E-025

対象となる名詞を選ばない規則を用いた可算／不可算判定手法

A Method for Distinguishing English Mass and Count Nouns with Target-Independent Rules

若菜　崇宏† 永田　亮‡ 河合　敦夫† 桝井　文人† 井須　尚紀†
Takahiro Wakana Ryo Nagata Atsuo Kawai Fumito Masui Naoki Isu

1．はじめに

日本人英語学習者が書いた英文に多く見られる，冠詞の誤りや単数／複数の使い分けに関する誤りを検出するためには，名詞の可算／不可算の判定が重要である．なぜなら，可算／不可算の情報が与えられると，表 1 の×で示される部分が誤りで有ることが分かり，上記の誤り（以下では，表記を簡単にするため，これらの誤りを冠詞誤りと呼ぶことにする）が検出できるからである．例えば“I have a furniture.”という英文で，“furniture”が不可算名詞であることが分かれば，表 1 から冠詞の余剰として検出できる．一方，表 1 の○で示される部分は，可算／不可算の情報からは誤りであるかどうかの判断は出来ず検出対象外となるが，学習者の書いた英文においては×で示される部分に比べ少ない誤りである．

表 1　可算／不可算に基づいた誤り検出ルール

	単数			複数		
	不定冠詞	定冠詞	無冠詞	不定冠詞	定冠詞	無冠詞
可算	○	○	×	×	○	○
不可算	×	○	○	×	×	×

誤りを含まない英文では，冠詞や単数／複数などの表層情報から可算／不可算の判定は比較的容易に行える．例えば，複数形の名詞は可算名詞であるし，無冠詞単数の名詞は不可算名詞である．一方，誤りを含む英文では，冠詞や単数／複数の用法が間違っている可能性があるので，これらの表層情報を用いることが出来ない．従って，冠詞誤りを検出する際には，これらの表層情報を利用しない可算／不可算の判定手法が必要となる．

可算／不可算を判定する先行研究として，英文コーパスから可算／不可算の判定規則（以下，表記を簡単にするため，単に判定規則とする）を生成する手法[4]がある．この手法では可算／不可算を判定の対象となっている名詞（以後ターゲット名詞と呼ぶ）の文脈情報（本論文では，ターゲット名詞周辺の単語のことを文脈情報と呼ぶ．冠詞や代名詞などの機能語は除外する）に基づいて可算／不可算の判定を行う．例えば，ターゲット名詞 paper に対する判定規則は，次のような形で生成される：

規則 1	read[-]	→	可算
規則 2	pencil[+]	→	不可算
:			
規則 10	pulp[+]	→	不可算
規則 11	author[+]	→	可算
:			
規則 n	*paper*	→	可算

ここで[-][+]はターゲット名詞の現れた名詞句からの位置関係であり，それぞれ名詞句の前，及び後ろに現れたことを示す．例えば規則 1 は「read が *paper* の現れた名詞句より前に現れたら可算と判定」という意味である．規則 n はデフォルト規則で，英文コーパス中で *paper* が現れたとき，可算／不可算のどちらが多かったかを示す．デフォルト規則は規則 1〜n-1 に適用可能な規則が無い場合にのみ用いられる．

この判定規則を用いて，可算／不可算の判定を行う．例えば，

例文（1）I read the *paper*.
例文（2）The *paper* is made of hemp pulp.

中の *paper* に対して可算／不可算の判定を行うことを考える．ここでターゲット名詞の文脈情報を見ると，例文（1）では規則 1 が，例文（2）では規則 10 が上記判定規則に当てはまることが分かる．よって例文（1）の *paper* は可算，例文（2）の *paper* は不可算と正しく判定される．

しかしながら，上記例で用いた判定規則はターゲット名詞を *paper* に限定し，*paper* と共に用いられた単語を規則としている．このため，学習データ中で，*paper* と共起しなかった単語は，規則に利用されない．例えば下記英文でターゲット名詞を *paper* とし下記英文

…She wrapped a thing in *paper*…

が与えられたとする．もし学習時データ中で *paper* が wrap，thing，in のいずれとも共起しなかった場合，デフォルト規則が適用され，可算と判定される．しかしながら，デフォルト規則は対象名詞が学習データ中で可算／不可算のどちらで多く使われているかに基づいて判定を行うので，文脈情報に基づいた規則より判定精度は低くなる傾向にある[4]．上記の例でも，デフォルト規則は誤ってターゲット名詞を可算と判定している．一方で，wrap，thing，in が判定規則に含まれていたならば，正しく不可算と判定される可能性が高いといえる．すなわち，判定規則を増加させることで，この問題が解決できると言える．

そこで本論文では，判定規則を効率良く増加させる手法を提案する．提案手法では，ターゲット名詞を特定の単語に限定せず，コーパス中に出現する全ての名詞を 1 つの名詞として扱う．そうすることで，あらゆる名詞に適用可能な判定規則（以下，一般判定規則）を学習する．一般判定規則と従来のターゲット名詞を限定した判定規

†三重大学，Mie University
‡兵庫教育大学，Hyogo University of Teacher Education

則（以下，限定判定規則）と組み合わせることで，規則の増加を図り，判定精度を向上させる．

2. 一般判定規則

2.1 学習データの自動生成

限定判定規則及び一般判定規則は，ターゲット名詞の可算／不可算の例からなる学習データから学習される．可算／不可算の例とは

She gave a ***paper***／可算 on wild animals.

の様に，ターゲット名詞に可算／不可算のタグが付与されたものである．

学習データはコーパスから以下の手順により自動生成される．

（1）ターゲット名詞の抽出
（2）ターゲット名詞のタグ付け
（3）タグ付けされたターゲット名詞の保存

（1）では主名詞として使用されているターゲット名詞をその周辺の単語とともにコーパスから抽出する．この処理は既存の構文解析などで行うことが出来る．ただし，本手法では一般判定規則を学習するので，ターゲット名詞は限定しない．よって，コーパス中のあらゆる名詞を 1 つの名詞とみなして抽出する．抽出の際には，単語を小文字かつ原形（例えば，Boxes から box）に変換する．ただし，表 2 中の単語，代名詞や助動詞などの機能語，基数，ターゲット名詞は抽出しない．

（2）では以下に述べるルールを用いて，抽出されたターゲット名詞に可算／不可算のタグを付与する．例えば，

She gave a ***paper*** on wild animals.

中の *paper* は単数形で不定冠詞が付いている事から

She gave a ***paper***／可算 on wild animals.

とタグ付け出来る．

図 1 と表 2 に，言語学の知見[1][2][3]に基づいて作成した可算／不可算のタグ付けのためのルールを示す．なお，詳細については文献[4]を参照されたい．

図 1 中のノードは，ターゲット名詞に適用される質問を表す．例えば，ルートノードは，「ターゲット名詞は複数形であるか．」と解釈される．また，図 1 中の葉は分類結果に対応する．例えば，ルートノードの質問の答えが“yes”であれば可算と分類される．“no”の場合は，次のノードの質問が適用される．図中の“?”は，ルールによって可算／不可算の分類が出来ないことを表す．

（3）で，上記ルールによって可算／不可算のタグ付けされたターゲット名詞とその周辺の単語を保存し，学習データとする．

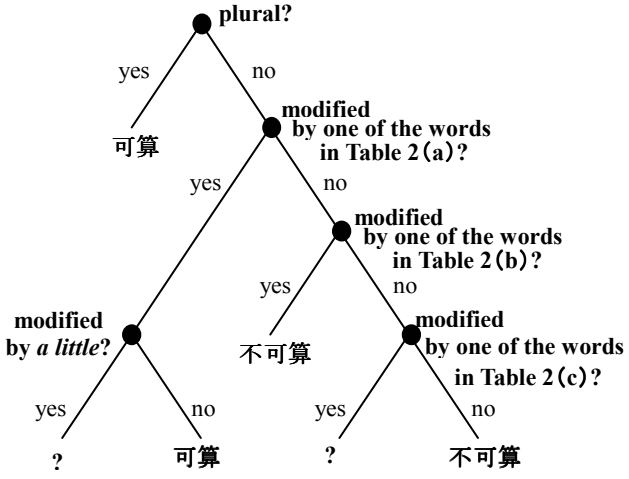

図 1　学習データ生成のためのルール

表 2　図 1 中で使用される単語群

(a)	(b)	(c)
the indefinite article	much	*the definite article*
another	less	*demonstrative adjectives*
one	enough	*possessive adjectives*
each	all	*interrogative adjectives*
-	sufficient	*quantifiers*
-	-	*'s genitive*

2.2 一般判定規則の学習

一般判定規則のテンプレートを定義するため次の記号を導入する．ターゲット名詞が可算／不可算になることを変数 *MC* を用いて表す． *MC* は可算／不可算を値にとると定義する．また，単語を *w*，ターゲット名詞周辺の文脈を *C* で表す．文脈 *C* として，NP（ターゲット名詞が主名詞となっている名詞句内の単語），±（その名詞句から左（-）または右（+）に 3 単語）の 3 種類を定義する．このときテンプレートは

単語 *w* が 文脈 *C* に現れたら *MC* と判定

と定義する．以下表記を簡単にするためテンプレートを

$$w_C \rightarrow MC \qquad (1)$$

で表すことにする．

次に 2．1 で説明した学習データを用いて一般判定規則の学習を行う．まず，学習データからターゲット名詞周辺の文脈に現れる単語を抽出しテンプレートに適合する規則を生成する．

以下に規則の生成例を示す． いま，学習データ

He cooked ***beef***／不可算 and fish for dinner.
I ate a piece of ***chicken***／不可算 with salad.

が与えられたとする．ターゲット名詞を *chicken* として限定判定規則を学習する場合，規則

eat[-] → 不可算, piece[-] → 不可算, salad[+] → 不可算

が生成される．一方，一般判定規則を学習する場合は，規則

cook[-] → 不可算, fish[NP] → 不可算, dinner[+] → 不可算
eat[-] → 不可算, piece[-] → 不可算, salad[+] → 不可算

が生成される．

次に，生成されたルールの重要度を決定するために対数尤度比を計算する．対数尤度比は，w_C が成立するときにターゲット名詞が MC となる条件付き確率 $p(MC|w_C)$ を用いて

$$\log \frac{p(MC|w_C)}{p(\overline{MC}|w_C)} \quad (2)$$

で計算される．ここで $\overline{MC}$ は MC の排反事象である．

条件付き確率 $p(MC|w_C)$ を学習データから推定する．いま，$f(w_C)$ を，学習データ中で w が C に出現した頻度とする．同様に，$f(w_C,MC)$ を，学習データ中で w が C に出現したときにターゲット名詞が MC となった頻度とする．このとき，条件付き確率を，

$$p(MC|w_C) = \frac{f(w_C,MC) + 0.5}{f(w_C) + 1.0} \quad (3)$$

で推定する．

2.3 限定判定規則のデフォルト規則

デフォルト規則とは，判定規則中の他の規則によって可算／不可算の判定が行えないときに使用される規則である．いま，ターゲット名詞を t で表すことにする．また，ターゲット名詞を限定した場合の学習データ中で，頻度が高い方の MC の値を $MCmajor$ で表す．このとき，デフォルト規則のテンプレートは

$$t \rightarrow MCmajor \quad (4)$$

と定義される．これは「ターゲット名詞が出現したら頻度の高い方の MC で判定」と解釈できる．

一般判定規則でもデフォルト規則は学習されるが，本論文では一般判定規則のデフォルト規則は扱わないものとした．

3. 可算／不可算の判定

ターゲット名詞の可算／不可算の判定は，2. で学習した一般判定規則と限定判定規則を組み合わせて行う．図 2 にターゲット名詞を *chicken* としたときの可算／不可算の判定の流れを示す．

まず，ターゲット名詞である *chicken* の限定判定規則の集合から，対数尤度比の高い順に適用可能な規則を検索する．この時点で適用可能な規則があればその規則に従って判定を行う．ただし，適用可能な規則がデフォルト規則のみである場合，この時点ではデフォルト規則は使用しない．限定判定規則に適用可能なものが無い場合は，一般判定規則から適用可能な規則の検索を行う．この時，限定判定規則のデフォルト規則よりも対数尤度比の低い規則は検索対象外とした．この時点でも適用可能な規則が無い場合は限定判定規則のデフォルト規則による判定を行う．図 2 では，限定判定規則には適用可能な規則が無く，一般判定規則の own[NP] → 可算が適用可能であることが分かるので，この時点で規則の適用を止め，可算と判定する．

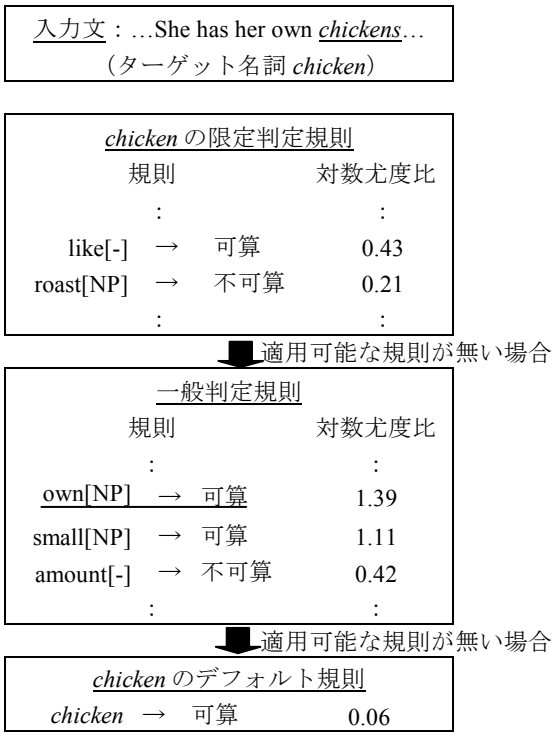

図 2　規則適用の流れ

4. 実験と評価

4.1 実験条件

本実験では，文献[3]に可算／不可算の両方で使用される名詞として上げられている 23 種類の名詞をターゲット名詞とした．

コーパスには，BNC を用いた．BNC 中のテキストタグで囲まれた部分を一つのテキストとして使用した．ただし，話し言葉のタグが付与されているテキストは除外した．また，長すぎるために，実験に用いたツールで解析できなかった文も除外した．

本実験では，精度を用いて提案手法の判定性能を評価した．精度は

$$\frac{\text{正しく判定されたターゲット名詞の数}}{\text{判定したターゲット名詞の数}} \times 100 \quad (5)$$

で定義した．

本実験では，tenfold cross-validation[5]を用いて，提案手法の性能を評価した．まず，上記コーパス中のテキストから 10 セットに分割した．ただし，分割はランダムに行い，各セットに含まれるテキストの数がほぼ等しくなるように行った．この結果，10 セットの合計コーパスサイ

ズは約 7400 万語となった．次に，10 セットのうち，1 セットを評価データ用のコーパス，残り 9 セットを学習データの生成，限定判定規則，一般判定規則の学習に使用し，1 セット中のターゲット名詞の可算／不可算を判定し性能を評価した．

4.2 評価結果と考察

表 3 に，評価結果を示す．表 3 中の“平均出現数”とは，評価データ中のターゲット名詞の平均出現数を表す．また，“BL”はターゲットを限定した場合の判定規則におけるデフォルト規則のみで判定したときの精度を示している．

表 3 から，提案手法は，従来手法よりも精度が良いことが確認出来る．また，両者の平均精度には有意水準 5%で有意差が見られた（paired *t*-test）．従来手法では限定判定規則のみを用いているので，限定判定規則の中に適用可能な規則が無い場合，文脈情報を用いていないデフォルト規則による判定を行う．しかし表 3 の“BL”の示す通り

表 3　評価結果

名詞	平均出現数	BL	従来手法	提案手法
advantage	628	61.3	89.8	88.8
aid	581	80.4	87.0	88.9
auth	1664	74.8	80.3	81.0
building	925	78.0	79.7	79.7
cover	228	63.8	72.9	74.0
detail	1301	74.0	88.5	88.6
discipline	230	61.3	76.1	76.4
duty	638	67.0	79.5	80.7
football	207	92.9	94.0	94.5
gold	252	92.8	92.8	92.9
hair	422	86.3	87.9	88.4
improvement	427	72.5	75.6	75.8
necessity	93	53.9	80.1	81.1
paper	1029	59.2	82.2	82.5
reason	1352	83.0	84.9	85.9
sausage	50	78.2	77.4	74.3
sleep	152	84.4	87.8	88.5
stomach	39	66.8	72.4	72.8
study	1683	74.6	79.8	80.7
truth	244	75.3	80.7	81.7
use	1474	87.4	88.2	88.4
work	3129	80.3	83.9	84.5
worry	122	79.6	84.1	84.5
平均	733	75.1	82.9	83.2

デフォルト規則による判定では，文脈情報を用いた規則に比べ精度が低下する傾向にある．そこで本手法では限定判定規則に適用可能な規則が無かった場合，すぐにデフォルト規則を適用するのではなく，一般判定規則の検索も行う．そうすることで従来手法に比べ，文脈情報を用いた規則を増加させることができ，判定精度が向上した．例えば，ターゲット名詞を *worry* とした下記英文

…stage could solve small ***worries*** before they become…

では，限定判定規則に適用可能な規則が無く，一般判定規則の small[NP] → 可算が使用された．small は名詞句内に現れると，名詞を可算にする可能性が高い単語である．このことは，small[NP] → 可算の対数尤度比が 1.11 と高いことからもわかる．この例のように，一般判定規則により，規則の不足が緩和され精度向上につながった．

提案手法の問題点として，精度改善率にばらつきがあることが挙げられる．1.9%と大きな改善を見せた aid に比べ，gold などは 0.1%しか改善していない．

提案手法の改善案として，学習データ中で一般判定規則が共起したターゲット名詞の種類数を考慮することが考えられる．より多くのターゲット名詞と共起し，なおかつ，対数尤度比が高い一般判定規則は，より一般的に名詞を可算または不可算にする規則といえる．そこで，共起したターゲット名詞の種類数に応じて一般判定規則の対数尤度比に重みをつけることで，一般判定規則の改善が期待できる

5. まとめ

本論文では，従来の名詞の可算／不可算判定手法における規則の不足を補う手法を提案した．具体的には，ターゲット名詞を特定の単語に限定せず，コーパス中に出現する全ての名詞を 1 つの名詞として扱う事で，判定規則の数を増加させる．この規則を従来手法と組み合わせて判定を行うことで，可算／不可算判定の平均精度を 0.3%向上することが出来た．今後の課題としては，この手法を冠詞誤り検出に応用した場合を調査することが挙げられる．

参考文献

[1] K. Allan, “Nouns and Countability,” Language: Journal of the Linguistic Society of America, 56（3）, pp. 541-567, Sep. 1980.

[2] B. Gillon, “The Lexical Semantics of English Count and Mass Nouns,” Proc.of the Special Interest Group on the Lexicon of the Association for Computational Linguistics, pp. 51-61, June 1996.

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

[4] R. Nagata, F. Masui, A. Kawai, and N. Isu, “An unsupervised method for distinguishing mass and count nouns in context,” Proc. of 6th International Workshop on Computational Semantics, pp.213-224, Jan. 2005.

[5] I. Witten and E. Frank,. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers. 2000.