# 可算/不可算名詞の判定に基づいた 科学技術論文における冠詞誤り検出の問題点と改善方法

深田 剛継，永田 亮‡，河合 敦夫，桝井 文人，井須 尚紀
（三重大学，兵庫教育大学‡）

## 1. はじめに

近年の国際化に伴い，我々日本人研究者でも英語で論文を執筆する機会が多くなった．しかし，その英語論文に毎回添削を依頼していると費用面でも時間面でも大きな負担となる．そこで英文中の誤りを自動で検出することで，費用，時間両面での効率化を計りたい．英文の誤り検出を行うにあたり，どのような誤りを検出できれば効果的かを考える必要がある．日本語には冠詞の概念がないため，日本人が書く英文には冠詞の誤りが多いことが報告されている[1]．

冠詞誤りを検出するためには，英語名詞の可算/不可算の情報が重要となる[1]．しかし，大部分の名詞は可算/不可算の両方で使用される．そのため永田ら[2]によって，可算/不可算の判定を行う手法が提案された．さらにこれを利用して，冠詞の誤りを検出する手法[3]が提案された．また，この検出手法では可算/不可算の判定を用いるため，冠詞の誤りだけでなく，単数/複数の誤りも検出することが出来る（本論文では冠詞誤りに単数/複数の誤りも含む）．この手法は学生の書いた趣味に関するエッセイを対象とした場合に有効であることが確認されている[4]．

この手法を用いて科学技術英文の冠詞誤りの検出を行った場合を考える．前述の通り，英文エッセイの筆者は学生であった．今回対象とする科学技術英文（論文）の著者は日本人研究者（大学教員）であるため，英語力に差が出ると考えられる．英語力に差が出る時の問題点を挙げる．和泉ら[7]によって，筆者の英語力で冠詞誤りの種類（a の欠落，the の余剰など．詳細は 4.1）の頻度が異なることが報告されている．例えば，英語力が上がるにつれて“冠詞の欠落誤りが冠詞の余剰と冠詞選択誤りに分散する”，“冠詞の余剰誤りの内，不定冠詞の余剰が減り，定冠詞の余剰が増える”ことが確認されている．可算/不可算判定による誤り検出手法では定冠詞の誤り検出は不可能な事が多い（詳細は 4.1）．また，技術論文と英文エッセイとでは使われる単語や表現が異なることが予想される．上記の理由より，英文エッセイを対象とした時と同じ手法で誤り検出を行っても同じ性能が出るとは考えにくい．

本論文では以下の構成をとる．2. で可算/不可算判定とそれを用いての誤り検出について説明する．3. で科学技術英文を対象とした場合の冠詞誤り検出の実験を行う．4. で実験結果を考察し，5. で改善方法の提案を行う．

## 2. 可算/不可算判定を用いた誤り検出

可算/不可算の判定を用いた冠詞誤りの検出は，可算/不可算を判定するための規則の学習，その規則を用いての可算/不可算の判定，可算/不可算判定を用いての冠詞誤り検出の順に行う．

### 2.1 可算/不可算判定用の規則の学習

まず最初に可算/不可算を判定するための規則を作成する．規則を作成するための学習データとして誤りを含まない英文に可算/不可算のタグを付与した文書集合を用いる．誤りを含まない英文では，可算/不可算の判定は表層情報を用いて比較的容易に行える[5]．学習データ中の対象名詞の可算/不可算の情報を基に周辺単語から学習していく．また，規則の作成は学習データ中の登場回数が 100 回以上だった名詞に対して行う．

例えば，Different alcohols mixed with... という文の場合，alcohol が複数形の alcohols となっているため，alcohol が可算名詞だということがわかる．そこでその周辺単語を見ることで different が alcohol の前に登場する場合や mix が alcohol の後に登場する場合，alcohol は可算名詞となりやすい，といった規則を学習する．

次に，生成された規則の優先度を計算する(詳細は[3])．この数値が高い程，その単語は可算または不可算になりやすい．

### 2.2 可算/不可算の判定

対象名詞の可算/不可算の判定は，判定規則中の規則を優先度の大きい順に適用し，適用可能な規則が見つかった時点で，その規則に従って判定する．適用可能な規則が見つからない場合は，学習データ全体でその単語が可算/不可算のどちらで使用されることが多かったかを元にした Default 規則と呼ばれる規則で判定を行う．

### 2.3 冠詞誤りの検出

冠詞誤りは可算/不可算の情報と対象名詞の表層情報に基づいたルール(表 1)によって検出する．表中の○が文法に基づくとその用法で使用可能，×が使用不可能を表す．

表 1. 名詞の可算/不可算に基づいた誤り検出ルール

| | 単数形 | | | 複数形 | | |
|---|---|---|---|---|---|---|
| | 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 | × | ○ | ○ | × | × | × |

## 3. 冠詞誤り検出実験

パターン認識分野と材料科学分野の論文に対して，それぞれの論文と同分野の学習データを用い実験を行った．

学習用データはScienceDirect社の論文誌の中から，パターン認識分野として「Pattern Recognition」より，597報/約270万語，材料科学分野として「Journal of Non-Crystalline Solids」より，620報/約170万語を用いた．

誤り検出の対象として，それぞれパターン認識分野，材料科学分野の日本人研究者が書いた文字認識，セラミックの論文を用いた．対象論文に含まれるネイティブの添削に基づく誤りの数は，パターン認識分野で61箇所，材料科学分野で89箇所となった．また，本システムが誤り検出の対象とする名詞は，名詞句の最後に登場する名詞とし，パターン認識分野で399箇所，材料科学分野で441箇所となった．これらの名詞に可算/不可算の判定と冠詞誤りを含むかどうかの判定を行う．

その結果は表2となった．表中のRecallは全ての誤りの内，どれだけの誤りを見つけられたかを表す．Precisionはシステムで誤りと判定した内，どれだけが実際に誤りだったかを表す．表2より可算/不可算の判定性能に比べて，冠詞誤り検出性能が低くなっていることがわかる．

例えばパターン認識分野の冠詞誤りは61箇所あった．規則の未作成など（詳細は4.3，4.4）の理由で10箇所が誤り検出対象外となった．可算/不可算の判定を行った51箇所中，44箇所はネイティブと同じ可算/不可算の判定結果となった．しかし，パターン認識分野のRecallは41.0%となっている．以下でこの性能低下の要因について考察する．

表2．誤り検出実験結果

| | 可算/不可算判定 | Recall | Precision |
|---|---|---|---|
| パターン認識 | 88.5% (299/338) | 41.0 (25/61) | 44.6 (25/56) |
| 材料科学 | 84.8 (263/310) | 39.3 (35/89) | 64.8 (35/54) |

## 4. 冠詞誤り検出失敗の考察

冠詞誤り検出がうまくいかなかった要因は大きく分けて４つであった．以下に各要因の詳細について述べる．

### 4.1 冠詞誤りの種類の違い

永田らの冠詞誤りの種類の調査[6]と本実験では，冠詞の誤りの種類毎の頻度が大きく異なる結果となった．詳細を表3に示す．冠詞誤りの種類毎の頻度が大きく異なった原因として，英文記述者の英語力や対象英文の違いなどが考えられる．

これらについて考察する．和泉ら[7]の調査から欠落の誤りが減り，theの余剰が増える予想を立てた．しかし，実際の結果では特にtheの欠落が増えたことが目立つ．これは特に材料科学分野で顕著に見られた．本実験の対象英文は技術論文であり，永田らの調査はエッセイであった．技術論文はエッセイ等の通常英文より，手法の説明等で繰り返し述べる名詞が多いなどの理由から同定可能な場合が多くなる事が予想できる．ただし，本実験では誤り検出用のデータ数が少なかったので，これを増やすことにより，これらの予想の妥当性を調査していきたい．

今回の誤り検出の実験は表1の規則に基づいて行った．表1より定冠詞が付与されている場合の判定は，不可算/複数形を除き全て使用可能である．表3を見ると不可算/複数形，つまりsの余剰が少ないことから，定冠詞が付与されている場合，誤りはほぼ検出不可能なことがわかる．よって，定冠詞の誤りが多くなった本実験では誤り検出の性能が下がったと見られる．

表3．永田らの調査と本実験との冠詞誤りの差違

| | 冠詞誤りの種類 | | | | | | |
|---|---|---|---|---|---|---|---|
| | aの欠落 | theの欠落 | sの欠落 | aの余剰 | theの余剰 | sの余剰 | 選択ミス |
| パターン認識 | 11.5 | 31.1 | 24.6 | 0.0 | 24.6 | 1.6 | 6.6 |
| 材料科学 | 16.9 | 69.6 | 4.5 | 0.0 | 5.6 | 0.0 | 3.4 |
| 永田らの調査 | 19.3 | 2.5 | 43.6 | 11.8 | 7.6 | 7.6 | 7.6 |

（数値は全て%）

### 4.2 冠詞の省略

本実験でシステムが誤りではない用法を誤りと判断した箇所には，冠詞の省略が目立つ結果となった．特にパターン認識分野では，システムが誤りとして検出した56箇所のうち，31箇所が誤りを含まない箇所を誤りと判定した箇所であった．この内17箇所が冠詞の省略によるものであった．

冠詞の省略は可算/不可算判定でも対応可能な場合がある．例えば，Figure 1.という用法をよく目にする．この場合の“figure”は可算名詞である．表1.より可算名詞の単数形が無冠詞で使われることは誤りである．しかし，この場合，1と言う記号によって特定されるため，通例，冠詞を省略する．　冠詞誤り検出がうまくいかなかった要因は大きく分けて４つであった．以下に各要因の詳細について述べる．

“Figure 1.”という用法から可算/不可算判定の規則を学習する場合を考える．この用法では“figure”は単数形/無冠詞で使われている．通常この用法で使われるのは不可算名詞の場合である．よって“figureは数字が後ろに登場する場合は不可算”という規則を学習することが出来る．この規則を用いて“Figure 1. ”と言う表現の誤りを検出すると，“figure”を不可算と判定することで，無冠詞での運用が誤りではないと判断することが出来る．

しかし，実際にはtheが余剰であるThe figure 1.のような表現が論文中に出現する可能性がある．このような場合は，figureを可算/不可算のどちらと判定しても表1からは誤りを検出することが出来ないため，可算/不可算判定以外の規則を作成する必要がある．

### 4.3 規則数の不足

本実験ではパターン認識分野で441箇所中87箇所（35異なり語）が，材料科学分野で399箇所中65箇所（26異なり語）が学習データ中の登場回数不足が原因で規則を作成しておらず，可算/不可算の判定が行えなかった．

また，何らかの規則が作成されている名詞であっても，それらの規則がその名詞の周辺単語と関係のない規則である事も多い．その場合はDefault規則を使うことになり，判定性能が低下してしまう．

### 4.4 チャンキングの失敗

本実験はチャンカーを用いて名詞句と判定された箇所に対して誤り検出を行った．しかし，科学技術英文中の専門用語に対応していない，括弧を含む言い換え表現などがうまくチャンキング出来ない，等の原因から名詞句の判定を間違えてしまうことがあった．

例えば誤り検出対象論文中に，... and document OCR. と言う英文がある．この文に対して誤り検出を行う場合，[document OCR]のように一つの名詞句として考えるべきである．しかし，OCRが単語辞書にない未知語であるため，[document][OCR]のようにそれぞれ名詞句として扱ってしまう．また，the input (binary) imageという文に対しても，[the input (binary) image]と考えるべきである．しかし，括弧が存在することにより，[the input][(binary)][image]と扱ってしまう．

## 5. 改善手法の提案

4.1の検出失敗は，可算/不可算の判定手法のみで対応することは困難である．そこで宮井らの文脈手法[8]を可算/不可算判定の手法に加えて取り入れる．可算/不可算を判定する規則の学習時に，対象名詞に定冠詞が付与されていた場合はその周辺単語を新たな規則として学習する．この規則で可算/不可算を判定するのではなく，定冠詞が付与されるかどうかを判定することで定冠詞による誤り検出失敗を低減できることが考えられる．

4.2で述べた冠詞の省略が起きる場合，表層情報からその理由を容易に推測できる場合が多い．例えば今回の場合は“表題”，“図表”，“手順の説明”，“例示”などがあった．これらの場合は，付近に数字（1. Introduction, Figure 1），括弧つき記号（(a) input image to...），i.e.（two modes, i.e. Japanese mode and English mode）がある，括弧で括られている等の記述の場合が多い．それらに着目し，可算/不可算の判定に加え，“付近に数字が登場する”，“i.e.を含む文である”，“括弧内の文である”等の場合は無冠詞で使用という規則を別途用意する．さらに学習前に(a) 等は手順を表す表現を特殊単語（仮にPROCとする）に置き換え，“PROCが前に登場すれば無冠詞”のような規則も用意する．表1に加えて上記の規則を用いることで，冠詞の省略が起きる場合でも正しく誤り検出が出来るようになる．

4.3の問題を解消するには学習データの規模を大きくすることが考えられる．ただし，ただ闇雲に学習データを追加していくだけでは時間的にも効果的にも得策ではない．そこで学習データを効果的に増やす方法として，今回は対象英文が論文であるということに着目する．論文は通常，参考文献を持つ．そこでその参考文献を学習データとすることを考える．参考文献中には，その論文と記述内容が似ている文献もあり，その場合は同じ単語が同じ用法で使用されている可能性が高い．すなわち，対象英文を添削するときに必要な規則を効率よく学習することが出来る．また，参考文献と言ってもその論文に密接に関わるものから，例えば検定方法を考えるための数学書などあまり関係のないものまで様々である．そこでそれらの各参考文献に登場する単語と対象英文に登場する単語の類似度を計ることでさらなる効果的な学習が行えると考えられる．

4.4の問題にはチャンカーの持つ単語辞書に専門用語を登録するという対処法が考えられる．また括弧の問題については，括弧内の語句は補足表現であることが多いため，チャンカーにかける前に括弧とその内容を削除する方法も考えられる．

参考文献


[1]河合敦夫，杉原厚吉，杉江昇：“英文の誤りを検出するシステム ASPEC-I”，情処学論，vol.25，no.6,pp.1072-1079,Nov.1984.

[2]永田亮，桝井文人，河合敦夫，井須尚紀：“英語名詞の可算/不可算判定手法”，「言語処理」特集号「コーパス言語学・言語教育と言語処理」，pp.3-15，2005.

[3]永田亮，若菜崇宏，河合敦夫，森広浩一郎，桝井文人，井須尚紀:“可算/不可算の判定に基づいた英文の誤り検出”，信学論(D-I)，Vol.J89-D-I NO.8，pp.1777-1790,2006.

[4]若菜崇宏，永田亮，河合敦夫， 桝井文人，井須尚紀：“可算/不可算の判定を用いた英文の誤り検出”，言語処理学会第11回年次大会発表論文集，Mar. 2005.

[5]永田亮，河合敦夫， 桝井文人，井須尚紀：“英語名詞の可算/不可算判定手法”，「言語処理」特集号「コーパス言語学・言語教育と言語処理」，pp.3-15，2005.

[6]永田亮，井口達也，脇寺健太，桝井文人，河合敦夫：“日本人英語学習者のための冠詞誤り検出”，信学論(D-I)，vol.J87-D-I，no.1，pp.60-68，Jan. 2004.

[7]和泉絵美，斎賀豊美，T. Supnithi，内本清貴，井佐原均：“エラータグ付き日本人英語学習者発話コーパスを用いた学習者の冠詞習得傾向の分析”，言語処理学会第9回年次大会発表論文集，pp.19-22，Mar. 2003.

[8]宮井俊也，永田亮，河合敦夫，桝井文人，井須尚紀：“文脈情報とイディオムを考慮した英文の自動冠詞付与手法”，言語処理学会第13回年次大会，PA3-5，2007