コーパスからの関連語獲得に基づく連想を加味した顕現性の推定

白松 俊　　駒谷 和範　　尾形 哲也　　奥乃 博
京都大学大学院情報学研究科
{siramatu,komatani,ogata,okuno}@kuis.kyoto-u.ac.jp

1　はじめに

■ **背景**　文ごとに推移する単語の顕現性 (salience) の推定は，談話解析・文脈解析のために必要不可欠な要素技術である．顕現性とは，談話参与者 (発話者，受話者，筆者，読者) が当該単語に向けている共同注意の度合 (目立ち具合) である．顕現性の推定を必要とする談話解析・文脈解析の例を以下に列挙する．

- **会話文脈の追跡:** 人間同士の会話の流れを対話システムが理解し，会話に参与するためには，会話内容に含まれる単語群の顕現性の変遷を時系列的に追跡する必要がある．
- **照応解析:** 代名詞やゼロ代名詞の指示対象を推定する照応解析では，当該文脈において目立っている単語が指示対象候補となるので，各単語の顕現性の推定が必要である．
- **文章の滑らかさの評価:** 文と文の繋がりの滑らかさを評価するためには，注意焦点への参照の連続性，すなわち参照結束性 (referential coherence) の計測と，そのための顕現性推定が必要である．

■ **従来研究**　われわれは以前，顕現性の推定のために**参照確率**という尺度を設計した [5, 7]．これは，「目立っている単語ほど継続的に参照されやすい」という性質を利用した尺度である．先行文脈における当該単語の参照パターンからコーパスに基づいて統計的に計算される．われわれは，参照確率の設計によって他の顕現性推定手法の問題点 (表 1) の多くを解決した．

■ **課題**　参照確率による顕現性推定手法を更に改良するためには，関連語による連想効果への対応が課題となる．例えば「アメリカ村」が主題になっている文脈では，たとえ「大阪」という語が未出であっても，参与者達の意識には大阪という概念が想起される．つまり，ある語が注目されるとその周辺概念も連想され，関連語の顕現性が励起される．参照確率は当該単語単独の参照パターンだけから計算されていたので，関連語の影響を反映していなかった．そのため，文脈と深く関連するが背景知識として省略されている語などの顕現性は，不当に低く推定されてしまっていた．

■ **アプローチ**　本稿では，参照確率による顕現性推定手法の上記の問題を解決する．具体的には，同時に想起されやすい関連語，つまり同時に顕現性が高くなりやすい関連語をコーパスから獲得することで，参照確率に連想を加味した顕現性推定手法を開発する．

2　従来の顕現性推定手法の問題点

表 1 に示した従来の手法の問題点について述べる．

■ **語の頻度**　文書における単語の重要度としては，TF·IDF のような語の頻度に基づく尺度が用いられることが多い．しかし，そのような尺度は以下の理由から，顕現性の尺度には不向きである．

- **時系列的遷移への対応の困難さ:** 顕現性は文ごとに時系列的に遷移するので，その遷移に対応するためには，頻度の計測範囲を細粒度にする必要がある．しかし，細粒度ではどの単語の頻度も 1 回や 2 回になり，単語間の顕現性の差が検出できなくなる．
- **他の因子との統合方法が無い:** 時系列的遷移を扱うには，語の頻度以外の因子，たとえば文中での文法役割の影響なども反映する必要がある．

■ **センタリング理論**　センタリング理論 (centering theory) [1] は，顕現性と代名詞化に関するルールで参照結束性を説明する理論である．センタリング理論では，顕現性を Cf-ranking という文法役割の順序によって定義している．日本語においては，以下のように定義される [6].

主題 (助詞ハ) > 主語 (助詞ガ) > 間接目的語 (助詞ニ) > 目的語 (助詞ヲ) > その他

表 1　従来の顕現性推定手法の問題点

	語の頻度	センタリング	RAP	参照確率
時系列的遷移	×	○	○	○
定量化	○	×	○	○
複数因子統合	×	△	○	○
コーパス適応	○	×	×	○
関連語連想	○	×	×	×

この定義にも，以下の問題点がある．

- **他の因子との統合方法が無い:** 正確な顕現性推定のためには，文法役割だけでなく，語の頻度や，最近参照された位置等も考慮する必要がある[*1]．
- **非定量的:** 文法役割の順序だけでは，たとえば主語と目的語の顕現性にどのくらいの差があるのかわからない．
- **コーパス適応手法の不備:** 正確な顕現性推定のためには，言語データから学習できる枠組が必要であるが，そのような統計的手法は備えていない．

■ <u>RAP</u>　Resolution of Anaphora Procedure (RAP) [3] は，センタリング理論に似た照応解析のアプローチである．RAP では，幾つかの因子を統合した顕現性の重み付けを用いるが，この重み付けにも以下の問題点がある．

- **コーパス適応手法が無い:** データから統計的に獲得した重みではなく，先験的なルールベースの重み付け手法である．したがって，やはり各言語のコーパスへの適応が困難である．

■ **<u>参照確率</u>**　参照確率は，「注目されている対象ほど継続的に参照されやすい」という性質を利用した顕現性の尺度である．センタリング理論がゲーム理論から導出できることを示す過程で設計された [2, 5, 7]．参照確率 $\mathrm{Pr}(e|\mathrm{pre}(U_i))$ は，「発話単位 U_i までの先行文脈 $\mathrm{pre}(U_i) = [U_1, \cdots, U_i]$ における実体 e の参照パターンが与えられた下で，後続する U_{i+1} で e が参照される条件付確率」と定義され，コーパスを用いて統計的に算出される (図 1)．これが，当該発話 U_i における e の顕現性を表す．

参照確率は, 他の従来研究における知見 (Cf-ranking の順序や，「語の頻度が多いほど顕著」といった知見) と整合性のとれた尺度であることが，日本語・英語の大規模新聞記事コーパスにおいて確認されている [5]．さらに，以下の点で参照確率は他の手法よりも顕現性の推定に適している．

- **時系列的遷移に対応:** 発話単位 U_i ごとに算出できる設計になっており，時系列的遷移を扱いやすい．
- **各言語のコーパスに適応可能:** 図 1 に示す手法により，コーパスを用いて統計的に計算されるので，データに適応可能である．
- **複数の因子を統合する枠組:** 複数の素性 (表 2) から構成される参照パターンを統合して計算され，他の新たな素性を加えることも可能である．

われわれは，参照確率の設計によってこれらの点の解決を得たが，参照確率による顕現性推定手法にも 1 つ，以下の課題が残されている．

- **連想の影響を反映していない:** 関連語からの連想を加味するためには，数多くある関連語の参照パターンも考慮して顕現性を推定する必要がある．しかし，参照確率の算出方法 (図 1, 表 2) に対象単語以外の素性を追加する枠組は無い．そのため，例えば先行文脈で未出の単語については，いくら文脈と関連が深い概念であっても顕現性を推定できない．

次節では，この問題の解決策を述べる．

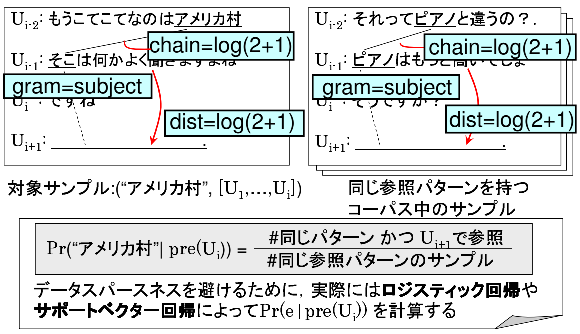

図 1　参照確率 $\mathrm{Pr}(e|\mathrm{pre}(U_i))$ の計算

[*1] Cf-ranking の拡張である SRL[4] は，文法役割だけでなく参照された位置も反映した顕現性推定手法である．しかし，他の新たな因子と統合する手法は備えていない．

3　連想を加味した顕現性推定手法

本研究では「同じ発話単位，同じ時点の注意状態に共起しやすい 2 つの単語は，互いの顕現性を上昇させる」と仮定する．ただし，注意状態とはその瞬間に注

表 2　$\mathrm{Pr}(e|\mathrm{pre}(U_i))$ の計算において e の参照パターンとして用いられる素性

dist	$\log((\#$ $[U_1, \cdots, U_i]$ 中で最近 e を参照した表現と U_{i+1} の間の発話単位の数$)+1)$
gram	$[U_1, \cdots, U_i]$ 中で最近 e を参照した表現の文法役割
chain	$\log((\#$ $[U_1, \cdots, U_i]$ 中の e への参照表現$)+1)$
exp	$[U_1, \cdots, U_i]$ 中で最近 e を参照した表現の名詞句種別 (代名詞/非代名詞)
last_topic	e が $[U_1, \cdots, U_i]$ 中での最近の主題か (yes/no)
last_sbj	e が $[U_1, \cdots, U_i]$ 中での最近の主語か (yes/no)
p1	e が一人称か (yes/no)
pos	$[U_1, \cdots, U_i]$ 中で最近 e を参照した表現の品詞
in_header	e がタイトルや見出し中で参照されているか (yes/no)

目されている単語の顕現性のリストのようなものと考える．具体的には，発話単位 U 時点における (連想を加味していない) 注意状態を以下の “**参照確率ベクトル**” $\boldsymbol{v}(U)$ で定義する．

$$\boldsymbol{v}(U) = [\Pr(n_1|\mathrm{pre}(U)), \cdots, \Pr(n_N|\mathrm{pre}(U))]^T \quad (1)$$

ここで $\boldsymbol{v}(U)$ は，コーパス中の全名詞 $n_1, \cdots, n_N$ を基底とする N 次元空間上の疎ベクトルであり，その要素は発話単位 U 時点における名詞 n_j の参照確率 $\Pr(n_j|\mathrm{pre}(U))$ である．

上記の仮定に基づき，(文書ごとの共起頻度ではなく) 発話単位 U ごとの注意状態 $\boldsymbol{v}(U)$ において単語同士が影響し合う傾向をコーパスから獲得し，それを用いて参照確率に連想を加味した顕現性推定を行う．

3.1 単語 n_j が注目されているときの平均的な注意状態 $\boldsymbol{b}_{n_j}$ の獲得

単語同士が顕現性を励起し合う傾向を獲得するために，ある名詞 n_j が注目されているときの**「平均的な注意状態 (参照確率ベクトル)」**$\boldsymbol{b}_{n_j}$ をコーパスから推定する[*2]．以下に，$\boldsymbol{b}_{n_j}$ 推定の手順を示す．

1. **名詞句抽出:** コーパスに含まれる全名詞句 $n_1, \cdots, n_N$ を抽出する．
2. $\boldsymbol{v}(U)$ **推定:** コーパスに含まれる全ての発話単位 U に対し，その瞬間の参照確率ベクトル $\boldsymbol{v}(U)$ を求める．
3. $\boldsymbol{b}_{n_j}$ **の計算:** 全名詞 n_j $(j = 1, \cdots, N)$ に関して以下 (3.a, 3.b, 3.c) を行う．

3.a. n_j **が顕著な注意状態の抽出:** n_j が注目されている発話単位 U の時点における参照確率ベクトル (注意状態) の集合として，$S_{n_j} = \{\boldsymbol{v}(U) \mid \frac{\Pr(n_j|\mathrm{pre}(U))}{\|\boldsymbol{v}(U)\|} \geq \theta\}$ をコーパスから抽出する．ただし，θ は所定の閾値である．

3.b. **総和:** n_j の参照確率による重み付き総和ベクトル $\boldsymbol{v}_{n_j} = \sum_{\boldsymbol{v}(U) \in S_{n_j}} \Pr(n_j|\mathrm{pre}(U))\boldsymbol{v}(U)$ を計算する．

3.c. **正規化:** $\boldsymbol{v}_{n_j}$ の長さを正規化した $\boldsymbol{b}_{n_j} = \frac{\boldsymbol{v}_{n_j}}{\|\boldsymbol{v}_{n_j}\|}$ を，名詞 n_j が注目されているときの平均的な注意状態と見なす．

コーパスから推定された $\boldsymbol{b}_{n_j}$ の k 番目の要素 $b_{j,k}$ は，**「n_j が注目されると同時に名詞 n_k がどれだけ注目されやすいか」**を表す．この $b_{j,k}$ は，名詞 n_j の顕現性が名詞 n_k の顕現性に影響する度合と見なせるはずである．また，$\boldsymbol{b}_{n_j}$ は，**「名詞 n_j の基底ベクトル** (名詞 n_j の軸だけ 1 で，他の軸が 0 であるベクトル) **を，n_j の関連語軸方向に回転させたもの」**と見なせる．

3.2 $\boldsymbol{b}_{n_j}$ を用いて参照確率に連想を加味

次に，コーパスから推定した $\boldsymbol{b}_{n_1}, \cdots, \boldsymbol{b}_{n_N}$ を用いて，名詞 n_k の参照確率 $\Pr(n_k|\mathrm{pre}(U))$ に連想を加味した顕現性推定を行う．$\boldsymbol{b}_{n_j}$ の k 番目の要素 $b_{j,k}$ を，名詞 n_j が名詞 n_k の顕現性を励起する度合であると見なすと，連想を加味した名詞 n_k の顕現性 $salience(n_k, U)$ は以下のように推定できる．

$$salience(n_k, U) = \sum_{j=1}^{N} b_{j,k} \Pr(n_j|\mathrm{pre}(U)) \quad (2)$$

ここで，発話単位 U 時点における各名詞 n_k の (連想を加味した) 顕現性 $salience(n_k, U)$ を k 番目の要素とする N 次元ベクトルを “**顕現性ベクトル**” $\boldsymbol{V}(U)$ と名づける．式 1,2 より，参照確率ベクトル $\boldsymbol{v}(U)$ に対して $\boldsymbol{b}_{n_1}, \cdots, \boldsymbol{b}_{n_N}$ を用いた以下の操作を加えることで $\boldsymbol{V}(U)$ を得ることができる．

$$\begin{aligned}\boldsymbol{V}(U) &= \begin{bmatrix} b_{1,1} & \cdots & b_{N,1} \\ \vdots & \ddots & \vdots \\ b_{1,N} & \cdots & b_{N,N} \end{bmatrix} \begin{bmatrix} \Pr(n_1|\mathrm{pre}(U)) \\ \vdots \\ \Pr(n_N|\mathrm{pre}(U)) \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{b}_{n_1} & \cdots & \boldsymbol{b}_{n_N} \end{bmatrix} \boldsymbol{v}(U)\end{aligned}$$

つまり $\boldsymbol{V}(U)$ は，$\boldsymbol{b}_{n_1}, \cdots, \boldsymbol{b}_{n_N}$ を基底とする斜交座標系における $\boldsymbol{v}(U)$ である．言い換えると，$\boldsymbol{V}(U)$ は $\boldsymbol{v}(U)$ を関連語軸方向へと回転させたものである．$\boldsymbol{v}(U)$ が連想を加味せずに推定した注意状態だとすれば，$\boldsymbol{V}(U)$ は連想を加味して推定した注意状態である．

4 実験と考察

■ **実験方法** 名詞 n_jを「**アメリカ村**」とした場合について 3.1 節の $\boldsymbol{b}_{n_j}$ をコーパスから獲得し，その結果を定性的に分析した．ただし，ここでは文単位を発話単位 U として扱った．以下に実験方法を示す．

0.a. **参照確率の回帰モデル学習:** 人手で照応タグが付与された GDA コーパス (毎日新聞 1,356 記事) から無作為抽出された 60,000 サンプルについて素性 (表 2) を抽出し，k=50 の k-NN 平滑化と TinySVM のサポートベクター回帰 (2 次多項式カーネル) を用いて参照確率の回帰モデルを学習した．

[*2] ここには，「注意状態を持つ主体の知識体系をコーパスが表せている」という仮定がある．

0.b. **コーパス収集:** 検索エンジン Google での検索語「アメリカ村」に対する検索結果から HTML 69 文書を収集し，CaboCha の構文解析結果を表す GDA タグを自動付与した．

1. **名詞句抽出:** 69 文書に含まれていた名詞句を，複合語も含めて $N = 6,293$ 語抽出した．
2. $\boldsymbol{v}_(U)$ **推定:** 参照確率の回帰モデルを用い，各文単位 U に対して $\boldsymbol{v}(U)$ を推定した結果を XML 属性として自動付与した．
3. $\boldsymbol{b}_{\text{アメリカ村}}$ **の計算:** 「アメリカ村」に着目している文 U における平均的な注意状態 $\boldsymbol{b}_{\text{アメリカ村}}$ を獲得した．ただし，3.1 節の手順 3.a における閾値 θ を 0.1 と 0.2 の 2 通りに設定し，試行した．

■ **実験結果** 閾値 θ を 0.1 と 0.2 にした場合に獲得した $\boldsymbol{b}_{\text{アメリカ村}}$ の要素を以下に示す．ただし，わかりやすいように値でソートした．

- $\theta = 0.2$ **のときの** $\boldsymbol{b}_{\text{アメリカ村}}$ **の要素:**
  アメリカ村:0.647, アメリカ:0.369, 大阪:0.258, 村:0.159, 防犯カメラ:0.139, カメラ:0.139, チェックアウト:0.129, アウト:0.129, 中:0.128, 女性:0.120, 男:0.102, 中央:0.098, 犯行:0.092, 人:0.087, たこ焼き:0.082, 心斎橋:0.075, ミナミ:0.074, 警察:0.073, $\cdots$
- $\theta = 0.1$ **のときの** $\boldsymbol{b}_{\text{アメリカ村}}$ **の要素:**
  アメリカ村:0.549, アメリカ:0.432, 大阪:0.280, 防犯カメラ:0.212, カメラ:0.212, 女性:0.141, 村:0.134, 人:0.119, 中:0.114, 犯行:0.114, 街:0.104, 被害:0.093, 男:0.092, 御津:0.086, 今回:0.084, 名前:0.084, 中央:0.084, 撲滅作戦:0.082, 作戦:0.082, 心斎橋:0.082, $\cdots$

■ **考察** この結果から，「大阪」「防犯カメラ」「心斎橋」「ミナミ」など，妥当と思われる関連語軸の方向へ $\boldsymbol{b}_{\text{アメリカ村}}$ が傾斜していることが確認できる．ただし，「防犯カメラ」などはコーパスの時事ニュース記事に大きく依存した例であり，応用に応じたコーパス選択の重要性を示唆している．

また，$\boldsymbol{v}(U)$ における「アメリカ村」の顕現性割合 $\frac{\Pr(\text{アメリカ村}\,|\text{pre}(U))}{\|\boldsymbol{v}(U)\|}$ の閾値 θ (3.1 節の手順 3.a) を変化させて試行したことにより，以下の知見を得た．

- $\theta = 0.2$ のときよりも $\theta = 0.1$ のときの方が $\boldsymbol{b}_{\text{アメリカ村}}$ における関連語軸の値が大きい (つまり，$\boldsymbol{b}_{\text{アメリカ村}}$ が関連語軸方向へ大きく傾いた)．
- θ を大きくして 1 に近づけると，$S_{\text{アメリカ村}}$ を抽出するとき「アメリカ村」に対する注目度が高い文に絞り込んで $\boldsymbol{b}_{\text{アメリカ村}}$ を獲得することになるので，連想の影響が小さくなる．また，θ が 1 に近づくほど $S_{\text{アメリカ村}}$ に含まれる $\boldsymbol{v}_{\text{アメリカ村}}(U)$ の数が少なくなるので，データの偏りに影響されやすくなる．
- 逆に，θ を小さくして 0 に近づけると，「アメリカ村」に少ししか関連しない文も含めて $\boldsymbol{b}_{\text{アメリカ村}}$ を獲得することになる．よって θ を小さくし過ぎると，連想の影響が過度に大きくなる恐れがある．

5 まとめと今後の課題

■ **まとめ** コーパスから獲得した $\boldsymbol{b}_{n_1}, \cdots, \boldsymbol{b}_{n_N}$ を用いて連想を加味した顕現性手法を開発し，従来の参照確率の問題点を解決した．また，$\boldsymbol{b}_{\text{アメリカ村}}$ の獲得実験を行い，関連語軸方向への傾きの妥当性を定性的に確認した．更に，n_j が顕著な注意状態の抽出に用いる閾値 θ を変えることにより，連想の影響の強さを変化させることができるという知見を得た．

■ **今後の課題** 式 2 で顕現性を推定する手法の有効性を，本稿では未評価である．今後は，その定量的な評価実験を行う．更に，注意状態を仮定せず単純な共起頻度を用いて $\boldsymbol{b}_{n_1}, \cdots, \boldsymbol{b}_{n_N}$ を獲得する可能性も検討し，それに対する本手法の優位性の有無を検証すべきである．

謝辞 新聞記事 GDA コーパス研究利用を許諾された三菱電機株式会社と調整して下さった GSK に感謝する．本研究の一部は科研費，21 世紀 COE，SCAT の支援を受けた．

参考文献

[1] B. Grosz, A. Joshi, and S. Weinstein. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, Vol. 21, No. 2, pp. 203–225, June 1995.

[2] K. Hasida. Issues in Communication Game. *Proc. of COLING'96*, pp. 531–536, 1996.

[3] S. Lappin and H.J. Leass. An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*, Vol. 20, No. 4, pp. 535–561, 1994.

[4] S. Nariyama. Grammar for Ellipsis Resolution in Japanese. *Proc. of the 9th TMI*, pp. 135–145, 2002.

[5] S. Shiramatsu et al. Meaning-Game-based Centering Model with Statistical Definition of Utility of Referential Expressions and Its Verification Using Japanese and English Corpora. *Proc. of DAARC2007*, March 2007. (to be appeard).

[6] M.A. Walker, M. Iida, and S. Cote. Japanese Discourse and the Process of Centering. *Computational Linguistics*, Vol. 20, No. 2, pp. 193–232, 1994.

[7] 白松, 宮田, 奥乃, 橋田. ゲーム理論による中心化理論の解体と実言語データに基づく検証. 自然言語処理, Vol. 12, No. 3, pp. 91–110, July 2005.