出する事に成功したが,「アメリカ」↔「米」,「大阪大学」↔「阪大」のような略語は条件 2 に当てはまらず，抽出することができない．この手法では表層上の文字のみを用いているため，獲得しうる略語に強い制限があるという欠点を持っている．また，辞書からの情報に強く依存しており，比較的新しい略語に関しては全く抽出することができない.

### 2.1.2　新聞記事からの略語抽出

岡崎らもやはり原語と略語の対を獲得しようとした [7]．対象はコーパスではなく新聞記事で，以下に示す手順により略語を抽出した.

1. 新聞記事から括弧表現「$X(Y)$」を抽出する.
2. $X$ と $Y$ が相互に言い換え可能であるか SVM により判定する.
3. SVM の素性には次のようなものを用いた.
   (a) パターン「$X(Y)$」の出現回数
   (b) 語 $X$ 単独での出現回数
   (c) 語 $Y$ 単独での出現回数
   (d) $\chi^2$ による共起度
   「$X(Y)$」の共起度を $\chi^2$ 値で測ったもの
   (e) 対数尤度比による共起度
   「$X(Y)$」の共起度を対数尤度比で測ったもの
   (f) 文字の包含
   $X$ が $Y$ の全ての文字を含む場合に 1
   (g) コンテキストの分布距離
   $X$, $Y$ と係り受けを持つ語の分布の Skew Divergence
   (h) 品詞コードのペア
   南瓜が $X$ と $Y$ それぞれに付与した品詞コードを並べたもの

(i) 品詞カテゴリのペア
南瓜が $X$ と $Y$ それぞれに付与した品詞カテゴリを並べたもの

(j) 固有表現タグのペア
南瓜が $X$ と $Y$ それぞれに付与した固有表現タグを並べたもの

(k) $X$ と $Y$ の文字種
漢字，ひらがな，カタカナ，アルファベット

(l) 言い換え発生率
「$X(Y)$」の表記に対し，言い換えと推測されるものの割合

なお，コンテキストの分布距離は語 $X$ と語 $Y$ が言い換え可能であれば，それぞれの周辺に存在する語も似たような分布を示すだろうという仮説に基づいている．$X$ と係り受け関係を持つ単語の頻度分布を $P$ とし，$Y$ と係り受け関係を持つ単語の頻度分布を $Q$ としたとき，確率分布 $P$ と $Q$ の距離を Skew Divergence で測定した．

$$\mathrm{SKEW}_\alpha(P||Q) = \mathrm{KL}(P||\alpha Q + (1-\alpha)P) \tag{2.1}$$

$$\mathrm{KL}(P||Q) = \sum_i P(i)\log\frac{P(i)}{Q(i)} \tag{2.2}$$

括弧表現「$X(Y)$」が言い換え「$X \to Y$」を導入するためのものであれば，その括弧表現の後では表現「$X$」よりも「$Y$」が好んで用いられると推測される．この状況を示したものが Fig 2.1 である．文書 (a) は「欧州連合 → EU」という言い換えを定義し，括弧表現以降では「EU」という表現を多く用いているのに対し，文書 (b) では固有名詞「ベッカム」の国籍の属性値として「イングランド」を挙げており，括弧表現以降でも「ベッカム」が多く用いられている．そこで,「$X(Y)$」というパターンが出てくる文書を集め，以下の 2 つの条件を同時に満たす文書は「$X \to Y$」の言い換えであると見なす．

**条件 1:** 「$X(Y)$」のパターンが出てくる前の文において，表現 $Y$ が出現しない

**条件 2:** 「$X(Y)$」のパターンが出てきた後の文において，表現 $X$ よりも表現 $Y$ の出現頻度が高い

言い換え発生率は以下のように定義する．

$$\mathrm{PR}(X,Y) = \frac{d_{para}(X,Y)}{d(X,Y)} \tag{2.3}$$

ここで，$d_{para}(X,Y)$ は条件 1, 2 を満たす文書の数，$d(X,Y)$ は括弧表現「$X(Y)$」を含む文書の総数である．この言い換え発生率 PR は表現 $X$，$Y$ に対して $\mathrm{PR} = 0$(言い換えの発生無し) から $\mathrm{PR} = 1$(全ての括弧表現が言い換えを導入している) までの値を返す関数である．

この手法では酒井らが獲得することができなかった「大阪大学 (阪大)」のような略語も獲得することができる．しかし品詞を用いるために，略語が辞書に登録されている必要があ

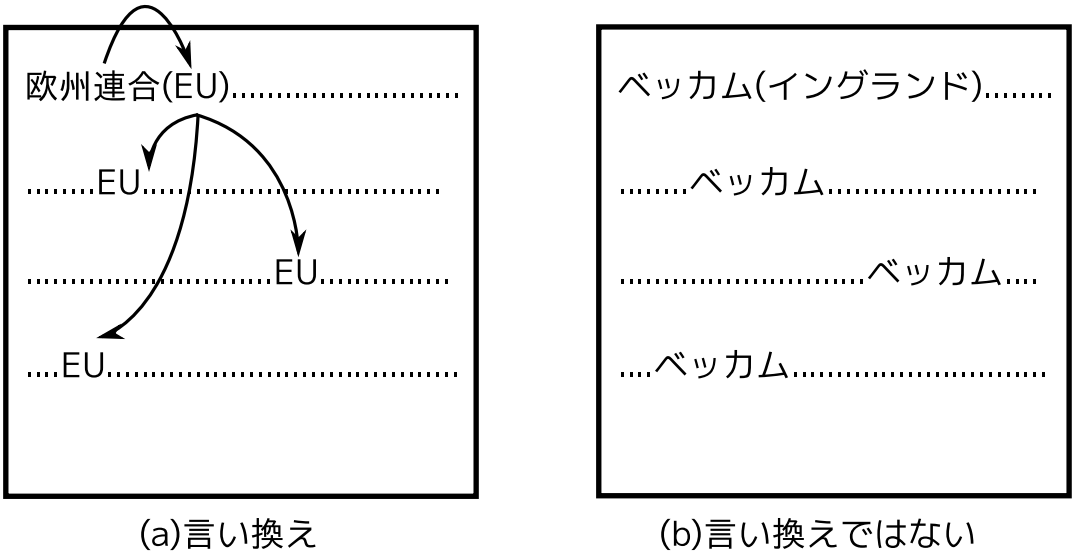


Fig 2.1: 括弧表現による言い換え

り，また，括弧表現にない略語を獲得することはできないという欠点がある．新聞記事のような形式がある程度整っている文章では有効だが，昨今の Web 上にあるような文章を対象とした場合，この手法が有効であるかには疑問符をつけざるを得ない．

## 2.2　略語の推定

原語が与えられたとき，その原語から略語を自動推定する手法も提案されている [8–10]．ここでは村山ら [8] の自動推定手法について説明する．

略語生成のモデルとして Noisy-Channel Model を用いている．ベイズの定理を用いて原語 $\boldsymbol{x}$ から尤もらしい略語 $\boldsymbol{y}$ を推定する．尤もらしい略語を $\hat{\boldsymbol{y}}$ とすると

$$\begin{aligned} \hat{\boldsymbol{y}} &= \operatorname*{argmax}_{\boldsymbol{x}} P(\boldsymbol{y}|\boldsymbol{x}) \\ &= \operatorname*{argmax}_{\boldsymbol{x}} \frac{P(\boldsymbol{x}|\boldsymbol{y})P(\boldsymbol{y})}{P(\boldsymbol{x})} \\ &= \operatorname*{argmax}_{\boldsymbol{x}} P(\boldsymbol{x}|\boldsymbol{y})P(\boldsymbol{y}) \end{aligned} \tag{2.4}$$

となる．ここで，$P(\boldsymbol{y})$ は略語らしさ，すなわち略語の言語モデルであり，$P(\boldsymbol{x}|\boldsymbol{y})$ は略語から原語への変換モデルである．

このようにして原語 $\boldsymbol{x}$ が与えられた時の条件付確率 $P(\boldsymbol{y}|\boldsymbol{x})$ を最大化するような $\boldsymbol{y}$ を選ぶことにより，尤もらしい略語を推定することができる．しかしこの手法は生成モデルを用いており，一般に識別モデルと比べて精度の面で劣るという欠点がある．実際，この研究では推定した略語の上位一位では再現率が 0.135，上位 10 位で 0.518 であり，高精度の手法であるとは言い難い．また，表層上の文字のみを用いて略語を推定しているので，やはり「アメリカ」↔「米」というような略語は推定することができない．

## 2.3　モーラとシラブル

略語の抽出 · 生成とは別に，略語とモーラ · シラブルの関係を調べた論文もある [11–13]．まずモーラとシラブル [11] について簡単に説明する．

モーラは拍とも呼ばれ，日本語の音のリズムを表すものである．俳句や短歌における 5·7·5 や 5 · 7 · 5 · 7 · 7 といった数は，このモーラ数を数え上げたものである．また，シラブルは一つの母音を中心とした音のかたまりの事であり，音節とも呼ばれている．Table 2.1 に実際に単語がどのようなモーラとシラブルで構成され，計何モーラ，何シラブルを持っているかを示す．なお，括弧内の数字がモーラ数及びシラブル数である．

Table 2.1: 単語のモーラ数とシラブル数

| 語 | モーラ | シラブル |
|---|---|---|
| はこ | は / こ (2) | は / こ (2) |
| しゃしん | しゃ / し / ん (3) | しゃ / しん (2) |
| きって | き / っ / て (3) | きっ / て (2) |
| おうさま | お / う / さ / ま (4) | おう / さ / ま (3) |

モーラでは拗音はその前の音とまとめて数えられるため,「しゃしん」は「しゃ」で 1 モーラとなる．シラブルでは拗音だけでなく撥音，長音，促音も前の音とまとめて数えられるため,「しゃしん」は「しゃ / しん」のように 2 シラブルとなる．

鈴木は外来語略語とモーラ · シラブルの関係について調査した [12]．計 607 語の略語のモーラ分布は単一語，複合語別に Table 2.2 のようになった．また，原語と略語のモーラ数との相関関係と割合は Table 2.3 に示すようになっている．Table 2.3 によると，全体的に右下がりの傾向を示し，2 モーラ語は原語のマイナス 1 モーラ語から始まり，略語のモーラ数が増えると徐々にマイナス 2 モーラ，3 モーラ，4 モーラと差が広がっている．略語の 2 モーラ語から 7 モーラ語までは原語から 3 モーラ短縮されると最も略語化の頻度が高くなる (Table 2.3 下線部)．

これらをまとめると以下に示すような事が明らかになった．

1. 略語のモーラ数は 2 モーラから 10 モーラである

Table 2.2: 略語のモーラ別分布

| モーラ数 | 単一語 | 複合語 | 計 (%) |
|---|---|---|---|
| 2 | 71 | 17 | 88 (14.50) |
| 3 | 66 | 89 | 155 (25.54) |
| 4 | 43 | 221 | 264 (43.49) |
| 5 | 2 | 51 | 53 (8.73) |
| 6 | 0 | 28 | 28 (4.61) |
| 7 | 0 | 2 | 2 (0.33) |
| 8 | 0 | 1 | 1 (0.16) |
| 9 | 0 | 1 | 1 (0.16) |
| 計 | 182 (29.98) | 425 (70.02) | 607 (100) |

Table 2.3: 原語のモーラ数別分布

| 略語モーラ数＼原語モーラ数 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 ～ | 計 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 1 | 17 | 35 | 22 | 8 | 2 | 3 | | | | | | 88 |
| 3 | | | 24 | 51 | 47 | 14 | 12 | 3 | 2 | 0 | 0 | 2 | 155 |
| 4 | | | 6 | 47 | 61 | 53 | 33 | 26 | 18 | 9 | 5 | 6 | 264 |
| 5 | | | | | 6 | 21 | 12 | 6 | 0 | 4 | 3 | 1 | 53 |
| 6 | | | | | 1 | 3 | 8 | 8 | 4 | 2 | 1 | 1 | 28 |
| 7 | | | | | | | 1 | 6 | 2 | 4 | 0 | 2 | 15 |
| 8 | | | | | | | | | | 2 | | | 2 |
| 9 | | | | | | | | | | | | 1 | 1 |
| 10 | | | | | | | | | | | | 1 | 1 |
| 計 | 1 | 17 | 65 | 120 | 123 | 93 | 69 | 49 | 26 | 22 | 10 | 12 | 607 |

2. 略語のモーラ数は 4 モーラ (43.5%) と 3 モーラ (25.5%) のものが多い
3. 略語は原語から 3 モーラ短縮するものが最も多い (29.0%)
4. 複合語の頭字語略は 2 モーラ+2 モーラのパターンが最も多い (83.6%)
5. 3 モーラの略語は 2 モーラ 1 シラブル+1 モーラで構成されている.
6. 2 シラブル 2 モーラで始まる略語のパターンが最も多い (52.1%)
7. 略語の最低成立条件は 2 モーラ 1 シラブルである.
8. 略語化に際し，語 · シラブル · モーラのどのレベルの機能が優先されるのかはまだ明確ではない

これらは外来語略語に関する調査ではあるが，この調査により略語とモーラ · シラブルの間には何らかの関係があることが明確になった．その関係は完全に明らかになったわけではないが，モーラとシラブルを略語推定の際に用いることは妥当であると言える.

## 2.4 CRF

本研究で用いる CRF(条件付確率場; Conditional Random Fields) について述べる．CRF はコスト最小法を一般化した手法であり，後述する日本語形態素解析のような自然言語処理や，バイオインフォマティクスの分野で用いられている [14].

### 2.4.1 構造タグ付け問題

CRF は観測された構造データ $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ に対するラベル $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)$ への写像を行う問題である.

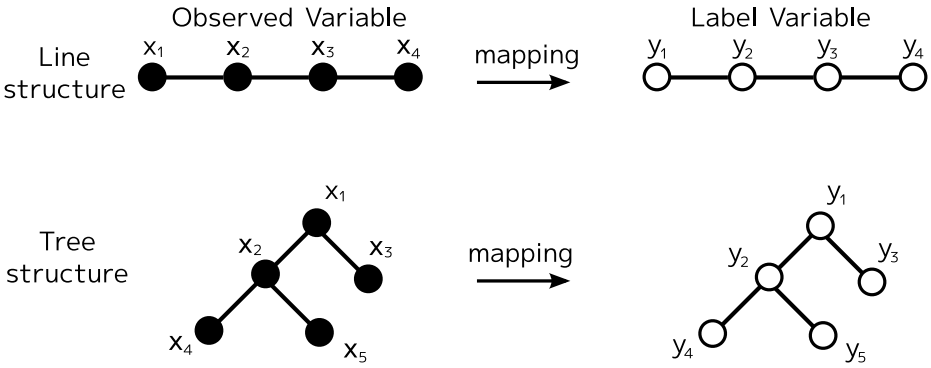


Fig 2.2: 構造ラベル付与問題

Fig 2.2 では列構造および木構造に対してラベルを付加している．左側が観測変数 $\boldsymbol{x}$ で，これに右側のラベル変数 $\boldsymbol{y}$ を付与する．具体的には Fig 2.3 のような形で用いられる．Fig

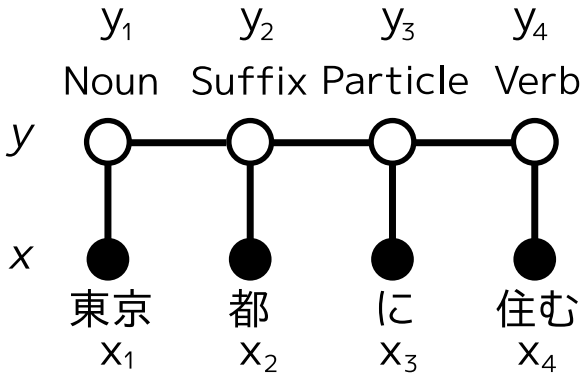


Fig 2.3: 日本語形態素解析における品詞ラベル付け

2.3 は，構造ラベル付与問題を用いた日本語形態素解析における品詞ラベル付けを示したものである [15]．形態素解析とはある文章を適切な形態素 (意味を持つ最小の語) に区切り，それに品詞を付加する問題であるが，これは列構造に対する構造ラベル付与問題として考えることができる．この場合では観測変数

$$\begin{aligned} \boldsymbol{x} &= (x_1, x_2, x_3, x_4) \\ &= (\text{東京}, \text{都}, \text{に}, \text{住む}) \end{aligned}$$

という文章に対して適切な品詞 (ラベルに相当)

$$\begin{aligned} \boldsymbol{y} &= (y_1, y_2, y_3, y_4) \\ &= (\text{名詞}, \text{接尾辞}, \text{助詞}, \text{動詞}) \end{aligned}$$

を付与するのが正しい.

### 2.4.2　教師付き学習

構造ラベル付与問題は，パラメータを推定するにあたり教師付き学習を行う．正しいラベルが付与されているデータ (学習用データ) を元にしてモデルのパラメータを推定し，ラベル無しデータに対してラベルを付与する．Fig 2.4 を例に取ると，$\boldsymbol{x}^{(1)} = (\text{東京}, \text{都}, \text{に}, \text{住む})$ に対してラベル $\boldsymbol{y}^{(1)} = (\text{名詞}, \text{接尾辞}, \text{助詞}, \text{動詞})$ が付与されているデータ，$\boldsymbol{x}^{(2)} = (\text{晩}, \text{ご飯}, \text{を}, \text{食べ}, \text{た})$ に対してラベル $\boldsymbol{y}^{(2)} = (\text{名詞}, \text{名詞}, \text{助詞}, \text{動詞}, \text{助動詞})$ が付与されているデータ，というような学習用データを大量に集め，それらからパラメータを推定し，$\boldsymbol{x} = (\text{試験}, \text{に}, \text{合格}, \text{する})$ というラベルが付与されていないデータにラベルを付与する事が目的である，この例では $\boldsymbol{y} = (\text{名詞}, \text{助詞}, \text{動詞}, \text{助詞})$ がラベルとして付与されれば正解である.

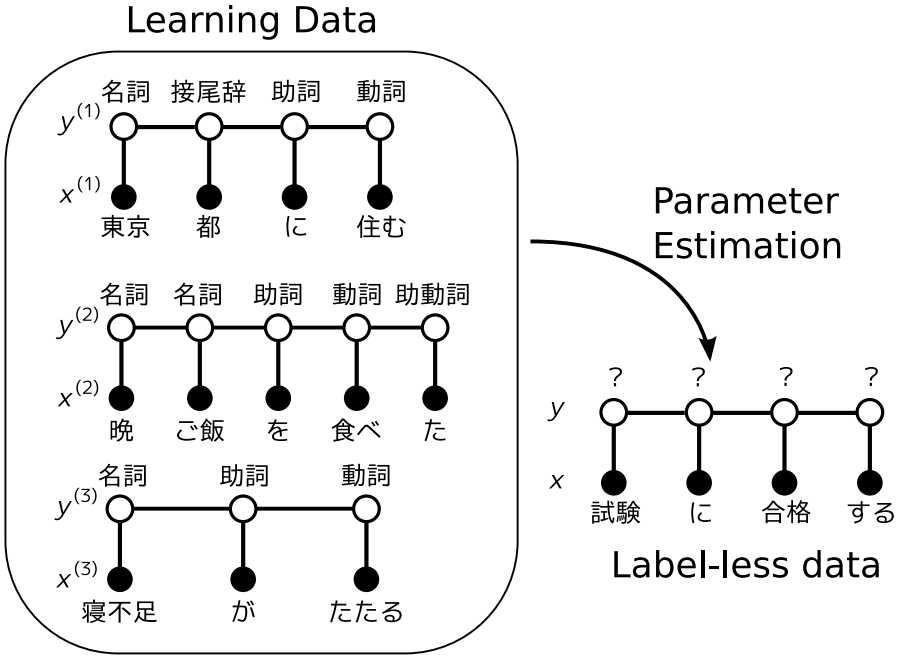


Fig 2.4: 教師付き学習

構造ラベル付与問題には識別モデルと生成モデルの二種類がある．

**識別モデル**：条件付確率 $P(\boldsymbol{y}|\boldsymbol{x})$ を直接モデル化し，$P(\boldsymbol{y}|\boldsymbol{x})$ によって未知の $\boldsymbol{x}$ に対する $\boldsymbol{y}$ を求める．CRF や隠れマルコフパーセプトロンなど．

**生成モデル**：ベイズの定理を用い，$P(\boldsymbol{x}|\boldsymbol{y})$ を学習する．隠れマルコフモデルなどがある．

識別モデルも生成モデルも求めたいものは同じであるが，識別モデルは直接解を求められ，生成モデルと比べて一般的に精度が高いというメリットがある．そこで，本手法では識別モデルである CRF を用いることとする．

# 第3章　CRFによる略語推定

前述した関連研究を踏まえ，本研究ではCRFの素性にモーラとシラブルを用い，識別モデルにより略語の自動推定を行う手法を提案する．この手法では「アメリカ」↔「米」のような原語に含まれていない文字が略語に使われるパターンや,「大阪大学」↔「阪大」のような頭文字が一致しない略語でも推定することができる．また，略語を推定するために「$X(Y)$」のような括弧表現に依存することがない．さらに，モーラとシラブルを考慮することにより，より人間の感覚にマッチした略語を推定する事ができる．

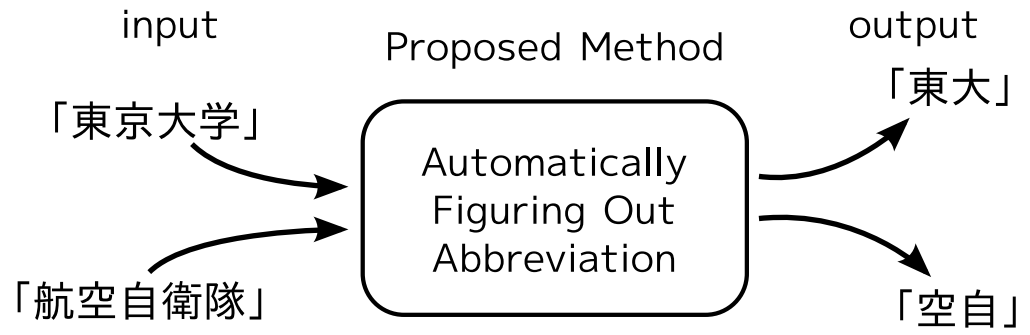


Fig 3.1: 提案手法

## 3.1　略語の推定方法

本研究で提案する手法の具体的な手順を述べる．複合語であればまず原語$\boldsymbol{x}$を$(x_1, x_2, x_3, \cdots)$と区切り，次にそれに対してCRFを用いて略語ラベル$\boldsymbol{y} = (y_1, y_2, y_3, \cdots)$を付与することにより略語の推定を行う．Fig 3.2にその例を示す．ここでは「住民基本台帳ネットワークシステム」に対する略語を推定する事を考える.「住民基本台帳ネットワークシステム」は複合語であるので，Fig 3.2に示すよう,「住民-基本-台帳-ネットワーク-システム」のように原語$\boldsymbol{x}$を$(x_1, x_2, x_3, x_4, x_5) =$ (住民, 基本, 台帳, ネットワーク, システム) と区切ることができる．これはMeCab [16] によって実現できる．その後，可能性として考えられる略語候補$y_i$を生成し，$y_i$の組合わせが尤もらしいものをCRFによって評価し略語を推定する．略語候補$y_i$には次のものを用いた．

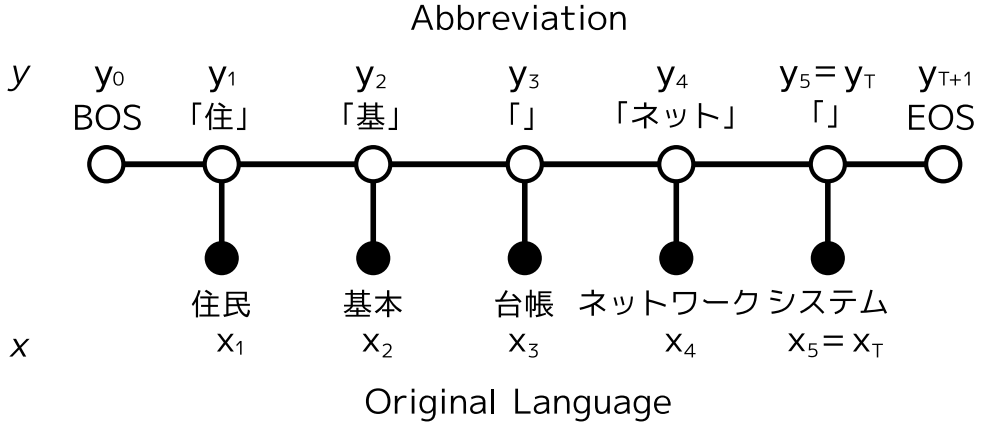


Fig 3.2: 略語推定

1. $x_i$ - $y_i$ の組み合わせが学習用データにある場合はその $y_i$ を用いる
2. $x_i$ そのものを $y_i$ に用いる
3. $y_i = \text{Null}$

この 1 が本研究の大きな特徴である．学習用データに「アメリカ」↔「米」という省略関係を持つ原語 · 略語対が 1 つでも存在していれば,「アメリカ」という文字を含む未知の原語から略語を推定する場合に $x_i =$ (アメリカ) に対する略語要素の候補として $y_i =$ (米) を提示することができる[1].

CRF の素性としては原語 · 略語の特徴を表すものとして次のものを用いる．

**観測素性** $f_O$ ：原語 · 略語間の特徴を表す．ある原語 $x_i$ に対してどのような略語 $y_i$ が付加されるかを示す．

**遷移素性** $f_T$ ：略語間のモーラ · シラブルのつながりを表す．略語 $y_i$ の持つモーラ数に続いて，$y_{i+1}$ が何モーラであるかを示す．シラブルについても同様である．

具体的には Fig 3.3 に示すように，観測素性 $f_O$ は「住民」に対して「住」を付加する素性であり，遷移素性 $f_T$ は「住」のモーラ数及びシラブル数 (2 モーラ 1 シラブル) に続く「基」が何モーラ, 何シラブルを持っているか (1 モーラ 1 シラブル) を表す素性である．

[1] このような原語 · 略語要素のペアが学習用データに全く含まれていない場合はこのような推定を行うことができない．しかし，我々人間もこのペアを知識として持っている場合しか適切な略語を推定することができないため，この手法の妥当性が無いという訳ではない．

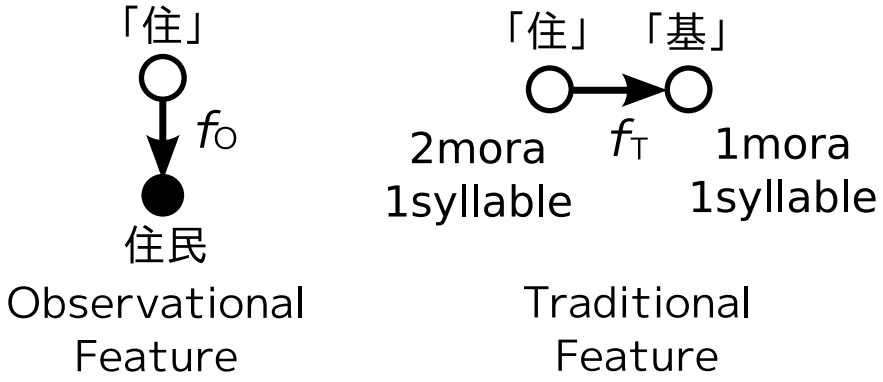


Fig 3.3: CRF の素性

素性の集合を $\boldsymbol{F}$ とし，これらの素性 $f_i \in \boldsymbol{F}$ が原語と略語のペア $(\boldsymbol{x}, \boldsymbol{y})$ にて成立する回数を $\phi_{f_i}$ とする．また，素性の重要度を考え，それを $\theta_{f_i}$ とする．さらに，$\phi_{f_i}$ および $\theta_{f_i}$ を並べたベクトルを $\Phi(\boldsymbol{x}, \boldsymbol{y})$ ならびに $\Theta$ と表記する．

$$\Phi(\boldsymbol{x}, \boldsymbol{y}) = (\phi_{f_1}(\boldsymbol{x}, \boldsymbol{y}), \phi_{f_2}(\boldsymbol{x}, \boldsymbol{y}), \cdots) \tag{3.1}$$

$$\Theta = (\theta_{f_1}, \theta_{f_2}, \cdots) \tag{3.2}$$

この $\Theta$ が CRF におけるパラメータである．なお，この $\Theta$ には確率的な制約は存在していない．

CRF を用いた略語の推定を行うにあたり，まず式 (3.3) に示す確信度を定義する．

$$C_r(\boldsymbol{x}, \boldsymbol{y}) \stackrel{\text{def}}{=} < \Theta, \Phi(\boldsymbol{x}, \boldsymbol{y}) > \tag{3.3}$$

$$= \sum_{f \in \boldsymbol{F}} \theta_f \phi_f(\boldsymbol{x}, \boldsymbol{y}) \tag{3.4}$$

これはつまり，ある素性 $f_i$ が推定略語において成立する回数 $\phi_f$ とその重要度 $\theta_f$ の積和であり，推定した略語が尤もらしいほど確信度は増加する．しかしこの確信度は負値をとることもあれば，1 を越えることもある．そこで指数の肩に乗せることにより条件付確率分布 $P(\boldsymbol{y}|\boldsymbol{x})$ を定義する．

$$P(\boldsymbol{y}|\boldsymbol{x}) \stackrel{\text{def}}{=} \frac{\exp < \Theta, \Phi(\boldsymbol{x}, \boldsymbol{y}) >}{\sum_{y \in Y} \exp < \Theta, \Phi(\boldsymbol{x}, \boldsymbol{y}) >} \tag{3.5}$$

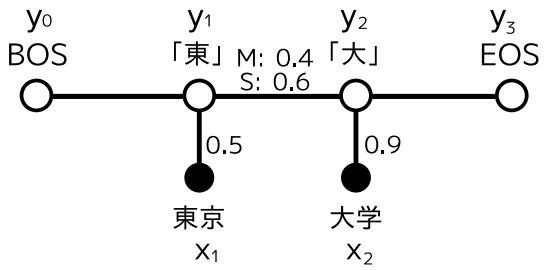


Fig 3.4: 「東大」を推定した場合

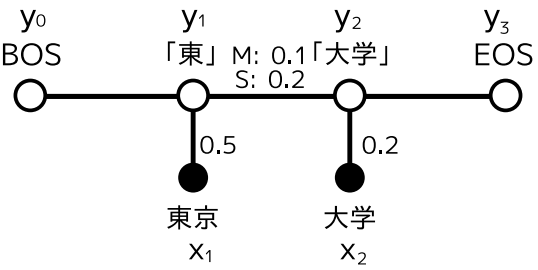


Fig 3.5: 「東大学」を推定した場合

式 (3.5) を最大化する際に分母は無関係であるため，式 (2.4) に代入することにより

$$\begin{aligned}
\hat{\boldsymbol{y}} &= \underset{\boldsymbol{y}\in Y}{\operatorname{argmax}} P(\boldsymbol{y}|\boldsymbol{x}) \\
&= \underset{\boldsymbol{y}\in Y}{\operatorname{argmax}} \frac{\exp < \Theta, \Phi(\boldsymbol{x},\boldsymbol{y}) >}{\sum_{y\in Y} \exp < \Theta, \Phi(\boldsymbol{x},\boldsymbol{y}) >} \\
&= \underset{\boldsymbol{y}\in Y}{\operatorname{argmax}} \exp < \Theta, \Phi(\boldsymbol{x},\boldsymbol{y}) > \\
&= \underset{\boldsymbol{y}\in Y}{\operatorname{argmax}} < \Theta, \Phi(\boldsymbol{x},\boldsymbol{y}) > \qquad (3.6)
\end{aligned}$$

となる．結果として，式 (3.6) を満たすような $\hat{\boldsymbol{y}}$ を求める (つまり確信度が最大になる $\boldsymbol{y}$ を求める) ことにより，原語 $\boldsymbol{x}$ から略語 $\boldsymbol{y}$ を推定することができる．

ここで「東京大学」の略語を推定する場合を例として挙げる．まず Fig 3.4 のように「東大」を推定した場合を考える.「東京」-「東」および「大学」-「大」の観測素性がそれぞれ 0.5 ポイント，0.9 ポイントの重要度 $\theta_f$ が与えられているとする．また,「東 (トウ)」(2 モーラ 1 シラブル),「大 (ダイ)」(2 モーラ 1 シラブル) の遷移素性がモーラでは 0.4 ポイント，シラブルでは 0.6 ポイントの重要度が与えられているとする．このとき「東京大学」から「東大」を推定した場合の確信度 $C_r(\boldsymbol{x},\boldsymbol{y})$ は

$$\begin{aligned}
C_r(\text{東京 / 大学}, \text{東 / 大}) &= (0.4+0.6)+(0.5+0.9) \\
&= 2.4
\end{aligned}$$

となる．一方，Fig 3.5 のように「東京大学」から「東大学」を推定した場合を考える．Fig 3.4 の場合と同じように観測素性が 0.5 ポイント，0.2 ポイント，遷移素性が 0.1 ポイント，0.2 ポイントで与えられていたとすると，確信度 $C_r(\boldsymbol{x},\boldsymbol{y})$ は

$$\begin{aligned}
C_r(\text{東京 / 大学}, \text{東 / 大学}) &= (0.1+0.2)+(0.5+0.2) \\
&= 1.0
\end{aligned}$$

となる．確信度 $C_r(\boldsymbol{x},\boldsymbol{y})$ が高いのは「東大」を推定した場合であり，結果として「東京大学」からは「東大」が尤もらしい略語として推定されることとなる．

## 3.2 パラメータの学習

次に，CRF を用いた場合のパラメータ学習について説明する．N 対の学習用データ，つまり原語 $\boldsymbol{x}$ と略語 $\boldsymbol{y}$ のペア $(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})$ $(i = 1, 2, \cdots, N)$ が与えられたとき，CRF では学習用データをもっともよく再現するようなパラメータを求めることになる．

学習用データに対する尤度は

$$\prod_{i=1}^{N} P(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(i)}; \Theta) \tag{3.7}$$

となる．したがって，これを最大化するようなパラメータ $\hat{\Theta}$ は

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \prod_{i=1}^{N} P(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(i)}; \Theta) \tag{3.8}$$

によって決定される．対数尤度によって表すと

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \sum_{i=1}^{N} \log P(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(i)}; \Theta) \tag{3.9}$$

$$= \underset{\Theta}{\operatorname{argmax}} \sum_{i=1}^{N} \log \frac{\exp < \Theta, \Phi(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) >}{\sum_{\boldsymbol{y} \in Y} \exp < \Theta, \Phi(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) >} \tag{3.10}$$

となる．

CRF において学習データに対する対数尤度を最大にするパラメータを求める．

$$\log P(\boldsymbol{y}|\boldsymbol{x}; \Theta) =< \Theta, \Phi(\boldsymbol{x}, \boldsymbol{y}) > - \log \left( \sum_{\boldsymbol{y} \in Y} \exp < \Theta, \Phi(\boldsymbol{x}, \boldsymbol{y}) > \right) \tag{3.11}$$

であるので，最大化のためにパラメータ $\Theta$ で偏微分すると

$$\begin{aligned} A &= \frac{\partial \sum_{i=1}^{N} \log P(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(i)}; \Theta)}{\partial \Theta} \\ &= \sum_{i}^{N} \frac{\partial}{\partial \Theta} \left\{ < \Theta, \Phi(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) > - \log \left( \sum_{\boldsymbol{y} \in Y} \exp < \Theta, \Phi(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) > \right) \right\} \\ &= \sum_{i}^{N} \left( \Phi(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) - \frac{\frac{\partial}{\partial \Theta} \sum_{\boldsymbol{y} \in Y} \exp < \Theta, \Phi(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) >}{\sum_{\boldsymbol{y} \in Y} \exp < \Theta, \Phi(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) >} \right) \end{aligned} \tag{3.12}$$

となる．ここで

$$\begin{aligned}
&\frac{\partial}{\partial\Theta}\sum_{\boldsymbol{y}\in Y}\exp<\Theta,\Phi(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)})>\\
=&\sum_{\boldsymbol{y}\in Y}\frac{\partial}{\partial\Theta}\exp<\Theta,\Phi(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)})>\\
=&\sum_{\boldsymbol{y}\in Y}\exp<\Theta,\Phi(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)})>\cdot\left(\frac{\partial}{\partial\Theta}<\Theta,\Phi(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)})>\right)\\
=&\sum_{\boldsymbol{y}\in Y}\Phi(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)})\exp<\Theta,\Phi(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)})> \qquad (3.13)
\end{aligned}$$

であるため，式 (3.12) に代入すると

$$\begin{aligned}
A =& \sum_{i=1}^{N}\left(\Phi(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)})-\frac{\sum_{\boldsymbol{y}\in Y}\Phi(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)})\exp<\Theta,\Phi(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)})>}{\sum_{\boldsymbol{y}\in Y}\exp<\Theta,\Phi(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)})>}\right)\\
=& \sum_{i=1}^{N}\left(\Phi(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)})-\sum_{\boldsymbol{y}\in Y}\Phi(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)})P(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(i)};\Theta)\right) \qquad (3.14)
\end{aligned}$$

となる．これは解析的に解を求めることはできないが，数値計算により解を得ることが出来る．単純な方法としては

$$\Theta^{new}\leftarrow\Theta^{old}+\eta\frac{\partial\sum_{i=1}^{N}\log P(\boldsymbol{y}^{(i)},\boldsymbol{x}^{(i)};\Theta)}{\partial\Theta}\Big|_{\Theta=\Theta^{old}} \qquad (3.15)$$

のようにしてパラメータを改善していく手法がある．

## 3.3 実験結果

実験には，wikipedia の漢字略語の項目に上げられているものを学習用データとして用いた (349 語)．評価には [17] の漢字略語の項目のうち，学習用データに含まれていない 70 語を用いた．また，学習用データの形式は Fig 3.6 のように作った．

再現率 (recall)，適合度 (precision)，$f$ 値を式 (3.16) ∼ (3.18) とした場合の実験結果を Fig 3.7 ∼ 3.9 に示す．CRF が本研究の結果，Noisy-Channel Model が先行研究である．なお，横軸は全て「上位 $n$ 位」に推定された略語の順位である．

$$recall(n) = \frac{\text{上位 } n \text{ 位の出力のうち，正解に含まれる数}}{\text{正解の略語数}} \qquad (3.16)$$

$$precision(n) = \frac{\text{上位 } n \text{ 位の出力のうち，正解に含まれる数}}{\text{上位 } n \text{ 位の全出力数}} \qquad (3.17)$$

$$f\text{-score}(n) = \frac{2\cdot precision(n)\cdot recall(n)}{recall(n)+precision(n)} \qquad (3.18)$$

learning_data.dat

```
#OP, Abb, Mora, Syllable, kana
BOS
東京, 東, 2, 1, トー
大学, 大, 2, 1, ダイ
EOS
BOS
国民, 国, 2, 2, コク
健康, Null, 0, 0, Null
保険, 保, 1, 1, ホ
EOS
```

Fig 3.6: 学習用データの形式

Fig 3.7 ～ 3.9 から本手法による略語推定では $f$ 値では先行研究を下回っていることが分かる．しかし冒頭に述べたように「アメリカ」↔「米」のような原語に含まれていない文字が略語に使われるパターンや,「大阪大学」↔「阪大」のような頭文字が一致しない略語でも本手法では推定することが可能であり，実際に「日本美術院展覧会」↔「院展」や「企業短期経済観測」↔「短観」といった略語を推定することができた.

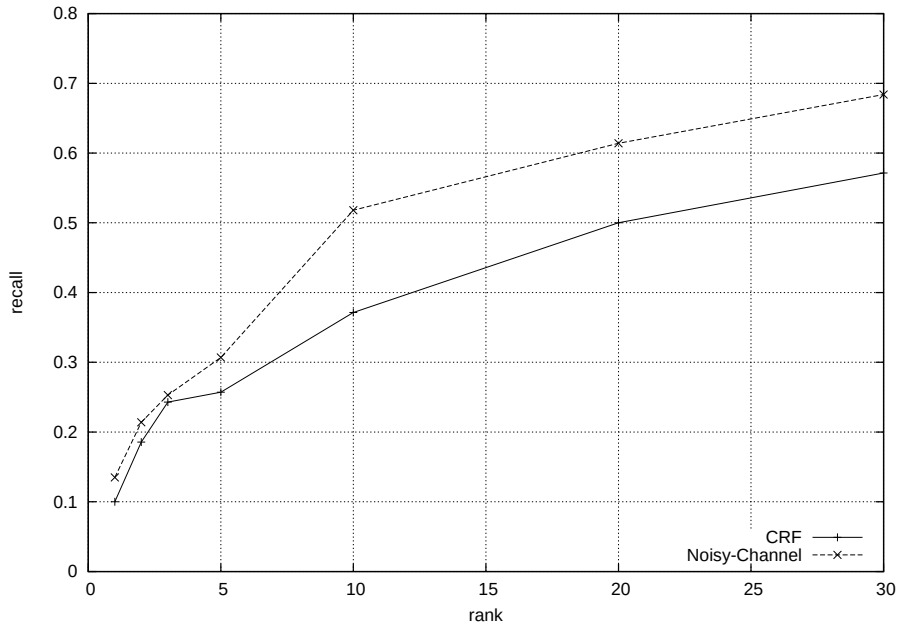


Fig 3.7: CRF による略語推定 (再現率)

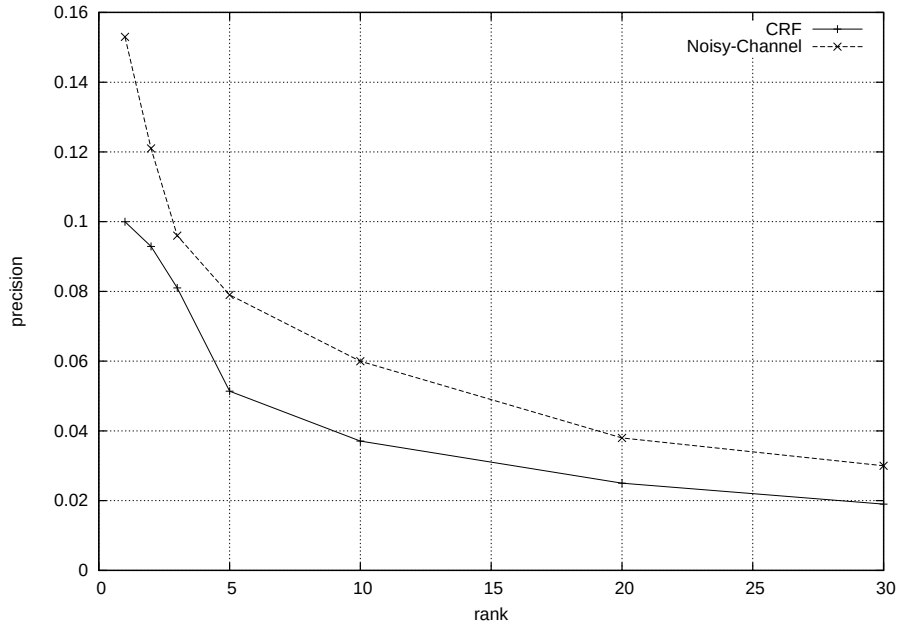


Fig 3.8: CRF による略語推定 (適合度)

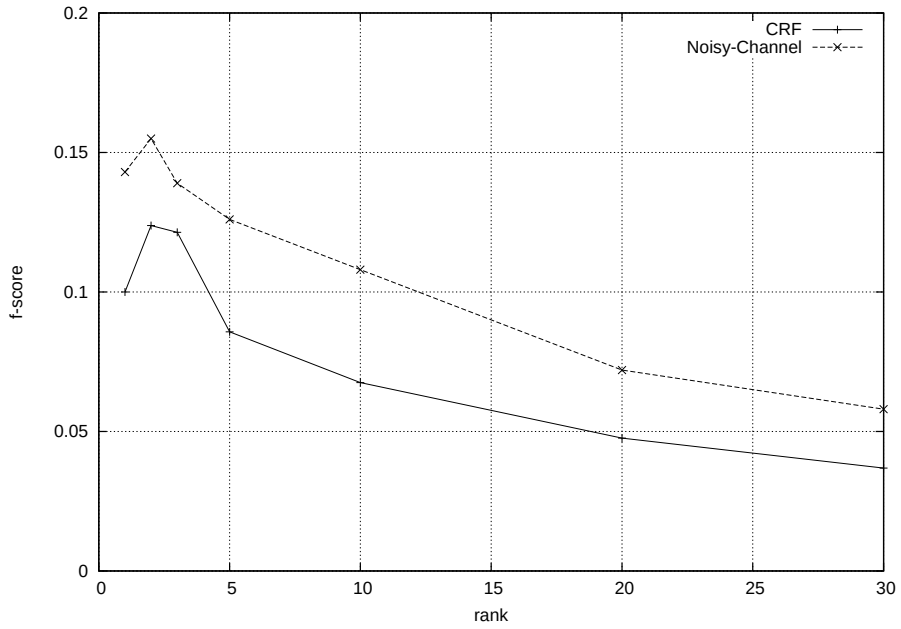


Fig 3.9: CRF による略語推定 ($f$ 値)

# 第4章　推定略語の絞り込み

前章では略語の推定手法を提案した．しかし，1 位に推定した略語のうち正しい略語は 10%であり，非常に推定の適合度が低い．そこで本節では推定した略語を絞り込み．正解略語をより上位に上げるための手法を提案する．

## 4.1　略語長による絞り込み

### 4.1.1　提案手法

CRF による略語の推定では，略語全体のモーラ数・シラブル数を考慮していない．CRF により推定した略語のモーラ数とシラブル数を調べたものが Fig 4.1, 4.2 である． それぞれ

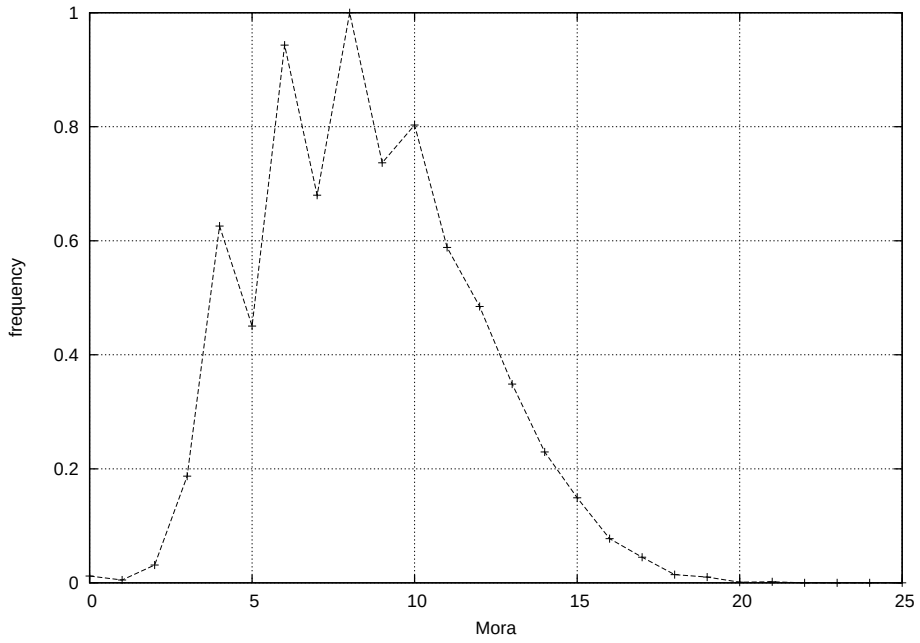


Fig 4.1: 推定略語のモーラ分布

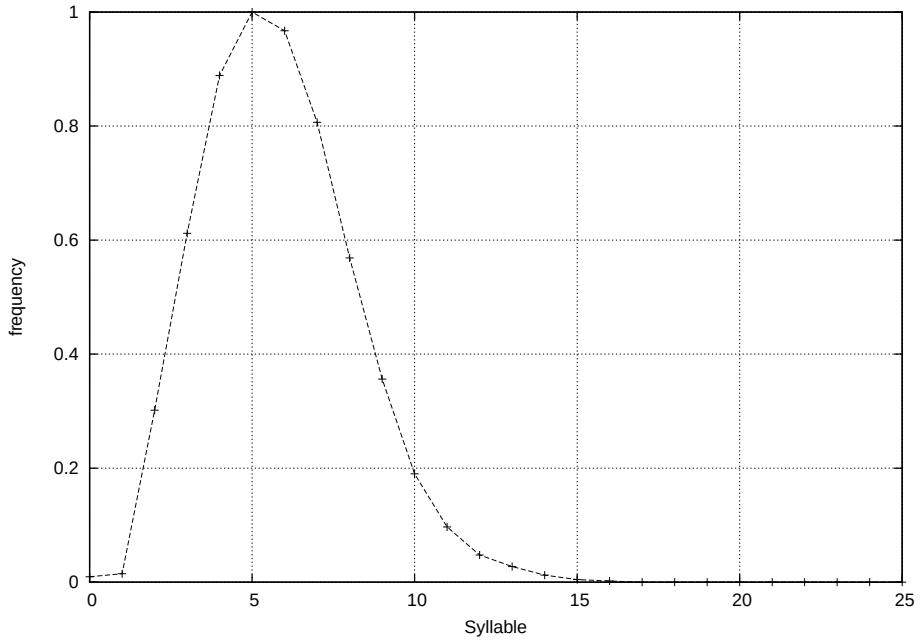


Fig 4.2: 推定略語のシラブル分布

横軸がモーラ数 (シラブル数)，縦軸がそのモーラ数 (シラブル数) の略語の出現頻度 (正規化してある) である．これらの図から明らかなように，CRF による略語推定ではモーラ数及びシラブル数が長い略語を推定していることが分かる．

これは以下のような理由により説明することができる．

Fig 4.3 では「住民基本台帳ネットワークシステム」に対して正解略語である「住基ネット」を推定しており，このとき確信度は素性の重要度を全て足したものと見なして良いので

$$
\begin{aligned}
C_r(\boldsymbol{x}, \boldsymbol{y}) &= \underbrace{(0.6 + 0.3 + 0.4 + 0.2)}_{\text{モーラの重要度}} \\
&\quad + \underbrace{(0.6 + 0.2 + 0.3 + 0.1)}_{\text{シラブルの重要度}} \\
&\quad + \underbrace{(0.4 + 0.5 + 0.6 + 0.9 + 0.8)}_{\text{観測素性の重要度}} \\
&= 5.9
\end{aligned}
$$

である． しかし，Fig 4.4 のように「住基ネットワーク」のような不正解略語を推定したと

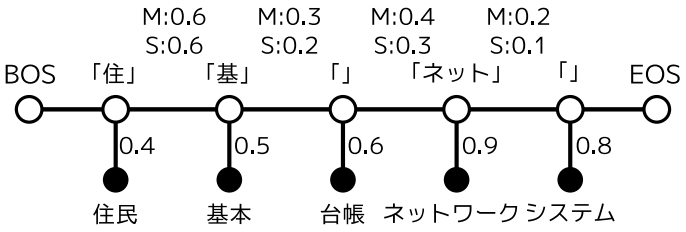


Fig 4.3: 正解略語を推定した時

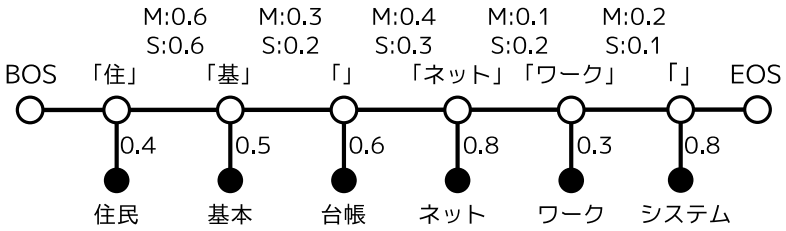


Fig 4.4: 不正解略語を推定した時

する．この時確信度は

$$\begin{aligned} C_r(\boldsymbol{x}, \boldsymbol{y}) &= (0.6 + 0.3 + 0.4 + 0.1 + 0.2) \\ &\quad +(0.6 + 0.2 + 0.3 + 0.2 + 0.1) \\ &\quad +(0.4 + 0.5 + 0.6 + 0.8 + 0.3 + 0.8) \\ &= 6.4 \end{aligned}$$

となる．

Fig 4.4 では不正解の略語を推定しているにも関わらず，推定略語の要素数が多くなり結果として確信度が高くなってしまう．このように CRF のみによる略語推定では要素数が多く長い (つまりモーラ数 · シラブル数の多い) 略語ほど確信度が高くなってしまい，必要以上に長い略語を推定することになる．これは CRF を用いて推定する以上避けられない現象である．

しかし，[12] によれば略語のモーラ数は 4 モーラ (43.9%) ないし 3 モーラ (25.5%) のものが多いと報告されている．そこで，略語推定を行ったのち，略語全体の長さを考慮して略語の絞り込みを行う．

略語のモーラ数に関する係数として

$$C_M(n) = \frac{(\text{学習用データのうち } n \text{ モーラの略語の数}) + 1}{(\text{学習用データの略語数}) + 1} \tag{4.1}$$

を定義する．つまりこれは

$$\frac{\text{学習用データのうち } n \text{ モーラの略語の数}}{\text{学習用データの略語数}}$$

という学習用データの略語に $n$ モーラの略語がいくつ含まれているか，という $n$ モーラの略語の割合をスムージングしたものである．同様にシラブルに対しても

$$C_S(n) = \frac{(\text{学習用データのうち } n \text{ シラブルの略語の数}) + 1}{(\text{学習用データの略語数}) + 1} \tag{4.2}$$

を定義する．さらに，

$$\mu = \mu(\boldsymbol{y}) \quad : \quad \text{略語 } \boldsymbol{y} \text{ のモーラ数} \tag{4.3}$$

$$\sigma = \sigma(\boldsymbol{y}) \quad : \quad \text{略語 } \boldsymbol{y} \text{ のシラブル数} \tag{4.4}$$

として $\mu$ および $\sigma$ を定義する．$C_M(n)$ と $C_S(n)$ を式 (3.3) の確信度に乗ずることによって式 (4.5) のように再定義し，略語の推定順位をリランキングする．

$$C'_r(\boldsymbol{x}, \boldsymbol{y}) = C_M(\mu) \cdot C_S(\sigma) \cdot C_r(\boldsymbol{x}, \boldsymbol{y}) \tag{4.5}$$

本研究では学習用データから $C_M$ および $C_S$ について Table 4.1 のような値が得られた．Table 4.1 によるとモーラは 4, 3, 2, 5 モーラの順に多く出現しており，Table 2.2 と結果が

Table 4.1: モーラおよびシラブル係数 $C_M, C_S$

| $n$ | $C_M(n)$ | $C_S(n)$ |
|---|---|---|
| 1 | 0.0400 | 0.0771 |
| 2 | 0.0628 | 0.4714 |
| 3 | 0.1542 | 0.3571 |
| 4 | 0.5457 | 0.0714 |
| 5 | 0.0542 | 0.0257 |
| 6 | 0.1142 | 0.0114 |
| 7 | 0.0200 | 0.0028 |
| 8 | 0.0200 | 0.0028 |
| 9 | 0.0085 | 0.0028 |

一致した．また，シラブルについては 2 シラブル，3 シラブルの順に多く出現していることが新たに判明した．

### 4.1.2　実験結果

実験結果は Fig 4.5 ∼ 4.7 のようになった．これより，略語長さを用いた絞り込みを行うことにより再現率，適合度ならびに $f$ 値が向上していることが分かる．Fig 4.6 から，ランクが 5 以上の時本研究においては適合度が先行研究を下回っていることが分かる．これは，先行研究が原語に対する正解略語は複数存在しているのに対し，本研究では正解略語を 1 つのみに限定しているためであると考えられる．この影響を受け Fig 4.7 のように $f$ 値も 5 位以上で先行研究よりも低い値となっている．

この絞り込み手法では，$C_M$, $C_S$ により生成する略語の長さは考慮しているが，原語の長さについては一切触れていない．Table 2.3 によれば原語のモーラ数が多くなるほど略語のモーラ数も多くなるという傾向が見られる．今後の課題として，原語の長さを考慮した絞り込みを行うことが考えられる．

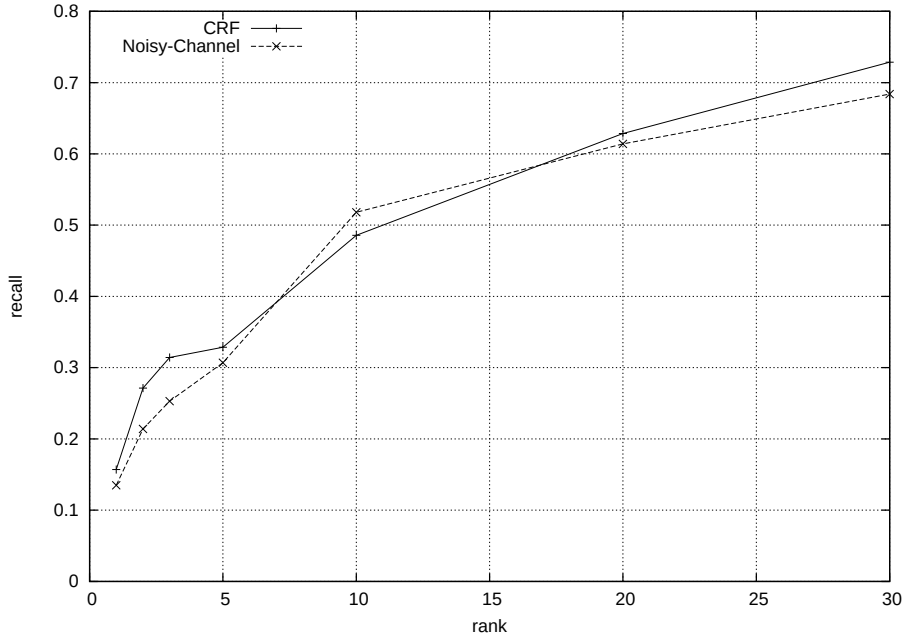


Fig 4.5: 略語長による絞り込み (再現率)

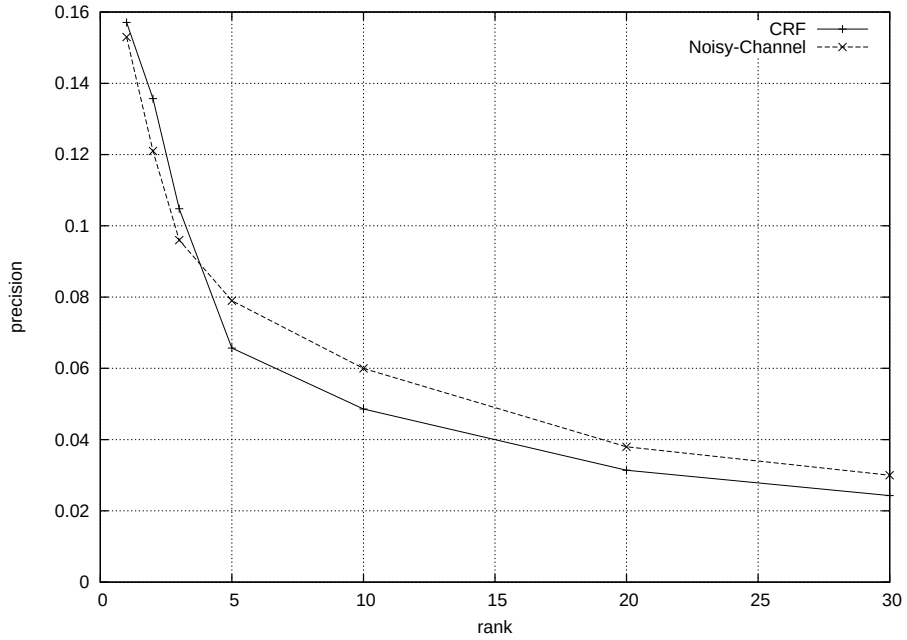


Fig 4.6: 略語長による絞り込み (適合度)

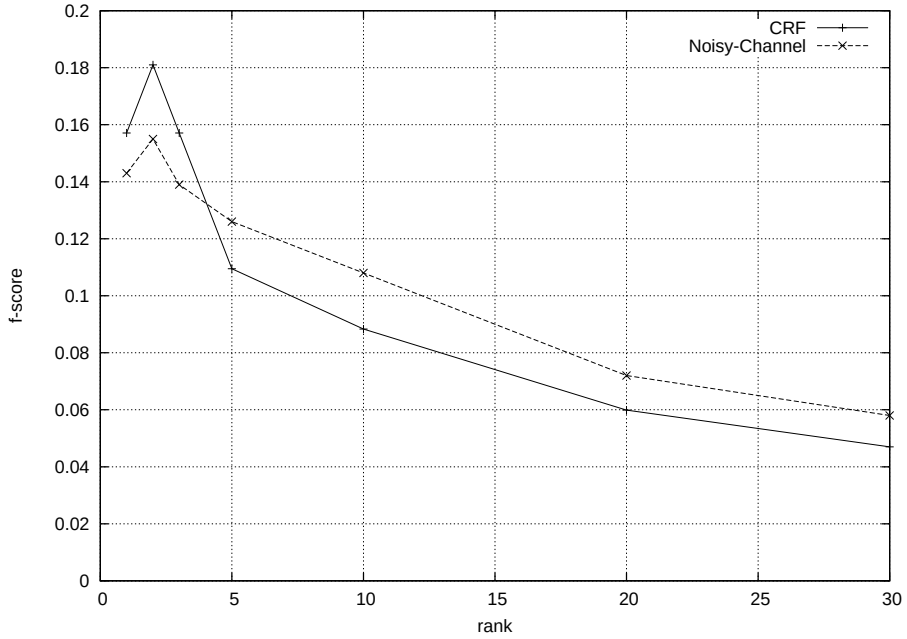


Fig 4.7: 略語長による絞り込み ($f$ 値)

## 4.2　検索エンジンを用いた絞り込み

本節では CRF により推定した略語が尤もらしい略語かどうか，検索エンジンを用いて絞り込みを行うことにより略語推定の精度を高める．

### 4.2.1　括弧表現による絞り込み

新聞記事などでは「東京大学 (東大)」もしくは「東大 (東京大学)」のような括弧表現を用いて略語の表記を行うことがしばし見受けられる．[7] によれば新聞記事で見つかる括弧表現のうち 10%以上は略語であるという．そこで，検索クエリとして「[原語]([推定した略語])」および「[推定した略語]([原語])」の 2 種類の括弧表現を用い，ヒット数の多い [原語] と [推定した略語] のペアを尤もらしい略語として採用する．

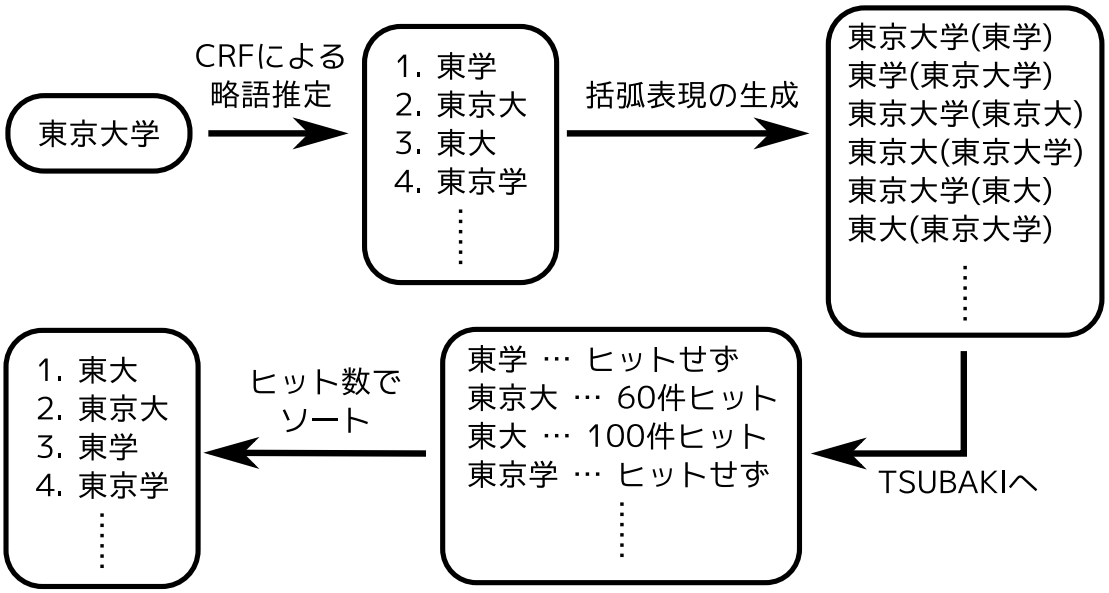


Fig 4.8: 括弧表現を用いた絞り込み

Fig 4.8 に具体例を示す．例えば「東京大学」に対して「東学」「東京大」「東大」「東京学」が略語候補として CRF により推定された場合,「東学 (東京大学)」「東京大学 (東学)」「東京大 (東京大学)」「東京大学 (東京大)」· · · を検索クエリとして用いる．

この手法では検索対象が括弧表現を含む文書に制限されてしまうという欠点はあるが，従来の括弧表現から略語を抽出する研究と比較すると

- 辞書 (品詞のデータ) が不要
- 文字種に依存しない
- 言い換え発生率を用いていない

というメリットがある．

### 4.2.2　実験結果

検索エンジンには TSUBAKI [18] を用いた．TSUBAKI は自然文で検索を行うことができる検索エンジンであり，科研特定領域研究「情報爆発」にて開発 · 運用が進められている．API が提供されており，ユーザ登録無し，API 制限無しに用いることができる．検索対象となる文書は 2007 年 5 月から 7 月にかけて収集した日本語の 1 億ウェブページである．TSUBAKI は最新のウェブページの内容をクローリングしているわけではないので最新のキーワードを検索することはできないが，特定の検索クエリに対して常に同じ検索結果を返すことになるため本実験の評価に適していると考え，今回は TSUBAKI を用いることにした．

括弧表現による絞り込みを行った結果は Fig 4.9 ∼ 4.11 のようになった．これより，括弧表現による絞り込みを行うことによって CRF のみを用いた推定よりも大幅に性能を向上させることができた事が分かる．

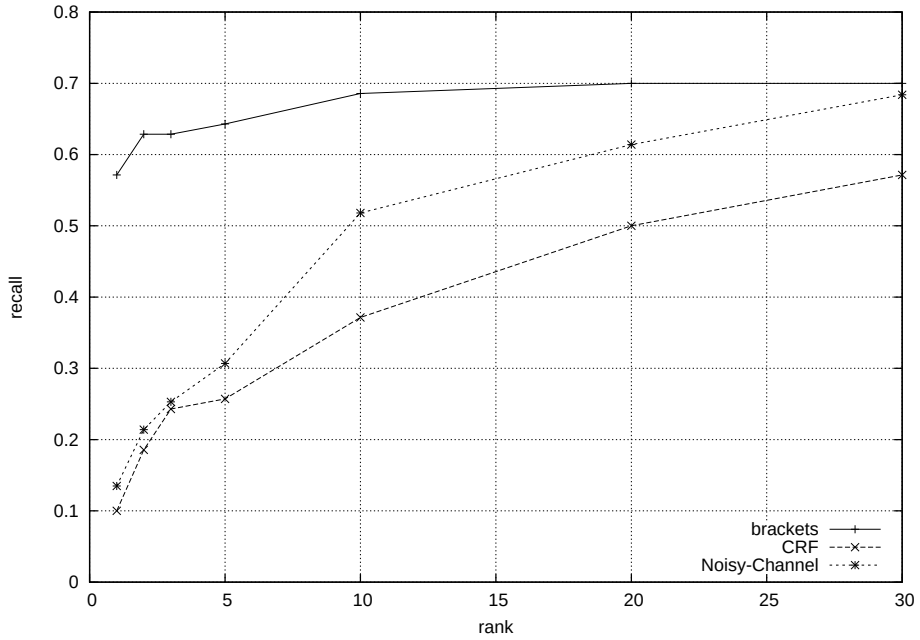


Fig 4.9: 括弧表現による絞り込み (再現率)

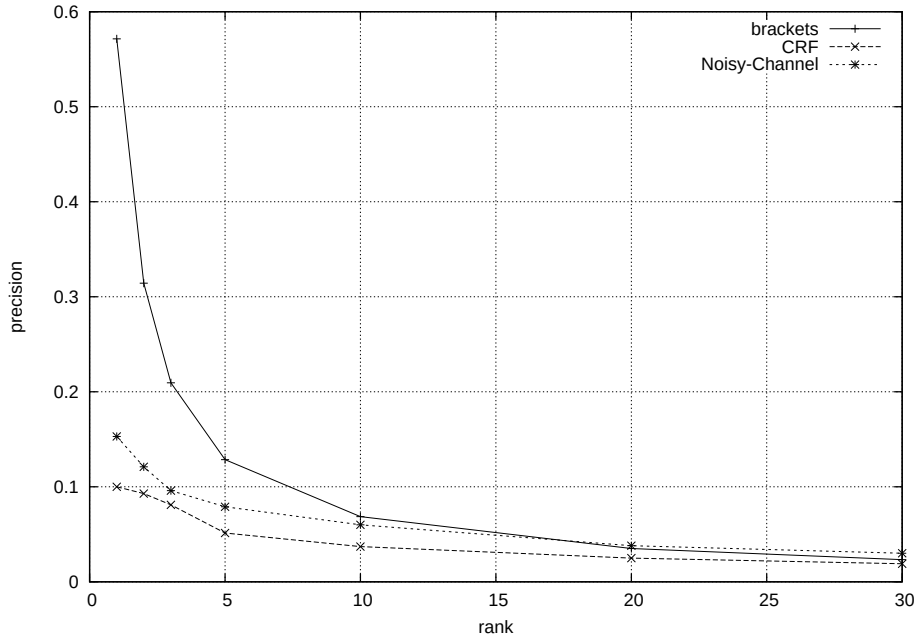


Fig 4.10: 括弧表現による絞り込み (適合度)

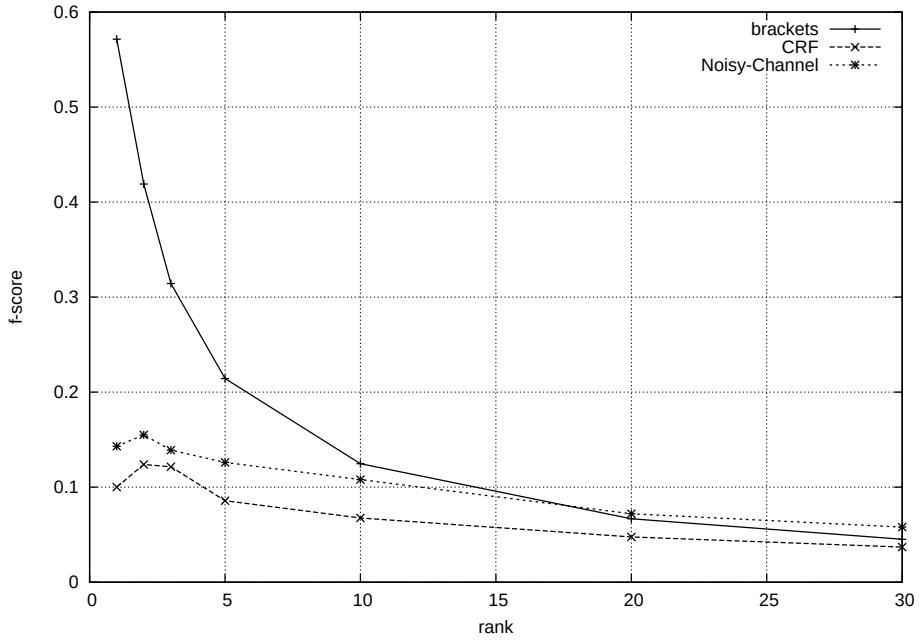


Fig 4.11: 括弧表現による絞り込み ($f$ 値)

なお，TSUBAKI を用いた括弧表現による絞り込みを行った際,「CRF で正しい略語を $n$ 位に推定したのにも関わらず，括弧表現による絞り込みで正解略語を $(n+1)$ 位以下に落としてしまう確率」は 4.34%であった．これはつまり，CRF で推定した略語のうち，明らかに略語として相応しくないものを括弧表現によってほぼ取り除くことが出来ることを意味している．従って，今後 CRF による略語推定の性能を向上させようとした場合は再現率よりも適合度を向上させる点に重点を置いた方が良い (つまり，略語推定の際に略語の取りこぼしが無いように) と考えることができる．

なお，括弧表現による誤判定の例には「世界銀行」↔「世銀」が例として挙げられる.「世界銀行」に対しては「世銀」を一位に推定するべきであるが，CRF で「世銀」と「銀行」の 2 つを略語候補として推定した場合，括弧表現による検索を行うとヒット数は以下のようになる．

- 世界銀行 - 世銀の括弧表現 · · · 2 件ヒット
- 世界銀行 - 銀行の括弧表現 · · · 8 件ヒット

従ってヒット数の多い「銀行」が「世界銀行」に対する尤もらしい略語として一位に推定され，正しい略語である「世銀」が二位に推定されてしまう．これは括弧表現による絞り込みを行った場合の失敗例である．

原因は括弧表現として

- 世界銀行 (世銀)
- 世銀 (世界銀行)

という文だけでなく

- 国際復興開発 銀行 (世界銀行) [1]

という表現が検索対象の文書に含まれているからである.

つまり，推定した略語が原語の接尾となっているとき,「[略語]([原語])」の括弧表現で検索したときに誤判定が出やすい.

では，誤判定の原因となる「[略語]([原語])」の括弧表現は検索クエリとして用いない方が良いのか．2 種類の括弧表現が表われる確率を調べたところ

- [原語]([略語]) のパターン ··· 73.3%
- [略語]([原語]) のパターン ··· 26.7%

であった．つまり,「[略語]([原語])」のパターンは全体の 1/4 以上を占めている．もし検索対象となる文書が充分に大きければ，このような誤判定を防ぐために「[原語]([略語])」の括弧表現パターンだけを検索クエリとして用いれば良いだろう．しかし，検索対象となる文書が充分な大きさを持っていない，もしくは原語や略語が比較的新しい語で検索対象となる文書の中に僅かしか含まれていない場合,「[原語]([略語])」の括弧表現パターンのみを検索クエリとして用いるのは 1/4 以上を占める「[略語]([原語])」の括弧表現をみすみす捨ててしまうことになる．この問題については今後なんらかの手法により改善する必要があると言える．具体的には「[略語]]([原語])」パターンの検索を行うときは [略語] の前の文字が平仮名 (すなわち助詞である可能性が高い文字) であることを条件として付与した検索方法が考えられる.

### 4.2.3　共起頻度による絞り込み

前節では括弧表現による絞り込みを行ったが，括弧表現というのはある程度限定された文書群でしか発見することができず，かつ，比較的新しい語については括弧表現で抽出するこ

[1] 世界銀行グループを形成する機関として
- 国際復興開発銀行
- 国際開発協会
- 国際金融公社
- 多国間投資保証基幹
- 国際投資紛争解決センター

の 5 つが存在するためにこれはこれで面白い検索結果であると言うこともできる.

とができない．そこで，本節では括弧表現に頼らず共起尺度による絞り込みを行い略語推定の精度を高める．絞り込みを行う際には略語と原語の共起尺度として Jaccard 係数ならびに Simpson 係数を用いる．ある語 $X$, $Y$ それぞれの検索ヒット数を $|X|$, $|Y|$，$X$ と $Y$ の AND の検索ヒット数を $|X \cap Y|$，$X$ と $Y$ の OR の検索ヒット数を $|X \cup Y|$ とすると，Jaccard 係数ならびに Simpson 係数は式 (4.6), (4.7) のように表わされる [19].

$$Jaccard = \frac{|X \cap Y|}{|X \cup Y|} \tag{4.6}$$

$$Simpson = \frac{|X \cap Y|}{\min(|X|, |Y|)} \tag{4.7}$$

Jaccard 係数と Simpson 係数は，分子は同じであるが分母が異なる．例えば $|X| = 1000$, $|Y| = 30$, $|X \cap Y| = 30$ という極端な場合を考えてみる．このとき Jaccard 係数は

$$\begin{aligned} Jaccard &= \frac{|OP \cap Abb|}{|OP \cup Abb|} \\ &= 0.03 \end{aligned}$$

となり，$Y$ からすると全てのページで $X$ と共起しているにも関わらず Jaccard 係数は小さくなってしまう．つまり，単独でのヒット件数が大きい $X$(もしくは $Y$) ほど他のキーワードとの関係が薄くなってしまうという欠点がある．一方，Simpson 係数は分母にて $|X|$ と $|Y|$ の min を取っているのでこのような問題は起きにくい．

さて，ここで原語を $OP$，略語を $Abb$ とすると，Jaccard 係数ならびに Simpson 係数は以下のように書き変えることができる．

$$Jaccard = \frac{|OP \cap Abb|}{|OP \cup Abb|} \tag{4.8}$$

$$Simpson = \frac{|OP \cap Abb|}{\min(|OP|, |Abb|)} \tag{4.9}$$

このようにして計算された Jaccard 係数および Simpson 係数を用い，Fig 4.12 に示すような手順で略語の絞り込みを行う．

### 4.2.4　実験結果

共起尺度による絞り込みでは検索エンジンとして Yahoo! Japan Web API 検索 [20] を用いた．Yahoo! Japan Web API 検索は TSUBAKI よりも結果を返すのが 10 倍以上速く，かつ検索対象となる文書群が多いため，今回は Yahoo! Japan Web API 検索[2]を用いた．しか

[2]なお，Google Web APIs の制限が 5000 query/day であるのに対し Yahoo! Japan Web API では 50000 query/day の制限である．

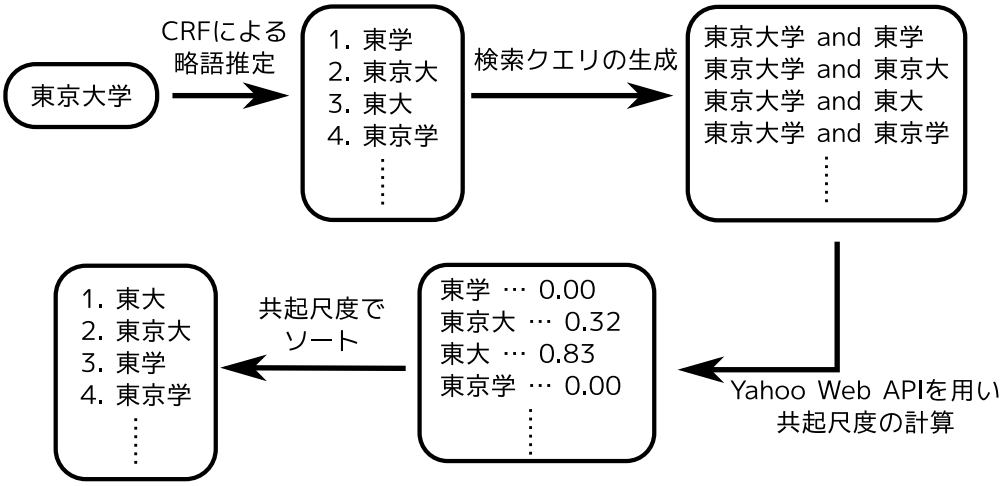


Fig 4.12: 共起尺度による絞り込み

し一方で，括弧表現は stop word として処理されてしまうため，4.2.1 で提案したような括弧表現での検索を行うことはできない.

この場合の実験結果は Fig 4.13 ∼ 4.15 となった．結果として，CRF そのものよりも共起尺度による絞り込みをすることにより格段に $f$ 値が向上し，Simpson 係数よりも Jaccard 係数を用いた方が良い結果を得られることとなった.

[19] では Jaccard 係数，Simpson 係数だけでなく式 (4.10) に示す閾値付 Simpson 係数を用いた場合についての考察を行っており，Jaccard 係数, Simpson 係数を用いた場合よりも閾値付 Simpson 係数を用いた場合の方が良い結果を得られたと報告している.

$$\mathit{Threshold\ \ Simpson} = \begin{cases} \frac{|X \cap Y|}{\min(|X|,|Y|)} & \text{if } \min(|X|,|Y|) > k \\ 0 & \text{otherwise} \end{cases} \tag{4.10}$$

従って，閾値付 Simpson 係数を用いた場合の絞り込みを行うことを今後検討したい.

CRF のみ，先行研究 (Noisy-Channel)，略語長による絞り込み (Length)，括弧表現による絞り込み，Simpson 係数による絞り込み，Jaccard 係数による絞り込みの全結果 ($f$ 値) を Fig 4.16 に示す.

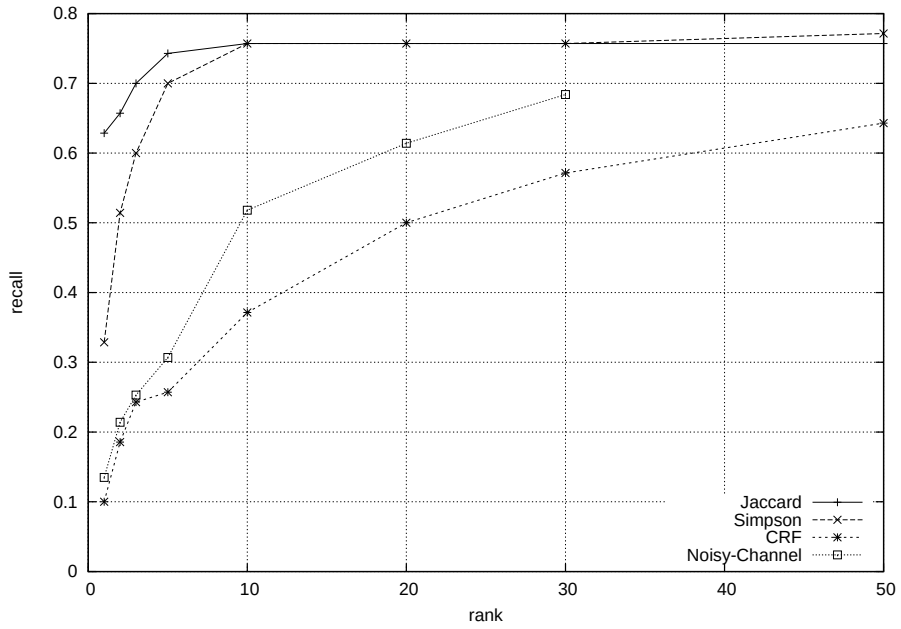


Fig 4.13: 共起頻度による絞り込み (再現率)

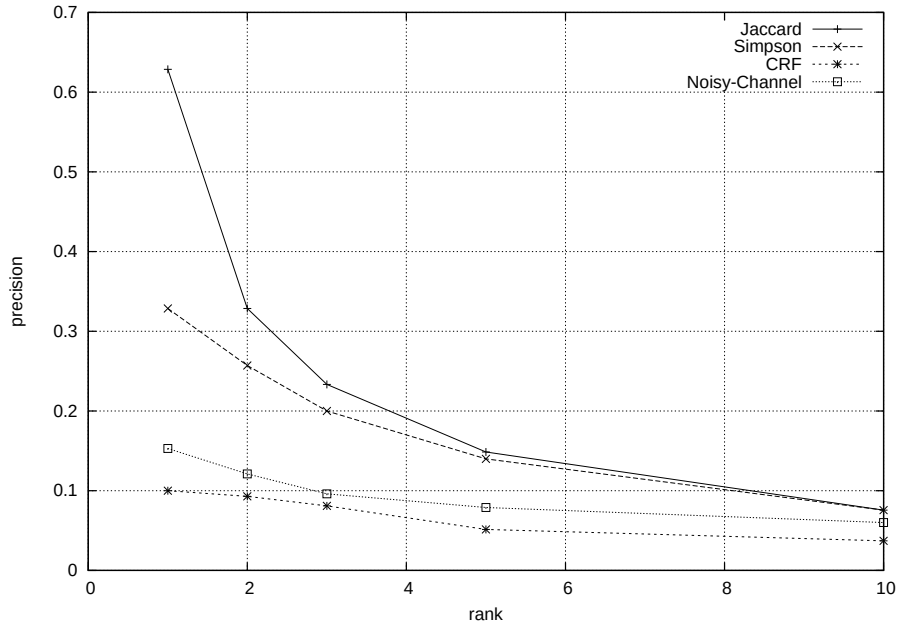


Fig 4.14: 共起頻度による絞り込み (適合度)

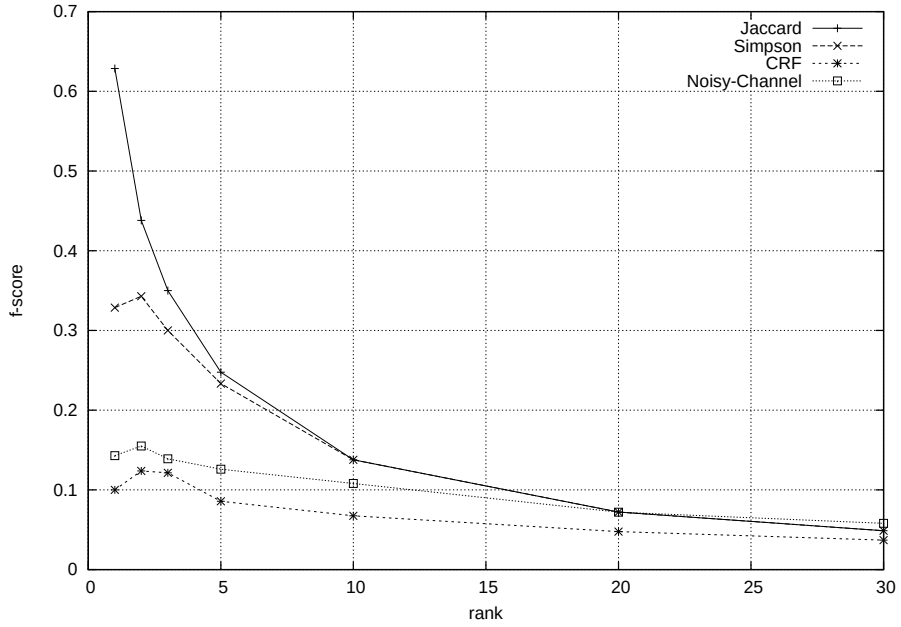


Fig 4.15: 共起頻度による絞り込み ($f$ 値)

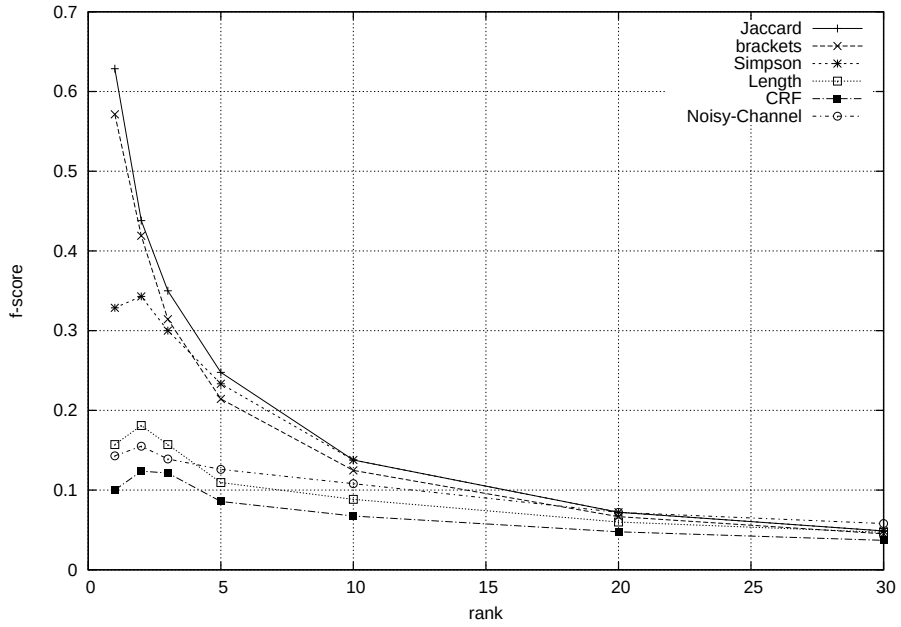


Fig 4.16: CRF および絞り込みを行った場合の $f$ 値

# 第5章　結論

## 5.1　まとめ

本研究では，CRF の素性にモーラとシラブルを用い，人間の感覚を取り込んだ略語推定の手法を提案した．既知の略語を学習用データとして用い，原語に対して未知の略語をラベリングする問題として解くことにより,「大阪大学」に対して「阪大」や「アメリカ」に対して「米」というような，表層上の文字に強く依存する従来手法では獲得できなかった略語を推定することが可能となった．また，そのようなメリットを持つ推定手法であるにも関わらず，従来の推定手法と比較してそこまで推定精度は劣っていない．また，括弧表現から略語を抽出する手法と比較すると，本提案手法での推定方法では括弧表現が存在しない略語を推定することができ，推定対象となる略語が多い．

さらに，本研究では推定した略語の絞り込みを行う手法を提案した．略語長 · 括弧表現 · 共起尺度による絞り込みの 3 つを試したところ，共起尺度のうち Jaccard 係数を用いた絞り込みを行うのが現時点でのベストであることが分かった．従来の括弧表現からの抽出では括弧表現に無い略語は抽出することができず，しかも括弧表現というのはある程度堅い文書 (主に新聞記事) に限定される表現である．従って，抽出できる略語に限りがあり，しかも抽出対象とできる文書群がかなり限定されてしまう．一方，共起尺度による絞り込みを行った場合，括弧表現に依存しないので表現が多種多様に及ぶブログのような Web 上の文書群も検索の対象とすることができ，比較的新しい略語に対しても有効な絞り込みを行うことが可能である．また，先行研究では原語と略語の先頭文字が一致しているか否か，といった表層上の文字を頼りにして略語抽出を行うことが一般的であったが，本研究ではそのような表層上の文字に依存しない絞り込みを行うことができる．

## 5.2　今後の課題

今後の課題としては以下のようなものが挙げられる．

1. CRF による推定の $f$ 値向上
   括弧表現や共起尺度を用いることにより，略語候補から正解略語をかなりの精度で絞り込むことができる．従って，4.2.2 にも述べたように主に適合度を向上させるように

推定を行うことが必要である．現在考えているのは CRF の観測素性に「原語要素 $x_i$ の何文字目を略語要素 $y_i$ として用いたか」という素性である．[12] によれば

> 略語のパターンは後略 (43.0%) が最も多く，前略 (7.58%) やその他 (3.62%) は少ない．後略が多いのは前半部を残すので復元可能性が高いためである．

と述べられている．そこで，原語要素 $x_i$ のより前方部分 (例えば「東京」であれば「東」) を略語要素 $y_i$ として採用することにより略語推定の際に正解でない略語の推定順位を下げることが可能であると考えられる．

2. 検索クエリとして使う括弧表現
   括弧表現を検索クエリとして絞り込みを行う場合,「[略語]([原語])」の括弧表現を用いた場合は検索対象が広がる反面，ノイズも多くなってしまう．この括弧表現を検索クエリとして用いた方が良いのか，それとも用いない方が良いのか検証を行う必要がある．

3. 共起尺度
   今回は Jaccard 係数および Simpson 係数を共起尺度として用いたが，閾値付 Simpson 係数，さらにコサイン距離

$$Cosine\ Distance = \frac{|X \cap Y|}{\sqrt{|X| \cdot |Y|}}$$

   を用いて絞り込みを行った場合どのような $f$ 値が得られるか検証したい．その上で略語推定に用いる共起尺度はどの指標が最も適しているのか比較を行いたい．

# 参考文献


[1] 窪薗晴夫, 「新語はこうして作られる」, 岩波書店, 2002

[2] Hiroyuki SAKAI, Shigeru MASUYAMA, “Knowledge Acquistion of Relation between Abbreviations and Their Original Words”, Institute of Electronics, Information and Communication Engineers. Vol.J85-D-II No.10(2002) pp.1624-1628

[3] Hiroyuki SAKAI, Shigeru MASUYAMA, “Improvement of the Method for Acquiring Knowledge from a Single Corpus on Correspondences between Abbreviations and Their Original words”, Journal of natural language processing Vol.12 No.5 pp.207-231, 2005

[4] 岡崎直観, 辻井潤一, 「Conditional Random Fields を用いた略語抽出」, NLP 若手の会 第 2 回シンポジウム, 2007

[5] Yuichiro SEKIGUCHI, Yoshihide SATOU, Harumi KAWASHIMA, Hidenori OKUDA, “Clipped word extraction using blog documents”, IPSJ 2007-DBS-143

[6] 岡崎直観, 石塚満, 「言い換え可能な括弧表現の抽出法」, 言語処理学会第 13 回年次大会 (NLP2007) pp.911-914, 2007

[7] Naoaki Okazaki, Mitsuru Ishizuka, “Abbreviation Recognition in Japanese Newspaper Articles”, The 21st Annual Conference of the Japanese Society for Artificial Intelligence, 2007

[8] 村山紀文, 奥村学, 「Noisy-channel model を用いた略語自動推定」, 言語処理学会第 12 回年次大会発表論文集 pp.763-766, 2006

[9] MIYAZAWA Kouki, HONDA Akiko, KIKUCHI Hideaki, “Automatic Estimation of Abbreviations using Phonological Generation Rules”, SIG-SLUD-A801-01, 2009

[10] Norifumi MURAYAMA, Manabu OKUMURA, “Abbreviation Estimation using a Probablistic Model with Web Information”, IPSJ 2008-NL-183

[11] 窪薗晴夫, 「音声学 · 音韻論」くろしお出版,2007

[12] 鈴木俊二,「外来語の略語の構造 –音節 · モーラ · フット · 語」, 国際短期大学 [編] Vol.11 pp.21-44, 1996

[13] 桑本裕二,「日本語におけるモーラ的鼻音の特徴」, 東北大学言語学論集 第 11 号, pp.93-104, 2002

[14] 坪井祐太, 鹿島久嗣, 工藤拓,「言語処理における識別モデルの発展 -HMM から CRF まで-」, 言語処理学会第 12 回年次大会 (NLP2006) チュートリアル, 2006

[15] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto, "Applying Conditional Random Fields to Japanese Morphologiaical Analysis", IPSJ SIG Notes Vol.2004, No.47 pp.89-96

[16] MeCab: Yet Another Part-of-Speech and Morphological Analyzer http://mecab.sourceforge.net/

[17] K's Bookshelf 辞典, http://ksbookshelf.com/DW/Ryaku/index.html

[18] 情報爆発プロジェクト検索エンジン基盤, http://tsubaki.ixnlp.nii.ac.jp/

[19] Yutaka Matsui, Tomobe Hironori, Hasida Koiti, Nakashima Hideyuki, Mitsuru Ishizuka, "Social Network Extraction from the Web information", The Japanese Society for Artificial Intelligence, Vol.20, No.1 pp.46-56

[20] Yahoo! Japan デベロッパーネットワーク, http://developer.yahoo.co.jp/webapi/search/websearch/v1/websearch.html

[21] Naoki ISHII, Tomonobu HIRAISHI, Shiho NOBESAWA, Hiroaki SAITO, Masakazu NAKANISHI, "Restoring Japanese Abbreviations", IPSJ SIG Notes Vol.2000 No.53 pp.61-68, 2000

[22] Terada Akira, Tokunaga Takenobu, "Automatic disabbreviation by using context information", IPSJ SIG Notes Vol.2001 No.69 pp.39-45, 2001

[23] ARITA Ippei, KIKUCHI Hideaki, SHIRAI Katsuhiko, "Word Clustering Using Concurrent Search Queries", IPSJ 2007-NL-180

[24] 新納浩幸, 佐々木稔,「検索エンジンを利用した未登録単語に関する単語間距離の測定」言語処理学会第 12 回年次大会, 2006

# 発表文献

1. 和田健太, 三輪誠, 横山大作, 近山隆, 「素性にモーラとシラブルを用いた略語の自動推定」, 情報処理学会 第 190 回自然言語処理研究会, 2009

# 謝辞

本研究を進めるにあたり，数多くの御指導と御助言を頂きました，近山 隆教授，田浦 健次朗准教授に心から感謝いたします．御二方の御指導がなければこうして修士論文を書き上げることができませんでした．また，勉強会などでたくさんのアドバイスを頂きました横山大作さん，三輪 誠さんに感謝いたします．研究の方針や詳細を詰めていく時にとても心強いアドバイスとなりました．

他，近山·田浦研究室の皆様に大変御世話になりました．勉強会にて何度も何度も意見をぶつけ議論した事は私にとって非常に素晴しい糧となりました．誠にありがとうございました．