

形態素周辺確率を用いた 確率的単語分割コーパスの 構築とその応用

岡野原 大輔 東京大学
 工藤 拓 Google Japan
 森 信介 日本 IBM 東京基礎研究所

背景 単語分割の難しさ

- 入力テキストを処理単位に分割する
 横浜市役所 横浜/市/役所 横浜市/役所
 横浜市役所
 応援団長 応援/団長 応援/団/長
 応援団/長
 東京大学情報理工学系研究科
 東京/大学/情報/理/工/学/系/研究/科
 どの分割が最適か一意に決定できない

形態素解析の結果を曖昧な形で保持したい

背景 全文検索

- 全文検索 (特に文字索引) の普及
 - 任意の部分列を漏れなく検索 **トロ** で検索

音韻体系や語彙はむしろ、南方系のオーストリア語族との近縁が見られるが、どうい
 う02年5月にキューバを訪れフィデル・カストロに会った。1959年の革命以来初めてキ
 ソル財閥のヴェステルなど、家電・エレクニクス部門の成長も期待されている。た
 点からは、液体や気体、熱エネルギー、エンピー、波といった巨視的な物理現象が研究さ
 規制も危惧されている。ネットを介してイの木馬やコンピュータウイルスの感染が広
 キターを担当していた)。ジョン作の「ストベリー・フォールズ・フォーエバー」でのイ
 リー・フォールズ・フォーエバー」でのイントロのピアノはポールの演奏である。主にピア
 ルズ・フォールズ・フォーエバー」でのイントロのピアノはポールの演奏である。主にピア
 「ゲッティンガ・ベター」でのタブラ、「ストベリー・フォールズ・フォーエバー」でのソ
 の等身大以上に巨大化し、もはや自らもコントロールできなくなってしまった「ビートルズ」
 本体、プレイヤーが入力に用いる装置(コントローラ等)、処理結果が出力される装置(主に

トロ(魚の部位)とは無関係の単語が結果に含まれる

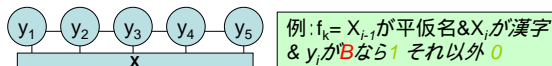
→ **検索結果に形態素解析結果を組み合わせたい**

提案手法

- SSC(確率的単語分割コーパス)を利用
 - 曖昧な分割情報を保持可能
- CRFを用いて各文字間の分割確率を求める
 - 文字列の文字種、Suffix、辞書情報、など
 リッチな情報を利用して分割情報を求める
- 全文検索結果と分割情報をオンラインで
 組み合わせる

CRF: Conditional Random Fields [Lafferty+ 01]

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_{i=1}^{|\mathbf{y}|} \sum_k \lambda_k f_k(\mathbf{x}, y_{i-1}, y_i, i)\right)$$



- P(Y|X) を単一の指数分布モデルで表現
 - 文字種、接尾辞等を利用した柔軟な素性設計が可能
 - 本タスクにおいては各 y_i は次の二つ
- B: 単語の開始を示す I: 途中を示す

今日 は と て も 良 い 天 気 で す
 B | B B | | B | B | B |

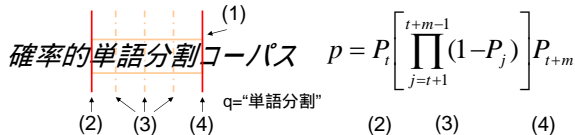
SSC: 確率的単語分割コーパス [Mori+ 04]

確率的単語分割コーパス
 0.99 0.01 0.99 0.89 0.18 0.85 0.19 0.95 0.0 0.0 0.0 0.99

- 文書 $x_{1..n}$ と連続する各2文字 x_i, x_{i+1} の間に単語
 境界が存在する確率 P_i からなる
 - 各 P_i は独立であり、作業領域量は $O(n)$ bit
 (全部分列の確率を保存するには $O(n^2)$ bits)
- 単語分割情報を factorize された形で保持

SSCにおける単語出現確率の計算

- SSCにおいて単語 $q[0\dots m-1]$ が $x[t\dots t+m-1]$ に出現している必要十分条件
- (1) 文字列が等しい $q[0\dots m-1]=x[t\dots t+m-1]$
 - (2) $x[t]$ の直前に単語境界がある
 - (3) 単語境界が $x[t\dots t+m-1]$ 中に無い
 - (4) $x[t+m-1]$ の直後に単語境界がある



全文検索+SSC

1. クエリと一致する全ての箇所を求める
 2. 1.で見つかった全候補について確率を求める
 3. 確率に基づいてソートする
- 2.の計算量は $O(\text{len} * \text{occ}) \#$
 - lenはクエリ長 occは一致箇所の個数
 - Succinct Data Structure [Jacobson 98]*を使うことで作業領域量を変えることなく計算量を $O(\text{occ})$ に減らすことが可能
 - * Bit列の部分和が定数時間で求まる
- #1. 3. の計算量はそれぞれ $O(\text{len}), O(\text{occ} * \log(\text{occ}))$

SSCの構築

| | 確 | 率 | 的 | 単 | 語 | 分 | 割 | コ | - | パ | ス |
|-----|------|------|------|------|------|------|------|------|------|------|------|
| (1) | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| (2) | 0.95 | 0.05 | 0.95 | 0.95 | 0.05 | 0.95 | 0.05 | 0.95 | 0.05 | 0.05 | 0.95 |
| (3) | 0.99 | 0.01 | 0.99 | 0.89 | 0.18 | 0.85 | 0.19 | 0.95 | 0.0 | 0.0 | 0.99 |

(1) 形態素解析結果 (2) 従来のSSC (=0.95) (3) 提案手法

- 従来[Mori+04,06]は P_j を同一パラメータを用いて決定していた
 - 形態素解析の結果単語境界と判定された位置を、それ以外を1-
 - は境界推定精度を用いる
- 文脈に依存した形で境界確率を推定できない
- CRFの結果を文字間分割確率に利用

CRFからのSSCの構築

$$P_i = \sum_{y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n} p(y_1, \dots, y_{i-1}, B, y_{i+1}, \dots, y_n | \mathbf{x})$$

$$= \frac{1}{Z(\mathbf{x})} \frac{\alpha_{i,B} \beta_{i,B}}{e_{i,B}}$$

$$\alpha_{i,y} = \sum_{y'} \left(\alpha_{i-1,y'} \cdot \exp\left(\sum_k \lambda_k f_k(y', y, x, i)\right) \right)$$

$$\beta_{i,y} = \sum_{y'} \left(\beta_{i+1,y'} \cdot \exp\left(\sum_k \lambda_k f_k(y, y', x, i)\right) \right)$$

- 各文字間の分割確率を $y_i=B$ 周辺確率で求める
 - $\alpha_{i,B}, \beta_{i,B}$ はforward-backward アルゴリズムで効率的に求められる

クエリに対する分割情報の場合

- クエリに対しても分割情報が与えられている場合にも効率的に求められる
 - クエリが 東京大学 と与えられた場合等

```
function calcProbQ (t, len, q[])
# q[]: クエリに対する確率的単語分割情報
  prevB = p[t]
  prevI = 0
  for i=1 to len-1
    tmp = prevB + prevI
    prevB = tmp * q[i] * p[t+i]
    prevI = tmp * (1-q[i]) * (1-p[t+i])
  end
  return (prevB + prevI) * p[t+len]
end
```

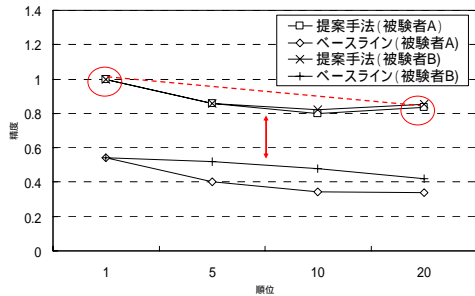
単語開始位置が一致した場合

単語中で一致した場合

実験環境

- Wikipediaの全文検索結果上位20件を被験者二人に提示し、クエリと関連しているかを判断してもらった
- 利用したクエリは次の通り
 - “京都”、“トロ”、“本部”、“国際”、“スター”、“パン”、“1月”、“かしい”、“はかた”、“国”
 - 1位、5位、10位、20位毎の結果を提示
- ベースラインは見つかった順に表示
- 提案手法はCRFを利用したSSCから求めた確率でソートした結果

実験



提案手法が一貫して、ベースラインよりも高精度高順位であるほど精度が高く、判別できている

実験結果例1/4(トロ) ベースライン

音韻体系や語彙はむしろ、南方系のオーストロネシア語族との近似が見られるが、どうい
02年5月にキューバを訪れフィデル・カストロに会った。1959年の革命以来初めてキ
ソル財閥のヴェステルなど、家電・エレクトロニクス部門の成長も期待されている。た
点からは、液体や気体、熱エネルギー、エンターテインメントの物理現象が研究さ
規制も危惧されている。ネットを介してインターネットの感染が広
ギターを担当していた)。ジョン作の「ストリープ・フィールズ・フォーエバー」でのイ
リー・フィールズ・フォーエバー」でのイントロのメロディはボールの演奏である。主に
「ゲッティング・ベター」でのメロディ、ペリー・フィールズ・フォーエバー」でのソ
の等身大以上に巨大化し、もはや自らもコントロールできなくなってしまった「ビートルズ」
本体、プレイヤーが入力に用いる装置(コントローラ等)、処理結果が出力される装置(主に

トロ(魚の部位)と関係の無い文脈で一致したとして報告される

実験結果例2/4(トロ) 提案手法

豚トロと呼ばれる部位が販売されている。トロという言葉は肉の種類に明確が
うとする販売戦略などから、今後も様々なトロは増えていくと思われる。フルネ
・中トロ以外の部分は「赤身」と称して、「トロ」とは別物とされる。一般に背肉よ
発売当初は供給量が少なかったところに、トロなどおしゃべりする「どこでもいっ
も、脂が乗っている肉の状態のことを総してトロということがある。例えば、カツオの
肉の種類に明確に定義がないこと、マグロのトロの高級品のイメージを借りようす
をトロカツオといったり、豚肉においても豚トロと呼ばれる部位が販売されている。
ていくと思われる。フルネームは「井上トロ」。また誕生日は5月6日とされ、当
劣るものを「中トロ」と称す。大トロ・中トロ以外の部分は「赤身」と称して、「
」、やや劣るものを「中トロ」と称す。大トロ・中トロ以外の部分は「赤身」と称
った部分を「大トロ」、やや劣るものを「中トロ」と称す。大トロ・中トロ以外の部分
含量が高い。「大トロ」は腹肉前部、「中トロ」は腹肉後部である。ただし、マグ
り表層肉のまわが脂質の含量が高い。「大トロ」は腹肉前部、「中トロ」は腹肉後
多い。特に、よく脂の乗った部分を「大トロ」、やや劣るものを「中トロ」と称す。

- トロを指し示す箇所が上位に来る
- 中トロ、大トロなど、形態素解析結果と完全マッチでは出なかった結果も上位に来る

実験結果例3/4(東大) ベースライン

しては倭王武が中国から俄国王安東大將軍に任せられたなどの記録がある。秀
の司令官として征夷大將軍、征夷大將軍(征夷將軍)、征西大將軍(征西將
、「大相撲」を参照。1923年には関東大震災によりシャープペンシル工場を焼失
画している旨の言表を行っている。「東大新聞」五月祭賞に入選した小説「奇妙な
な。更に昭和2年には、関東大震災の手形の魚づき付きが累積し、それ
;非政党内閣が続いた。その後、関東大震災や虎ノ門事件の発生は、それまで
;対策が後手後手に回った。更に関東大震災による京浜工業地帯の壊滅と緊急
なる。1923年(大正12年)には関東大震災が生じた。この未曾有の大災害に
;り退位を余儀なくされた。また、関東大震災で東京市内のほとんどの教会が破
学風を特色としている。この学風は東大 安田講堂事件に代表されるような学生
;漫画等でネタにされることもある(東大 一直線やドラゴン桜など) 法文学
別荘は後に小田原に移築され、関東大震災で焼失しているため現存しないが、
た。兄に冶金学者の小川芳樹(東大 教授、東洋史学者の貝塚茂樹(京大名

- 東大(大学名)と関係の無い文脈で一致したとして報告される

実験結果例4/4(東大) 提案手法

金沢医科大学の法医学教授となる。以後、東大、東京医科大学の教授を歴
究科博士課程修了。2001年に法学博士(東大)。所属学会は、日本行政学会、
月までドイツ滞在了。戦後農林省から東大、日大教授になる。多くの技術
東海圏の国立大学に多くの生徒が進学し、東大、京大合格者も三重県内では最
立し、その会長となった。1931年に東大では夏目漱石と同級。ちなみに
、また共通の友人として藤田主計がいる。東大、京大、阪大への進学者も多い
ルの高さ「3本の指に入る」と言われ、東大、京大、阪大への進学者も多い
992年東京大学経済学部経営学卒業。東大ではエースとして活躍も、六六が
なっている。南に外堀(御茶の水)、北に東大、東に湯島、西に後樂園。現在
東京帝国大学法学部政治学科を卒業する。東大では矢部貞治の門下で、同期
の類出事項・盲点事項を熟く語り続ける。東大では「天文学」と、京大では「宇
に生まれた。熊本の第五高等学校を経て、東大を卒業後、改造社で雑誌編集に
(東京帝国大学助教授。1928年4月に東大を追放され、以後は講壇ジャー-

東大(大学名)として正しいものだけが上位に来ている

まとめ

- 入力テキストの基本単位への分割情報を文字間分割情報(SSC)で扱う
 - 曖昧な単語分割情報を埋め込む
- 文字間分割確率は、CRFを用いた周辺確率として求められる
- 全文検索の際、任意部分列の出現確率がオンラインで求められる
 - クエリに分割情報が与えられている時も効率的に求められる

今後の目標

- より複雑な確率情報のfactorization
 - ・ 係り受け
 - ・ 構文木
 - ・ (階層のある)固有表現
- 速度、作業領域量の向上
 - ・ 確率の高い候補から提示可能な索引