Vol.2011-NL-204 No.10
2011/11/22

英文への自動冠詞付与における前方照応の考慮

竹内裕己[†1]　河合敦夫[†2]
永田亮[†3]　乙武北斗[†4]

英文中において，既出の名詞句が 2 回目以降出現，かつ同一の内容を表すと認められる場合，その名詞句には定冠詞 the が付与される．このことを我々は前方照応による the と呼び，これを考慮した英文への自動冠詞付与を目指している．本稿では，前方照応による the を考慮した自動冠詞付与を行うために必要な要因分析について述べる．そのために，最大エントロピー分類器を用いた疑似的な前方照応の判定モデルを作成し，性能評価実験を行った．また，その判定結果を従来の自動冠詞付与モデルで利用する手法を提案し，効果の分析を行った．

Usage of Coreference in Automatic Generation of Articles in English Composition

Hiromi Takeuchi,[†1] Atsuo Kawai,[†2] Ryo Nagata[†3]
and Hokuto Ototake[†4]

Typical usage of an article “ the ”is anaphoric: an entity in question has been referred to previously in a same document. We try to incorporate this anaphoric reference to a conventional article generation system. In order to achieve this precisely, which noun phrases are co-referents must be analyzed based on deep understanding of a document. However a computer can’t understand the document content; we constructed a judging model of pseudo co-reference by using Maximum Entropy method and evaluated the effect of the pseudo co-reference model on the article generation performance.

1. はじめに

近年，英語非母国語話者による英文執筆の機会が増加しているが，それにはしばしば誤りが含まれる．特に，日本語のように冠詞の概念がない言語話者においては冠詞の誤用が多く報告されている[1]．

このような冠詞の誤用を人手に頼らず機械的に訂正するために，様々な自動冠詞付与手法が提案されている[2)3)4)]．しかしそれらの手法においては，不定冠詞 a と無冠詞の付与精度には優れているが、定冠詞 the の付与精度はそれらと比較して低いことが報告されている[5]．その大きな要因の 1 つとして，前方照応による the の付与の考慮を行っていないことが挙げられる．前方照応による the とは，その文章中で既出の名詞句に対しては the を付与するという文法的な規則である．冠詞付与に対して前方照応の考慮を正確に行うためには，既出かどうかの判定を行う必要があり，それには複数文に渡る照応解析を必要とするため単純には利用できない[6]．さらに，従来の冠詞付与手法では，冠詞決定の情報として，付与対象名詞句周辺の文内情報のみを用いるため，複数文に渡る前方照応による the の可能性は考慮されていない．しかしながら，我々の自動冠詞付与の目的は冠詞付与誤りの訂正であるため，一般的な文法規則である前方照応は考慮すべき事項である．また，前方照応よる the の判定が完全にできない場合でも，その可能性を提示することができれば，英語非母国語話者にとって大きなメリットであると考えられる．

このような問題を解決するために，本稿では前方照応を考慮した定冠詞の付与手法を提案する．ただ，前述の通り，前方照応を厳密に行うことは難しいため，提案手法では前方照応の解析を疑似的に行う．前方照応を疑似的に行うことにより，the を付与できる可能性の高い名詞句を判定するモデルの作成を行う．さらに，その判定結果を利用することにより，従来の文内情報による冠詞付与手法が受ける影響について考察する．従来の冠詞付与手法としては，比較的付与精度が高く，疑似的な前方照応の考慮が行いやすい最大エントロピー分類

†1 三重大学大学院工学研究科
Graduate School of Engineering, Mie University
†2 三重大学工学部
Faculty of Engineering, Mie University
†3 甲南大学知能情報学部
Faculty of Intelligence and Informations, Konan University
†4 福岡大学工学部
Faculty of Engineering, Fukuoka University

器による手法[3] を用いる．本稿の目的は，疑似的に前方照応を扱うことにより，従来の冠詞付与手法に対し，前方照応による the の考慮を行うために必要な要因を明らかにすることである．

以下，2 章では，前方照応判定を行う上での問題点と疑似的な前方照応判定モデルについて，3 章では，比較実験に用いる文内情報のみを用いた冠詞付与システムについて述べる．4 章では，性能評価実験について，5 章で実験結果の結果と考察，最後に 6 章でまとめを述べる．

2. 疑似的な前方照応判定モデル

本章では，前方照応を機械的に判定する上での問題点を挙げる．更に，それを考慮し作成した疑似的な前方照応判定モデルについて述べる．

2.1 前方照応判定の問題点

前方照応が起きるのは，**図 1** に示すように冠詞付与対象の名詞句に対して，同じ実体を指す名詞句が存在する場合である．前者を照応詞，後者を先行詞を呼ぶ．このような前方照応の判定および，それによる冠詞付与を機械的に行うための問題点を以下に示す．

(1) 先行詞と照応詞の関係には複数のパターンが存在

先行詞と照応詞の字面が一致する場合や，照応詞が一部省略，言い換え表現などで出現する場合などがある．特に言い換え表現は多数存在するため，すべてを網羅することは難しい．この他にもパターンが存在するが，今回考慮したパターンについては次節で詳しく述べる．

(2) 異なる実体を指す場合は非照応

(1) のパターンにマッチした場合でも，名詞句同士が異なる実体を指す場合は照応しない．例えば，図 1 の「company」がそれぞれが別の実体 (会社) を指している場合は非照応となる．照応，非照応の判断には深い文脈理解を要するものもあるため，機械的にすべてを判定することは非常に困難である．

(3) 名詞句によって有効な照応のパターンが異なる

どのパターンで出現すると照応しやすいかなどは，名詞句ごとに性質が異なる．したがって，すべての名詞句に対し一律の処理をすべきではない．

(4) 機械学習に使用するコーパスの問題

冠詞が the に決定される要因は前方照応の他にも文内情報など複数存在する．またそれらは排他的ではないため，どの要因により the になっているかは一意に決まらない．よって，機械学習時に正解を与えるには照応情報付きの学習コーパスが必要となる．しかし，MUC7[⋆1] などの照応情報付きのコーパスは存在するが，人手で作成しているため量が少なく機械学習には不十分である．さらに，照応情報の多くは代名詞等の冠詞とは関係のないものとなっているため本研究には適していない．

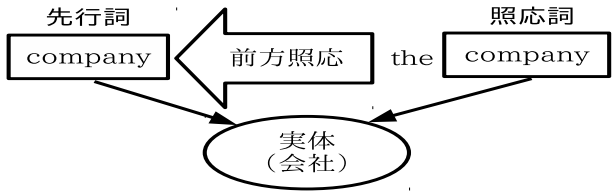

図 1 前方照応

2.2 疑似的な前方照応判定モデル

前節で述べた前方照応判定の問題点を踏まえた上で，最大エントロピー分類器による疑似的な前方照応判定モデルを提案する．この判定モデルでは，最大エントロピー分類器の学習時に与える正解として，前方照応かどうかではなく，判定対象名詞句の冠詞が the かどうかを使用する．よって，冠詞が the であれば照応，the 以外 (a,an, 無冠詞) であれば非照応とする．このことから，前方照応ではなく，疑似的な前方照応の判定モデルとした．本稿では 1 章で述べた通り，本来複雑な問題である前方照応の考慮を疑似的に利用し，冠詞付与を行うための要因を明らかにすることが目的である．したがって，前方照応と判定する条件を細かく設定せず，照応の可能性があるものをより多く考慮するような判定モデルの作成を行った．

2.2.1 使 用 素 性

使用した素性を**図 2** に示す．図 2 において，最も右の列の要素は素性値を表しており，例文の該当要素が入っている．素性値より左の要素は素性名およびカテゴリ名，素性を識別するための素性記号である．照応条件カテゴリに属する素性の素性値は，該当要素が存在するか，しないかの 2 値で表す．図 2 の例文の場合，素性値 co-I により素性名の言い換えが存在することを表し，素性値が“ - ”となっているものは該当要素が存在しないことを表す．素性値 co-I の co は照応条件である識別子，I は素性記号を表す．基本的に一致するかどうかの判定は文字列一致で行うが，一致の場合により素性を分けている．素性名に含まれる単

⋆1 http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_toc.html

Vol.2011-NL-204 No.10
2011/11/22

名詞とは，名詞１語からなる名詞句を表す．また，それぞれの素性についての補足事項は以下の通りである．

- 本稿では実験に新聞記事を用いるため，照応判定の適用範囲は，同一の新聞記事内とする． また，先行詞候補と対象名詞句の距離の差は考慮しない．先行詞候補が 2 文前でも 4 文前でも同じ素性値とする．また，適用範囲内において複数の照応条件とマッチした場合は，すべて素性に加える．但し，同じ条件に 2 回以上マッチした場合は．1 回目のみ素性に加える．
- 素性記号 C では対象名詞句が主語，補語，目的語，その他のいずれに属するかの判定を行う．その判定には、対象名詞句前後が前置詞か動詞かという簡易的な情報を用いる．
- 素性記号 D〜H はそれぞれ対象名詞句と先行詞候補が一致しているかの条件を表す．2.1 節でも述べたように，一致のパターンは複数存在し，名詞句によって有効なパターンが異なると考えられるので，一致のパターンにより素性を分ける．
- 素性記号 D〜H それぞれにおける一致条件の具体例を**図 3** に示す．ここで，素性記号ＦとＧの（z）は 1 単語以上からなる単語列の意味で，3 単語以上からなる名詞句にも一 致することを表す．
- 図 2 において，素性記号 D〜G は，単純に文字列として一致する単複一致の場合に加え，文字列としては一致しない単複不一致の場合にも素性を設けている．この理由は，先行詞候補が複数形の名詞句で，対象名詞句が同じ名詞句の単数形だった場合，複数あるうちの１つとして対象名詞句が出現することを考慮するためである．よって単複不一致は，原形が同じ名詞句で，先行詞候補が複数形，対象名詞句が単数形の場合のみしか適用しない．
- 素性記号 I の言い換え表現については，その規則を人手によって作成した．今回は，実験に使用したコーパス中に頻出する company と companies の 2 語どちらかを含む名詞句のみに適用する．その言い換え表現として，会社名に使われる Co.，Inc.，Corp. の内どれかを含む名詞句を言い換えとした．図 2 の例文では，Mie Dai Kotsu Co., Ltd. が bus company に言い換えられたものと判定している．
- 素性記号 D〜I の素性は照応条件であり，これらに 1 つでもマッチする名詞句のみを照応判定の対象とし，マッチしない名詞句は判定対象外とした．

素性記号	カテゴリ	素性名	素性値
A	名詞句情報	名詞句	bus company
B		主名詞	company
C		名詞句の役割	主語
D	照応条件	単名詞ー致(単複ー致)	−
E		単名詞ー致(単複不ー致)	−
F		名詞句部分ー致(単複ー致)	−
G		名詞句部分ー致(単複不ー致)	−
H		先行詞の主名詞以外にー致	−
I		言い換え	co_I

図 2 前方照応素性リストと例

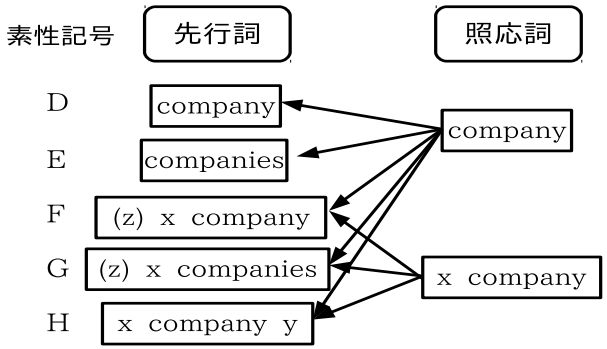

図 3 前方照応の一致条件

(例文) ···. This is the good company's guide of the random walks.

カテゴリ	素性名	素性値
対象名詞句	主名詞	guide
	主名詞POS	NN
	名詞句部分文字列	good company's guide company's guide
	主名詞以外の名詞	company
	修飾語	good
	修飾語POS	JJ
	所有格を含む	yes
名詞句の前	文章の先頭	no
	句の種類	VP
	名詞句直前の語	be
	上のPOS	VBZ
	イディオム	—
名詞句の後	句の分類	PP
	名詞句直後の語	of
	上のPOS	IN
修飾句	句の分類	NP
	主名詞	walks
	主名詞以外の名詞	—
	修飾語	random
	修飾語POS	JJ

図 4 文内情報素性リストと例

3. 文内情報による定冠詞 the の付与モデル

2.2 節で説明した疑似的な前方照応判定モデルの効果を検証するために，文内情報のみを用いる定冠詞 the の付与モデルを作成した．前方照応は定冠詞 the の決定のみに影響し，冠詞が a，an，無冠詞の判定には影響しない．したがって，本付与モデルでは the と the 以外 (a,an, 無冠詞) の 2 値での判定行う．以後本稿では，冠詞付与と表記した場合，特にことわりがなければ，the と the 以外 (a,an, 無冠詞) どちらかの付与を意味する．

使用した素性は**図 4** に示す通りであり，これらは乙武ら[5] が用いた無冠詞判定用の素性に the の決定に影響を及ぼすと考えられる素性を新たに追加したものである．図 4 において，最も右の列の要素は素性値を表しており，例文の該当要素が入っている．素性値より左の要素は素性名および素性の分類を表すカテゴリ名である．図 4 中のカテゴリが名詞句の前である素性名イディオムには，独自の定義を行った．本稿の定義としては，対象名詞句の前 2 単語が，1 語からなる名詞と前置詞である場合をイディオムとして扱う．例えば，number of や terms of などが挙げられ，これらの後に続く名詞句は冠詞は無冠詞になりやすいなど，冠詞の用法が通常とは異なる場合が多いため，セットで 1 つの素性として考慮した．

4. 実　　験

本章では，以下の 3 つのモデルについての評価実験について述べる．

照応判定モデル： 2 章で述べた疑似的な前方照応判定モデル
モデル 1　　： 3 章で述べた文内情報による定冠詞 the の付与モデル
モデル 2　　： 前方照応の考慮を追加した定冠詞 the の付与モデル

モデル 2 は**図 5** に示すように，モデル 1 に対し，照応判定モデルにおける照応，非照応の 2 値の判定結果を追加素性として使用することにより，前方照応の考慮を行う．但し，照応判定モデルにおいて，2.2 節で述べたような照応条件にマッチせず判定対象外となる名詞句は，前方照応の素性値にはなにも入らない．

4.1 実験データ

本実験では，全モデルの評価実験において Reuters Corpus[7] を用いた．Reuters Corpus 中には 175,249 の冠詞付与対象箇所が含まれる．冠詞付与対象箇所とは，コーパス中で冠詞が {a，an，the，無冠詞} のいずれかの名詞句である．素性抽出のために品詞タグ付けを行

Vol.2011-NL-204 No.10
2011/11/22

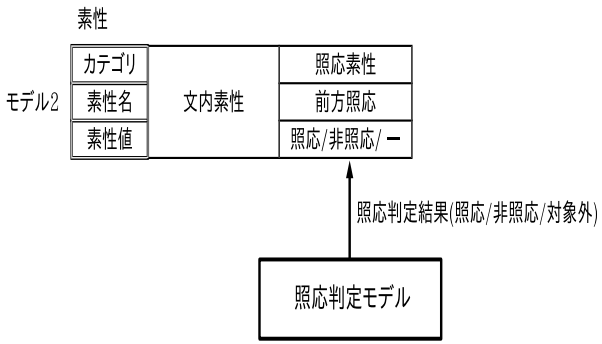

図 5 モデル 2 の素性

うツールとして OAK System[⋆1] を用いた．

最大エントロピー分類器は，機械学習アルゴリズムの実装の一つである Classias[⋆2] の L2 正則化ロジスティック回帰モデルを用いた．また，使用した Reuters Corpus 内の冠詞分布状況を**表 1** に示す．本実験では，a，an，無冠詞の判別は行わないため同じ種類としてカウントした．

4.2 実験手順

本実験では，作成した 3 つのモデルを用いて以下の 2 つの実験を行った．

実験 1： 照応判定モデルの精度評価

実験 2： モデル 1，2 の比較

実験 1 では，まず照応判定モデルにおいて全体での精度評価を行う．さらに，名詞句の出現頻度や冠詞生起確率によって前方照応判定精度が異なると予想されるため，名詞句の分類を**図 6** のように定義するカテゴリ A〜B それぞれに対しての精度評価も行う．また，図 6 中の出現頻度や冠詞生起確率は照応判定を行った名詞句を対象に算出する．

実験 2 では，モデル 1，2 の精度評価を行い比較する．ここで，最大エントロピー分類器の判定結果を信用するかどうかを決定する指標として，classias が出力するスコアの閾値 ($\theta \geqq 0$) を考える．classias が出力するスコアと閾値 θ の関係を**図 7** に示す．判定結果のスコアの絶対値が θ よりも小さかった場合，判定は行わないとものとする．

⋆1 http://nlp.cs.nyu.edu/oak/
⋆2 http://www.chokkan.org/software/classias/

表 1 Reuters Corpus 中の冠詞分布状況

冠詞	分布
the	28.6%
a, an, 無冠詞	73.4%

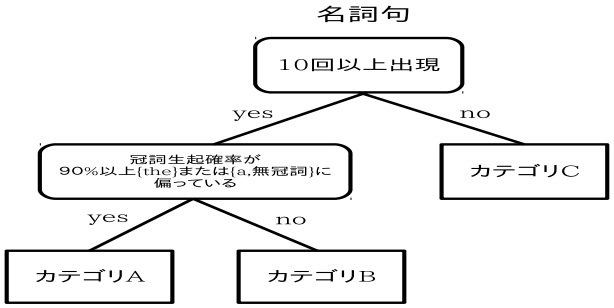

図 6 名詞句分類決定木

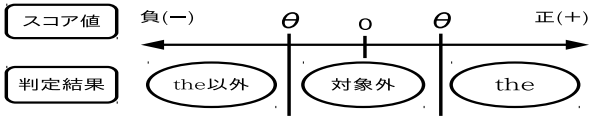

図 7 スコアと閾値 θ の関係

4.3 評価指標

実験 1 の評価尺度として，Recall-coref (R_c) と，Precision-coref (P_c) を用いた．これらの評価尺度は式 (1)，(2) で定義される．式 (1)，(2) における正しく照応判定を行った数には，学習時と同様に疑似的な照応判定の正解を用いた．すなわち，対象名詞句の冠詞が the であれば照応，the 以外 (a,an, 無冠詞) であれば非照応が正解となる．

$$R_c = \frac{\text{正しく照応判定を行った数}}{\text{冠詞付与対象箇所の総数}} \tag{1}$$

$$P_c = \frac{\text{正しく照応判定を行った数}}{\text{照応判定を行った数}} \tag{2}$$

実験 2 では，モデル 1，2 の評価尺度として，Recall-article (R_a) と，Precision-article (P_a) を用いた．これらの評価尺度は式 (3)，(4) で定義される．

Vol.2011-NL-204 No.10
2011/11/22

$$R_a = \frac{\text{正しく冠詞を提示した数}}{\text{冠詞付与対象箇所の総数}} \quad (3)$$

$$P_a = \frac{\text{正しく冠詞を提示した数}}{\text{冠詞を提示した数}} \quad (4)$$

これらの評価尺度に基づき，10 分割交差法によりトレーニングデータとテストデータの分割を行った．よって，最終的な数値は 10 回の平均値を示す．

5. 結果・考察

実験 1 の結果を**表 2** に示す．表 2 には，照応判定全体とカテゴリごとの結果に加え，名詞句カテゴリごとの分布も示した．照応判定全体での Recall-coref 値が 0.1633 と低い理由としては，2.2 節で定義した照応条件にマッチする名詞句数に依存していることによる．今後照応条件にマッチする名詞句数を増やすためには，言い換え表現を増やす必要があると考えられる．今回は言い換え表現による照応を用いたのは company と companies のみである．これらの名詞句が図 2 の照応条件にマッチした総数のうち，言い換え表現にのみマッチした数の割合を**表 3** に示す．表 3 から，6 割以上が言い換え表現によってのみ照応条件にマッチしていることがわかる．よって，今後照応判定の適用数を増やすためには，言い換え表現を増やす必要がある．

また，図 2 の各名詞句カテゴリごとにおける Precision-coref と，Recall-coref をみると，冠詞生起確率に偏りが大きいカテゴリ A の名詞句が，両値ともに最も高い値を示しており，全体の精度に大きく影響ことがわかる．しかし，冠詞生起確率に偏りの小さいカテゴリ B では，他のカテゴリに比べ分布が小さいものの，Precision-coref がかなり低い値となっている．このことから，作成した照応判定モデルの照応条件のみでは十分に対応しきれていないと考えられる．

実験 2 において，モデル 1 とモデル 2 の判定精度の比較結果を**図 8** に示す．グラフの横軸は，最大エントロピー分類器におけるスコア値の閾値 θ を表している．図 8 より，Recall-article に関しては θ の上昇とともにモデル 1 と 2 の差が広がっていることがわかる．また，Precision-article は，$\theta = 0$ のときに 0.39% 程度モデル 2 が高い値を示しているものの，それ以降は現状維持に留まった．この要因としては，表 2 のカテゴリ A や C の高い精度の判定結果は，ほぼモデル 1 でも既に正しく判定されているということが考えられる．このことにより，既に正しい判定でも，スコア値が低い判定に対しては，スコア値の底上げにつながったため，Recall-article 値のみが上昇したと考えられる．よって今後 Precision-article の改善には，カテゴリ B、C のような，より細かな照応条件が必要となる名詞句への対応が必要である．

表 2 全体と名詞句カテゴリ別の実験結果

名詞句カテゴリ	分布	Recall-coref	Precision-coref
照応判定全体	-	16.33 %	93.12 %
A	55.57 %	52.29 %	94.1 %
B	6.41 %	4.04 %	62.91 %
C	38.02 %	31.56 %	83.01%

表 3 言い換え表現でのみ照応条件にマッチする割合

名詞句	言い換え表現のみの割合
company	67%
companies	70%

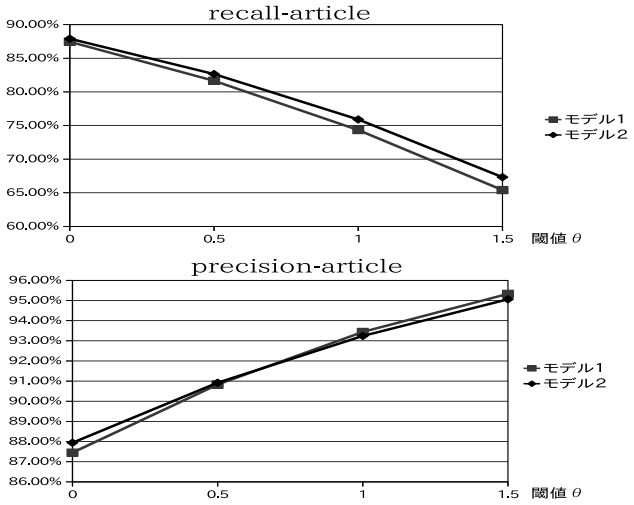

図 8 モデル 1 とモデル 2 の比較実験

6. まとめと今後の課題

本稿では，従来の最大エントロピー分類器による文内情報のみの冠詞付与手法に対して前方照応の考慮を試みた．前方照応を考慮する手法として，疑似的な前方照応判定モデルを作成し，その判定結果を従来の付与モデルで利用することを提案した．そして，疑似的な前方照応判定モデルを利用した場合の冠詞付与モデルと，利用しない場合の冠詞付与モデルを比較し，その影響の検証を行った．実験結果としては，冠詞付与に高い信頼性を要する場合において，前方照応判定モデルを用いない場合に比べ，Precision-article を維持したまま，Recall-article が改善した．

今後は，照応条件をさらに細かくし，冠詞生起確率に偏りが少ない名詞句への対応が必要である．よって，センタリング理論など文同士の意味的つながりなども考慮に入れ，照応の判定精度向上を目指す．さらに，今回 company と companies のみ考慮した言い換え表現についても，実験結果で述べたように照応判定の対象となる名詞句数の増加に有効であるため拡張が必要である．その手法として，WordNet[8] 等でシソーラスを利用することにより自動的な言い換え表現の抽出を行いたいと考えている．また，前方照応判定モデルの評価を the かその他かという疑似的な正解で行っていたが，今後は人手で照応判定を行ったテスト用コーパスを作成し，それによる評価も行いたいと考えている．

今回作成した疑似的な前方照応判定モデルでは，基本的に人手で見つけやすい照応条件を使用している．しかし一見，前方照応による the が付与できそうな場合でも，全く the が付与されない場合が多数存在した．これらは，名詞句自体の性質であったり，イディオム表現の一部であることが考えられる．よってこれらの判定に特化することにより，英語学習者が前方照応により the を付与できると判定した中から，実際には the を付与出来ない場合を提示する補助的ツールとしての利用も考えている．

参 考 文 献

1) R. D. Felice and S. G. Pulman, : A classifier-based approach to preposition and determiner error correction in L2 English, , *Proc. 22nd International Conference on Computational Linguistics*, pp.169–176, Manchester, UK (2008).
2) A. Kawai,K. Sugihara,N. Sugie, : ASPEC-I : An error detection system for English composition, Proc. IPSJ Journal (in Japanese) , vol. 25, no. 6, pp. 1072-1079 (Nov.1984)
3) Han, N. Chodorow, M. and Claudia, L., : Detecting errors in English article usage with a maximum entropy classifier trained on a large, diverse corpus, Proceedings of the 4th international conference on language resources and evaluation (2004)
4) R. Nagata, T. Wakana, F. Masui, A. Kawai , and N. Isu, : Detecting article errors based on the mass count distinction, Lecture Notes in Artificial Intelligence, Dale, R., Wong, K.F., Su, J., Kwong, O.Y. (Eds.), Springer-Verlag, pp. 815-826 (Oct. 2005)
5) 乙武北斗，荒木健治，吉村賢治：自動獲得されるルールに基づく英文冠詞誤り校正手法における最大エントロピー分類器の利用，言語処理学会　第 17 回年次大会 （2011）
6) 樋口昌幸：英語の冠詞 -その使い方の原理を探る- ，開拓社， 東京 （2009）
7) D. Lewis, : Reuters-21578 text categorization test collection (1997)
8) C. Fellbaum, : WordNet : An Electronic Lexical database, The MIT Press (1998)