英語例文オーサリングのための可算性決定プロセスの可視化

永田　亮† 河合　敦夫†† 森広浩一郎† 井須　尚紀††
† 兵庫教育大学 †† 三重大学
E-mail: †{rnagata,mori}@hyogo-u.ac.jp, ††{kawai,isu}@ai.info.mie-u.ac.jp

1. はじめに

本論文では，英語名詞の可算性を可視化する手法を提案する．正確には，文中の名詞の可算生が決定されて行くプロセスを可視化する．また，可視化対象の名詞の周辺単語が，そのプロセスに寄与する度合いも可視化する．

可算性の可視化には，様々な応用可能性がある．例えば，英語教師に対する例文オーサリングの支援が挙げられる．可算生の決定プロセスを視覚的に把握できれば，任意の名詞に対して，典型的な可算／不可算の例，可算としても不可算としても捉えられる例，希少例など多様な例文が効率良く作成できる．また，可算生の可視化には，学習支援としての応用も考えられる．可算性は，適切な冠詞の撰択や単数／複数の決定に重要な役割を果たす [4] 一方で，日本語の名詞にはない概念であるため，日本人には理解が難しい [6]．名詞の可算生が決定されて行くプロセスを，周辺単語と共に可視化できれば，可算性のメカニズムを理解する上で大きな支援となる．

これまでに，可算生に関連した言語処理の研究は盛んになされている（例えば，文献 [3], [4], [9] など）．これらの研究の多くは，可算性クラス（または，Countability Preference [2]）に基づき名詞を分類することを目的としている．例えば，Bond ら [4] は，オントロジーを用いて，英語名詞を “fully countable”，“strongly countable”，“weakly countable”，“uncountable”，“plural only” の 5 種類の可算性クラスに分類する手法を提案している．この手法では，可算性の転換をほとんど受けない名詞を，“fully countable”，“strongly countable”，“plural only” として分類する．名詞 “cake” のように，容易に不可算に転換される可算名詞は，“strongly countable” として分類する．同様に，名詞 “beer” のように，容易に可算に転換される不可算名詞は，“weakly countable” と分類する．また，Baldwin ら [3] は，コーパスから可算性クラスを学習する手法を提案している．Nagata ら [9] は，別アプローチとして，文中の名詞が可算名詞であるか不可算名詞であるかを判定する手法を提案している．

しかしながら，可算性を可視化するためには，従来手法は不充分である．第一に，大部分の名詞は，文脈やその意味に応じて，可算名詞としても不可算名詞としても使用されるため，名詞を可算性クラスに分類する手法は不充分であるといえる．更に，文中の名詞を，単に可算であるか不可算であるかを判定する手法 [9] も不充分である．一般に，高い教育効果を挙げるためには正しい例を示すだけでなく，誤りの原因や解答プロセスの説明などが必要である [1]．したがって，可算性の学習支援や例文オーサリングの支援のためには，対象とする名詞の可算性決定に，何がどの程度寄与しているかを示すことが重要である．

そこで，本論文では，文中の名詞の可算性を可視化する手法を提案する．提案手法の可視化では，対象名詞の周辺単語が可算性の決定にどの程度寄与するかを視覚的に把握できる．結果的に，可算性を学習するための良い支援となる．また，提案手法では，対象名詞が可算／不可算であることを強く示す語を容易に抽出することが可能である．抽出した単語と可視化を利用することで，可算名詞および不可算名詞の例文オーサリングが効率的に行える．

以下，2. では，提案手法の基本アイデアを述べる．3. で，可視化手法を提案する．4. で，可算／不可算を強く示す語をコーパスから抽出する手法を提案する．5. では，提案手法を評価した実験について述べる．

2. 基本アイデア

英語名詞は，通常，文中では可算か不可算のどちらかで使用される [4] が，同じ名詞でも文脈によりその可算性に順位が付けられることがある．例えば，

Where is the *paper?*

では，名詞 “*paper*” は，可算であるか不可算であるか曖昧であるが，

Where is the important *paper?*

では，可算と認識するのが自然であるといえる．更に，

Where is the important *paper* she published?

では，より強く可算であると認識される．

この現象は，対象名詞が，ある文脈中で可算と認識される確率で表すことができる．上記一番目の例文では，その確率が 0.5 に近い値であるため，可算であるか不可算であるかが曖昧であるといえる．また，二番目と三番目の例文では，一番目の例文より高い確率値を持つことになる．

提案手法の基本アイデアは，この確率を利用して，可算性の可視化を行うというものである．上記の確率は，学習データを用いて推定する．推定した確率を，対象名詞の周辺単語

と共に，グラフとして描くことで可算性を可視化する．これが，提案する可視化手法の基本アイデアである．次章で，確率の推定方法と可視化手法について詳しく述べる．

3. 可視化手法

3.1 確率の推定

既に述べたように，提案手法では確率を可視化に利用する．この確率を定式化するために，以下では，対象名詞が可算と不可算であることを，それぞれ記号 c と u を用いて表す．また，対象名詞周辺の文脈を記号 x で表す．このとき，対象名詞がある文脈中で可算と認識される確率は，

$$P(c|x) = \frac{P(c)P(x|c)}{P(x)} \tag{1}$$

で定式化される．ここで，$P(c)$ は文脈について何の情報も与えられないときに，対象名詞が可算と認識される確率，すなわち事前確率である．事前確率 $P(c)$ は，文脈から得られる情報 $P(x|c)/P(x)$ により更新される．

ここで，Naive Bayes assumption [8] を利用して式 (1) を，

$$P(c|x) \approx \frac{P(c)\prod_{w_i \text{ in } x} P(w_i|c)}{\sum_{z\in\{c,u\}} P(z)\prod_{w_i \text{ in } x} P(w_i|z)} \tag{2}$$

と近似する．ただし，w_i は文脈 x 中の単語とする．式 (2) の右辺の確率は，学習データから推定する．事前確率 $P(c)$ は，

$$P(c) = \frac{o(c)}{N} \tag{3}$$

で推定する．ただし，$o(c)$ と N は，それぞれ学習データ中で可算とラベル付けされた対象名詞の数と対象名詞の総数とする．確率 $P(w_i|c)$ は，

$$P(w_i|c) = \frac{o(w_i,c)+\alpha}{\sum_t (o(w_t,c)+\alpha)} \tag{4}$$

で推定する．ただし，$o(w_i,c)$ と α は，それぞ学習データ中での w_i の頻度と平滑化パラメータである．

3.2 学習データの生成

式 (2) 中の確率を推定するためには，対象名詞が可算であるか不可算であるかラベル付けされた学習データが必要となる．幸い，文献 [9] で，名詞に可算／不可算のラベルを自動的に付与する手法が提案されているので，本論文でもその手法を利用して学習データを自動生成する．以下，学習データの自動生成について説明する．なお，名詞に可算／不可算のラベルを付与する手法の詳細については，文献 [9] を参照されたい．

学習データの生成は，対象名詞ごとに行う．まず，主名詞として使用された対象名詞とその周辺の単語をコーパスから収集する．収集は，既存の構文解析器や句解析器を用いて行う．次に，手法 [9] を利用して，収集した対象名詞に，可算／不可算のラベルを付与する．例えば，通常，不定冠詞は可算名詞のみを修飾するので，不定冠詞に修飾された主名詞には可算のラベルが付与できる．同様に，無冠詞単数の場合は，不可算のラベルが付与できる．次に，可算／不可算のラベルが付与された主名詞の周辺単語から，以下の単語を抽出する(注 1)：(i) 主名詞が存在する名詞句内の単語，(ii) その名詞句から左 5 単語，(iii) その名詞句から右 5 単語．抽出の際に，全ての単語は，小文字かつ原形に変換する．また，可算性に関係しないと考えられる語は，ストップワードとして抽出しない．最後に，抽出した単語に上記 (i) 〜 (iii) のうち該当する文脈情報を付加し，可算性のラベルと共に保存し，学習データとする．例えば，例文（対象名詞：*paper*），

She read new *papers*/可算 in her room.

からは，学習データ，

-5=read, NP=new, +5=in, +5=room, LABEL=可算

が生成できる（“-5=”，“NP=”，“+5=” は，文脈情報を表す）．

3.3 可算性の可視化

提案手法では，推定した確率を線グラフとして描く．グラフの横軸に，対象名詞の文脈中の単語を取り，縦軸に確率を取る．プロットする座標は，$(i, P(c|x = w_1 w_2 \cdots w_i))$ で与えられる．ただし，w_i は，対象名詞の文脈中の i 番目の単語とする．グラフの可読性を高めるため，縦軸の 0，0.5，1 に対応するラベルを，それぞれ “Uncountable”，“Neutral”，“Countable” と表示する．同様に，横軸の i に対応するラベルに単語 w_i を表示する．更に，横軸の $i = 0$ に，文頭記号 ϕ を表示し，事前確率 $P(c)$ をプロットできるようにする．

図 1 に，対象名詞を “*paper*” とした可視化の例を示す(注 2)．横軸上の記号*が付いた単語（She*, the*など）は，ストップワードとして，確率の推定に用いなかった単語である．

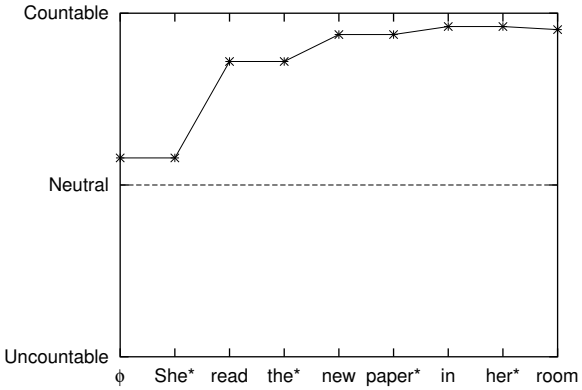

図 1： 可視化の例（対象名詞：*paper*）

(注 1)：ただし，文を越えては抽出しない．

(注 2)：確率の推定には，British National corpus [5] を用いた．詳細は，5. に示す．

図 1 から，対象名詞 “*paper*” は文脈なしでは，ほぼ中立の可算性を持つと認識されることがわかる（i.e., ϕ）．最初の単語 “She*” は，ストップワードとして，確率の推定から除外された単語である．次の単語 “read” で，グラフは大きく上昇し，対象名詞が可算であることを強く示している．その後，グラフは安定して上昇し続け，最終的に，非常に強く可算であると認識されることがわかる．

この例のように，提案する可視化手法では，対象名詞が可算／不可算と認識されるプロセスを視覚的に把握できる．その結果， 可算性の決定に寄与する要因を特定でき，学習支援や例文オーサリングの支援に有益となる．

4. 可算性に寄与する語の抽出

対象名詞の可算性の決定に大きく寄与する語を集めることができれば，可算性に関する例文オーサリングの支援に利用できる．すなわち，可算性の可視化と集めた語に基づき，例文に対して語の追加／置換／削除を行うことで，効率良くオーサリングが行える．

提案手法の枠組みでは，対象名詞が可算と認識されるのに大きく寄与する単語は，$P(w|c)$ が大きく，$P(w|u)$ が小さい単語である．逆に，不可算と認識されるのに寄与する語は，$P(w|u)$ が大きく，$P(w|c)$ が小さい単語である．

このことに基づき，単語 w のスコアを，

$$g(w) = \log \frac{P(w|c)}{P(w|u)} \qquad (5)$$

で定義する．大きな正のスコアを持つ単語 w は，対象名詞が可算と認識されるのに大きく寄与する語である．大きな負のスコアを持つ単語 w は，対象名詞が不可算と認識されるのに大きく寄与する語である．したがって，スコアの降順で単語をソートすれば，可算と認識されるのに寄与する単語のリストが得られる．昇順でソートすれば，不可算と認識されるのに寄与する単語のリストが得られる．ただし，“a” や “each” などの denumerator [2] は，ソートしたリストに含めない．なぜなら，denumerator は，殆んどの場合，対象名詞の可算性を可算と決定するためソートしたリストに含めても，得られる情報が無いからである．

表 1 にリストの例を示す．対象名詞は，“*paper*” である．左側のカラムは，可算と認識されるのに寄与する単語のリストである．右側のカラムは，不可算と認識されるのに寄与する単語のリストである．表 1 から，“*paper*” の可算の例文を作るときには，例えば，“researcher” を主語にしたり，“scientific” を修飾語にして例文を作れば良いことがわかる．

表 1： 対象名詞 “*paper*” の可算性に寄与する語

Countable		Uncountable	
w	$g(w)$	w	$g(w)$
-5=researcher	3.83	-5=piece	-4.26
NP=scientific	3.79	-5=sheet	-3.83
+5=thesis	3.72	NP=kitchen	-3.78
NP=research	3.52	NP=greaseproof	-3.57
-5=geology	3.45	-5=tin	-3.55
+5=bias	3.42	NP=waste	-3.48
NP=publish	3.40	-5=spin-dry	-3.48

5. 評価実験

5.1 実験条件

本実験では 2 種類のタスクを行い，提案手法 を評価した．第一のタスクでは，式 (2) で計算した確率で，対象名詞の可算／不可算の判定を行い，間接的に確率の推定精度を評価した．第二のタスクでは，可視化と 4. の手法で作成した単語リストで，可算／不可算の例文を修正した．修正結果を，中学校で英語教師をするネイティブスピーカに提示し，改善率を求めた．

確率の推定は，British National Corpus [5] を用いた．また，名詞句の抽出には，OAK system(注 3)を用いた．

テストデータとして文献 [7] に示される 25 種類の名詞とその例文を用いた．各名詞とも，それぞれ可算と不可算の例文が一つづ集録されており，計 50 例文が得られた．

5.2 実験手順

まず，25 種類の名詞に対して，学習データを生成し，確率を推定した．平滑化パラメータは，$\alpha = 1$ とした．

次に，各例文を OAK System で解析し，$P(c|x)$ を計算した．解析誤りのため対象名詞が抽出できない例文に対しては，解析誤りを人手で修正した．計算された確率が $P(c|x) > 0.5$ を満たしたときは，対象名詞を可算と判定した．そうでなければ，不可算と判定した．判定結果が，文献 [7] に示されている可算性と一致した場合に，判定成功とした．最終的には，判定成功数を判定数 50 で除した精度で評価を行った．

精度評価後，第二のタスクを実施した．50 の例文から，$P(c|x) \leqq 0.85$ を満たす可算の例文と $P(c|x) \geqq 0.15$ を満たす不可算の例文を取りだした（その結果，18 の例文が取りだされた）．これらの 18 の例文に対して，可視化結果を見ながら，4. の手法で作成した単語リストで例文の修正を行った．リスト中の単語を上から順に，意味的に適合するかを確認し，適合した場合は，その単語を例文に追加した．また，その単語で例文中の単語を置換可能であり，かつ，可算性のグラフが改善する場合は，置換を行った．この作業を，可算の例文の場合は $P(c|x) > 0.90$，不可算の例文の場合は $P(c|x) < 0.10$ を満たすまで繰り返した．

最後に，修正結果を評価した．修正前と修正後の例文のペアを，ネイティブスピーカに提示し，どちらの例文が可算／

（注 3）：OAK System homepage:
http://www.cs.nyu.edu/ sekine/PROJECT/OAK/

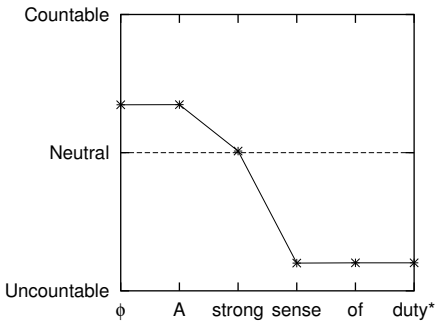

図 2： 可視化の例（対象名詞：*duty*）

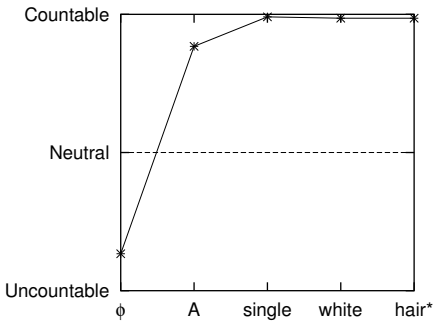

図 3： 可視化の例（対象名詞：*hair*）

不可算の例文として良いか選択してもらった．

5.3 実験結果と考察

第一のタスクでは，提案手法は判定精度 80%を達成した．Naive Bayes assumption にもかかわらず，対象名詞が可算／不可算に認識される確率を高い精度で推定できたことになる．判定に成功した例文に対する可視化結果は，第一著者の直感に合うものとなった．その一部を図 2(注 4)と図 3 に示す．図 2 では，“strong” と “sense” で，グラフが急激に下降しているのがわかる．単語 “strong” や “sense” は，対象名詞 “*duty*” が，概念的な「義務」を意味しており，不可算であることを強く示唆しているといえる．図 3 では，対象名詞 “*hair*” は，通常，不可算として認識されるが，一本一本の髪に言及するときは，可算に転換されることがわかる．これらの例のように，提案手法の可視化は，人間の直感に良く適合した．この適合により，提案手法は高い判定精度を達成した．

二番目のタスクでは，18 の例文のうち，8 例文のみが修正後のほうが良いと判定された（7 例文については，修正前のほうが良いと判断され，残り 3 例文は同程度良いと判断された）．すなわち，修正による例文の改善はほとんど見られないたことになる．しかしながら，実験後，ネイティブスピーカにインタビューを行ったところ，短い例文が好まれていることが判明した．その理由として，例文が短ければ，対象名詞と周辺単語の関係が把握しやすく，対象名詞の可算性の認識が容易になるからであるということが得られた．ところが，修正後の例文では，$P(c|x) > 0.90$ または $P(c|x) < 0.10$ を満たすため，複数の単語が追加され，文の長さが長くなる傾向にあった．

そこで，例文の長さを考慮して再度評価を行った．修正した例文のうち，単語リストから追加した単語数が 1 である例文のペアに対して評価を行った．その結果，約 80%（9 ペア中 7 ペア）で，修正後のほうが良いと判断されていることが判明した．すなわち，提案手法により，大部分の例文が改善されたことになる．以上の結果から，可算／不可算の例文の良さには，少くとも二つの重要な要因があるといえる：(i) 可算性を強く示す語を含むこと，(ii) 比較的短い例文であることである．

6. おわりに

本論文では，文中の名詞の可算性を可視化する手法を提案した．提案手法では，対象名詞が，可算／不可算として認識される確率をコーパスから推定し，その確率をグラフとして描くことで可視化を行う．実験の結果，精度良く可視化が行え，可算／不可算の例文オーサリングを支援可能であることを確認した．

(注 4)：著作権を考慮し，例文の一部を示す．

参考文献

[1] 奥畑健司, 島崎克也, 太田義一, 野村康雄, and 溝口理一郎, “教育戦略の観点から見た教材知識の分類,” 電子情報通信学会論文誌, vol.J 75-A, no.2, pp.305–313, 1992.

[2] K. Allan, “Nouns and countability,” J. Linguistic Society of America, vol.56, no.3, pp.541–567, 1980.

[3] T. Baldwin and F. Bond, “A plethora of methods for learning English countability,” Proc. of 2003 Conference on Empirical Methods in Natural Language Processing, pp.73–80, 2003.

[4] F. Bond and C. Vatikiotis-Bateson, “Using an ontology to determine English countability,” Proc. of 19th International Conference on Computational Linguistics, pp.99–105, 2002.

[5] L. Burnard, Users Reference Guide for the British National Corpus. version 1.0, Oxford University Computing Services, Oxford, 1995.

[6] Y.G. Butler, “Second language learners’ theories on the use of English articles,” Studies in Second Language Acquisition, vol.24, no.3, pp.451–480, 2002.

[7] R. Huddleston and G. Pullum, The Cambridge Grammar of the English Language, Cambridge University Press, Cambridge, 2002.

[8] C.D. Manning and H. Schütze, Foundations of Statistical Natural Language Processing, The MIT Press, Massachusetts, 1999.

[9] R. Nagata, T. Wakana, F. Masui, A. Kawai, and N. Isu, “Detecting article errors based on the mass count distinction,” Proc. of 2nd International Joint Conference on Natural Language Processing, pp.815–826, 2005.