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

宮井 俊也†，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
E-mail: †{miyai, kawai, masui, isu}@ai.info.mie-u.ac.jp　‡rnagata@hyogo-u.ac.jp

１．はじめに

日本人英語学習者にとって，冠詞の用法は最も誤りを犯しやすい文法項目の一つである．冠詞の用法には厳密なルールが無い場合が多く，冠詞を正しく使用するためには，辞書や多くの用例を調べることが必要となる．特に専門用語の場合，辞書・用例共に少なく，正しい冠詞の選択が困難となる．一方で，英語論文では，文脈を明確にするために，冠詞を正しく使用することが重要となる[1].

この問題を解決するために，冠詞を自動付与する手法が提案されている．井口ら[2]は，文中において対象名詞の出現が何度目となるか，名詞を修飾する形容詞や前後の前置詞から限定的となりやすいかなどの情報に基づいて冠詞の生起確率を推定し，冠詞付与を行う手法を提案している（以降，この手法を「文脈手法」と呼ぶ）．また，Lee[3]や Han ら[4]は，冠詞の前後の単語や品詞情報などの情報に基づいて冠詞の生起確率を推定し，冠詞付与する手法を提案している．さらに，Nagata ら[5]は，冠詞を中心とした単語列をイディオムとしてコーパスから抽出し，そのイディオムを利用して冠詞を付与する手法を提案している（以降，この手法を「イディオム手法」と呼ぶ）．文脈手法[2]とイディオム手法[5]は，それぞれ付与できる冠詞の傾向が異なるため，両者を組み合わせることで付与性能を向上させる手法[6]も提案されている．手法[6]では，両手法でそれぞれ得られる冠詞生起確率を重み付きで混合する．しかしながら，それぞれの手法で確率が最大となる付与規則しか用いないため，冠詞ごとに重み付きで混合する一方で，付与する冠詞が異なる場合や，片方の手法で付与規則が見つからない場合には，混合に用いる値を 0 として計算してしまう．よって，この場合には，統合手法が逆効果となってしまう．

本論文では，この問題点に対応するため，用いる付与規則を冠詞ごとに精度が最大となる規則に拡張し，得られた冠詞生起確率を重み付きで混合する．また，付与規則が無い場合にも，最も分散した状態である値（0.33）で計算を行う．また，パラメータ設定の簡単化のため，組み合わせに必要となる重みを自動設定する手法を提案する．この提案手法を，科学技術英文（計算機科学，材料科学）に適用し，付与性能を評価する．

以下，2. で文脈手法とイディオム手法について説明する．3. で両手法を効果的に統合する手法とその際に用いる重みを自動設定する手法を提案する．4. で評価実験について述べる．5. で実験結果を考察する．

２．文脈手法とイディオム手法

文脈手法では，コーパスから，「a」「the」「φ（無冠詞）」3 つの冠詞について，文脈情報を基に冠詞生起確率を推定し，冠詞の付与規則として利用する．ここでの文脈情報とは，「文中において対象名詞の出現が何度目となるか」，「名詞を修飾する形容詞」，「前後にある前置詞」を指す．図 1 に，文脈手法での付与規則の取得例を示す（取得対象名詞”… ultimate goal of …”の「goal」）．例えば，名詞「goal」が文中で初めて出現する場合において，無冠詞 ϕ となる割合を求めることで，名詞「goal」の初出時における無冠詞 ϕ の生起確率が推定できる．名詞「goal」が既出の場合，直前に出現した「goal」に使用されていた冠詞を文脈情報として用い，生起確率の推定を行う．形容詞「ultimate」が修飾する場合や，前置詞「of」が後に接続する場合についても，同様に推定できる．また，「a」「the」についても同様に，コーパスから生起確率を推定する．推定した確率とその文脈情報を付与規則として辞書に登録し，冠詞付与に利用する．ただし，冠詞全体の頻度が θ_f 未満のものは登録しない．

取得対象（goal）…ultimate goal of …

$$\text{文中で初出の名詞「goal」の冠詞生起確率}(\phi) = \frac{\text{文中で初出の場合の名詞「goal」が無冠詞となる出現頻度}}{\text{文中で初出の場合の名詞「goal」の出現頻度}}$$

$$\text{直前では無冠詞}\phi\text{が使われた既出の名詞「goal」の冠詞生起確率}(\phi) = \frac{\text{名詞「goal」が直前では無冠詞}\phi\text{となり今も無冠詞}\phi\text{となる出現頻度}}{\text{直前の名詞「goal」では無冠詞}\phi\text{となる名詞「goal」の出現頻度}}$$

$$\text{形容詞「ultimate」が修飾する場合の冠詞生起確率}(\phi) = \frac{\text{形容詞「ultimate」に修飾される場合で無冠詞となる出現頻度}}{\text{形容詞「ultimate」に修飾される名詞の出現頻度}}$$

$$\text{前置詞「of」が後に接続する場合の冠詞生起確率}(\phi) = \frac{\text{名詞「goal」の後に前置詞「of」が接続する場合に無冠詞となる出現頻度}}{\text{名詞「goal」の後に前置詞「of」が接続する場合の出現頻度}}$$

（a, theの場合も同様に取得し，分母の出現頻度がθ_f未満の場合には辞書に登録しない）

図1：文脈手法での冠詞生起確率の取得例

イディオム手法では，冠詞を中心とした単語列のうち，頻出し，かつ冠詞生起確率の偏ったものをイディオムとし，そのイディオムごとに冠詞生起確率を推定する．図２に，イディオム手法の例を示す．図２において，単語列「is altered as 〈a/the/φ のいずれか〉」がコーパス中に多く出現する場合，冠詞が「a」となる割合を求めることで，不定冠詞の生起確率を推定できる．この値が他の２つ（the，φ）に比べ大きい場合，イディオムとする．同様に，他の全ての単語列についても，出現数が閾値 θ_f 以上のものをイディオムとして抽出する．抽出したイディ

オムを付与規則として辞書に登録する．

冠詞の付与は，両手法ともに，付与箇所にあてはまる規則のうち，最も高い生起確率となる付与規則で行う．冠詞付与できる生起確率かの判定には閾値θを用いる．冠詞付与の際に，規則が見つからない，あるいは全ての規則の生起確率が閾値θ未満の場合には付与を行わない．

… function is altered as **a** result of **the** gravitational …

「is altered as〈冠詞〉」での冠詞生起確率(a) ＝ 単語列「is altered as **a**」の出現頻度 / 単語列「is altered as〈冠詞〉」の出現頻度

「is altered as〈冠詞〉result」での冠詞生起確率(a) ＝ 単語列「is altered as **a** result」の出現頻度 / 「is altered as〈冠詞〉result」の出現頻度

「of〈冠詞〉gravitational」での冠詞生起確率(the) ＝ 「of **the** gravitational」の出現頻度 / 「of〈冠詞〉gravitational」の出現頻度

(〈冠詞〉は(a/the/ϕ)のいずれかで，冠詞全体の出現頻度がθf未満の場合は辞書に登録しない)

図2：イディオム手法での冠詞生起確率の取得例

3．提案手法

提案手法では，文脈手法とイディオム手法を組み合わせて冠詞付与を行う．新たな冠詞生起確率として，両手法それぞれで得られる生起確率を冠詞（a/the/ϕ）ごとに重み付きで混合する．新たな冠詞生起確率の導出法を図３に示す．より効果的に両手法を組み合わせるために重み α を導入する．αは０から１の間の値をとり，０に近いほど文脈手法が，１に近いほどイディオム手法が優先される．どちらかの手法で付与規則が辞書に登録されていない場合は，その生起確率は全て0.33（約三分の一）として計算する．加えて，この重みαを様々な値に変化させることで，最適な重みを調査する．

新たな冠詞生起確率(ϕ) ＝ (1−α)×文脈手法で得た冠詞生起確率(ϕ)の最大値 ＋α×イディオム手法で得た冠詞生起確率(ϕ)の最大値

αには0〜1の値が入る
冠詞生起確率が得られなかった場合は0.33とする
a / the についても同様に求め，最も値の大きい冠詞を付与する

重みαの自動設定：

$$\text{重みα} = \frac{Ni}{Nc+Ni}$$

Ni：最大となった冠詞付与規則の学習コーパス中での出現頻度（イディオム手法）
Nc：最大となった冠詞付与規則の学習コーパス中での出現頻度（文脈手法）

図3：統合手法での冠詞生起確率の導出法と重みαの自動設定方法

また，冠詞付与を行う文章が，文脈手法とイディオム手法のどちらを優先すべきか判断しにくい場合も考えられる．そこで，重みαを自動で設定する手法を提案する．両手法から得られた規則の使われやすさで優先すべき手法を推定できる．そこで，学習コーパス中での登場回数，つまり冠詞生起確率の推定に用いた出現頻度（図１，図２の式の分子部分）の比を求めて重みαとする．この場合，冠詞ごとに最大となった付与規則によってαが決まるので，付与対象箇所，冠詞ごとにその値は異なる．

図４で，実際に名詞「extraction」に冠詞を付与する場合の例を示す．両手法でそれぞれ取得できる付与規則のうち，生起確率が最大となるものを抽出し，それらを冠詞（a/the/ϕ）ごとに重み付きで混合する．「a」の場合，文脈手法では直前の「extraction」が無冠詞となるときに2.0％，イディオム手法では”〈a〉feature”が9.3%と最大となるため，この値を用いて新たな冠詞生起確率を算出する．同様に，「the」では文脈手法，イディオム手法それぞれから28.6%，64.7%を，「ϕ」では78.3%，65.4%を用いる．

αを自動設定する場合には，それぞれで用いられる冠詞生起確率の推定に用いた頻度から重みαを算出する．図４の場合，「a」では，文脈手法の規則の頻度が2，イディオム手法での頻度が95より，α＝95／（2＋95）＝0.98となる．同様に「the」では，両手法それぞれの頻度が4，75から，αは0.95となる．「ϕ」では，それぞれ115，68から，α＝ 0.37となる．よって，「a」の新たな生起確率は，$0.02 \times 2.0 + 0.98 \times 9.3 = 9.2$ となる．
同様に，「the」は，$0.05 \times 28.6 + 0.95 \times 64.7 = 62.9$ となり，「ϕ」では，73.5となる．したがって，この場合には無冠詞「ϕ」が付与される．

… The main trend in〈冠詞〉feature extraction has been …

付与規則：文脈手法	生起確率(頻度)
1 〈ϕ〉extraction 直前の「extraction」には冠詞ϕ	**78.3**% (**115**/147)
2 〈ϕ〉extraction 前に前置詞「in」	71.4% (9/14)
3 〈the〉extraction 前に前置詞「in」	**28.6**% (**4**/14)
…	
5 〈a〉extraction 直前の「extraction」には冠詞ϕ	**2.0**% (**2**/147)

付与規則：イディオム手法	
1 〈ϕ〉feature extraction	**65.4**% (**68**/104)
2 in 〈the〉feature	**64.7**% (**75**/116)
3 in 〈the〉	47.3% (6518/13782)
…	
9 〈a〉feature	**9.3**% (**95**/1022)

重みα=0.6の場合：

冠詞生起確率(a)	=	0.4	×2.0 ＋	0.6	×9.3 =	6.4
冠詞生起確率(the)	=	0.4	×28.6 ＋	0.6	×64.7 =	50.3
冠詞生起確率(ϕ)	=	0.4	×78.3 ＋	0.6	×65.4 =	70.6

重みαを自動設定する場合：

冠詞生起確率(a)	=	$\frac{2}{2+95}$	×2.0 ＋	$\frac{95}{2+95}$	×9.3 =	9.2
冠詞生起確率(the)	=	$\frac{4}{4+75}$	×28.6 ＋	$\frac{75}{4+75}$	×64.7 =	62.9
冠詞生起確率(ϕ)	=	$\frac{115}{115+68}$	×78.3 ＋	$\frac{68}{115+68}$	×65.4 =	73.5

図4：提案した統合手法での冠詞生起確率の取得例

4．評価実験

4．1 実験条件

評価対象として，材料科学（セラミック），計算機科学（人工知能）の２つの分野を選んだ．それぞれの分野から１つの論文誌（材料：Journal of Non-crystalline Solids，計算機：Pattern Recognition）を選択し，200論文の英文（材料：約60万語，計算機：約90万語）から前述の２つの手法で生起確率辞書を取得した．また，評価用として別に取得した各分野10論文（冠詞付与箇所は，材料：6856箇所，計算機：10127箇所）を対象に冠詞付与を行った．コーパスとなる英文は，インパクトファクター値（ある論文誌ごとの，１論文あたりに引用される回数の平均値）を参考に分野内で値の大きい論文誌を選択した．引用されることの多い論文誌であるため，冠詞

の用法などの文法的誤りは比較的少ないと考えられる．インパクトファクター値は，Journal Citation Reports（JCR）[7]が提供しているものを用いた．

性能の評価のために，入力文中にある，冠詞付与箇所のうち，正しい冠詞を付与した割合を示す冠詞付与率（Recall），付与できた冠詞のうち，正しく付与できた割合を示す冠詞付与制度（Precision），その両方を考慮した総合的な性能値である F 値の 3 種類の尺度を用いた．また，辞書に登録する規則の頻度の閾値 θ_f は 10 とした．付与に用いる冠詞生起確率の閾値 θ は，0，0.4，0.7，0.9 と変化させた．提案手法で用いる重み α の値は，0 から 1 までを 0.1 刻みで変化させたものと，自動設定させる手法を用いた．また，α=0 では文脈手法のみでの冠詞付与を，α=1 ではイディオム手法のみでの冠詞付与を示すため．これらの値をベースラインとした．

4．2 実験結果

冠詞付与率（Recall）を材料科学，計算機科学それぞれ図 5，図 6 に示す．また，自動設定した場合の重み α は AUTO としている．Recall は．生起確率の閾値が低い場合（θ≦0.4）には，重み α=0.6 付近とややイディオム手法に重みを持たせる場合にわずかに高い値を示す．閾値が高くなるにつれ，イディオム手法のみを用いるほうが，値が高くなった．また，α を自動設定とした場合では，どの閾値においても平均値より高い値を示していることがわかる．

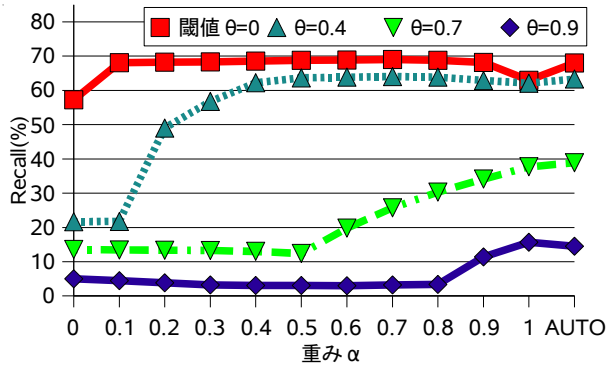

図 5：**Recall**（材料科学）

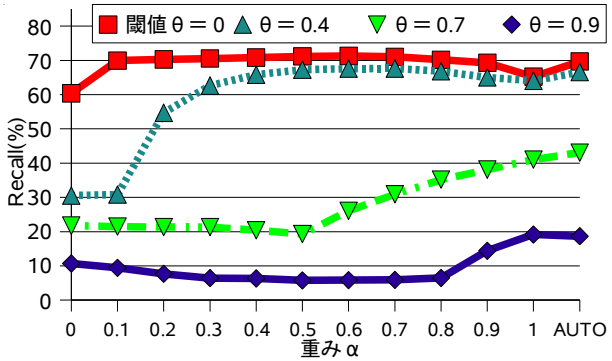

図 6：**Recall**（計算機科学）

同様に，冠詞付与精度（Precision）を図 7，図 8 に示す．自動設定の重み α は AUTO と記載している．閾値 θ=0 では，ほぼ横ばいながら重み α=0.6 付近で精度が最も高い．閾値 θ=0.4 でも横ばいに近い形で，α=0.2 付近のとき精度が高くなる．また，θ=0.7 では，α=0.6 付近で，θ=0.9 では α=0.4 付近で精度が高くなる．α を自動設定とした場合には，α を変化させた場合での最低値を下回ることは無いものの，あまり高い精度となっていない．

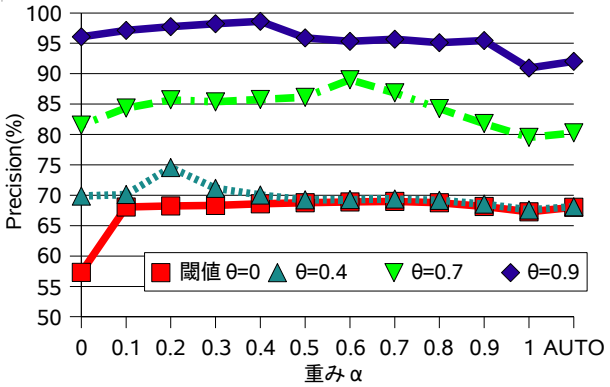

図 7：**Precision**（材料科学）

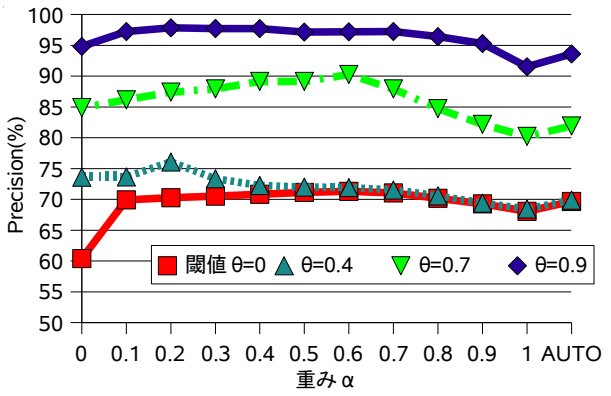

図 8：**Precision**（計算機科学）

最後に，総合的な性能値を示す F 値を図 9，図 10 に示す．自動設定した場合の重み α は AUTO と示す．F 値も Recall と同様に，閾値が小さい場合（θ≦0.4）では，ほぼ横ばいに近いかたちで α=0.6 付近で最も高い値となる．また，ベースラインであるイディオムのみと比較してもほとんど高い値となる．しかし，閾値が高くなるにつれ，イディオム手法のみのほうが性能が高くなることがわかる．α を自動設定とした場合も，Recall と同様，高い値を示している．

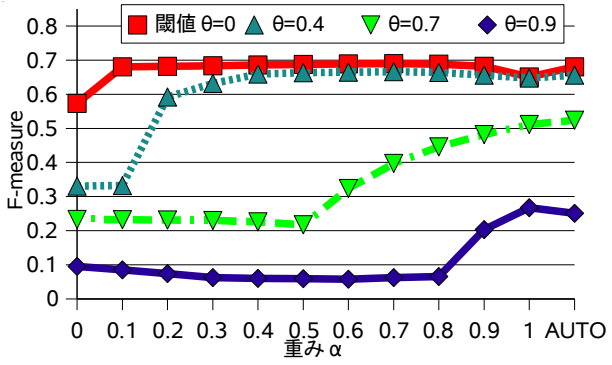

図 9：**F 値**（材料科学）

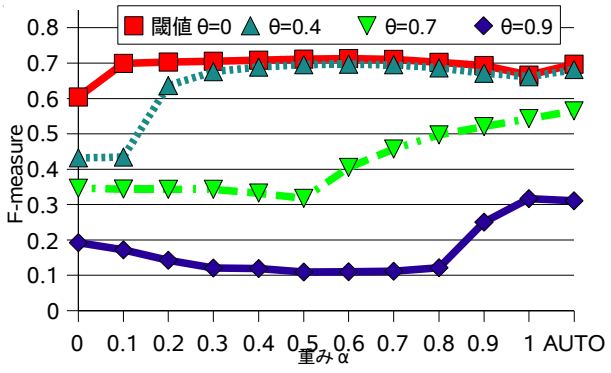

図 10：F 値（計算機科学）

5．考察

提案した統合手法は，閾値 θ が低い場合には，文脈手法イディオム手法単独で用いるよりも高い性能となる．中でも，イディオム手法にやや重みを持たせる（α＝0.6 ~ 0.7）ほうが，わずかに性能が高くなる．閾値 θ が高くなると，イディオム手法のみ用いるほうが性能が高くなり，閾値によって最適な α の値は変化することがわかる．精度（Precision）でも，閾値 θ＝0.7 の場合，イディオム手法にやや重みを持たせると高い値となり，θ＝0.9 の場合にはやや文脈手法に重みを持たせたほうが高い，と最適な α の値は変化している．また，分野における影響も少しながら存在する．したがって，最適な α の値は，冠詞付与を行う文章や閾値ごとに異なり，その値を求めるために何度も冠詞付与を行うのは効率的ではない．

一方で，α を自動設定とした場合には，閾値の値に関わらず高い性能を示している．また，閾値 θ＝0.7 の場合には，重み α を変化させた場合よりも高い値となっている．したがって，閾値 θ＝0.7 付近では，文章単位で α の値を固定するよりも，冠詞付与箇所ごとに設定するほうが，高い値を示すことがわかる．これは，両手法の付与規則の出現頻度により，文脈手法とイディオム手法のどちらの傾向が適しているかを推定でき，それは冠詞付与箇所ごとに異なっているためである．したがって，α を自動設定する手法は，パラメータの設定の必要が無くなると同時に，より実際の冠詞決定に近づけるため，冠詞付与の性能向上につながり，θ＝0.7 前後で最も効果的である．

しかし，精度（Precision）のみを見た場合には，高い値となっているとは言えない．これは，イディオム手法において，頻度は大きいが，イディオムと見なせない規則（図 4 における規則「in 〈the〉」など）が最大の値となった場合に，頻度比から単純に，イディオム手法に重みを持たせることが影響していると考えられる．これは，α の自動設定に，イディオムの長さを考慮に入れることで対応できる．また，冠詞付与規則が見つからなかった場合に，全て 0.33 として新たな生起確率を計算しているが，これも，より適した値があると考えられる．これには，冠詞の分布を用いるとよい．すなわち，名詞の情報や文脈情報が全くない状態での冠詞の生起確率となるため，冠詞付与規則取得の際に得ることができる．

また，本論文においては，正しい冠詞は原文で用いられた 1 つであるとみなして評価を行っている．しかし，「the」の省略など，複数の冠詞（the／φ）が正しいとされる場合も考えられる．実際に，英文添削を職業とする母語話者が冠詞を選択した結果について，図 11 に示す．これは，実験に用いた論文誌の文章から冠詞を消去し，科学技術分野の添削経験のある英文添削者 8 人が冠詞付与を行ったものに，原文を合わせた 9 つの英文での冠詞選択の状況を示している．この調査では，全員が選択した冠詞が一致する場合は半数に満たず，選択する冠詞に揺れが存在することがわかる．これを考慮することも，性能評価には必要であると考えられる．

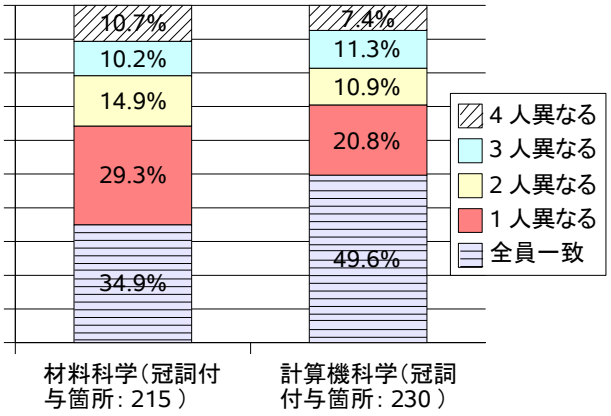

図 11：ネイティブによる冠詞付与結果の揺れ

6．まとめ

本論文では，以前に提案された文脈手法，イディオム手法を統合する新たな手法を提案した．また，その際に用いる重みを自動で設定する手法を提案した．性能は片方の手法単独で用いるより高くなることが示せた．また，統合における重みを自動的に決定する手法は，設定しにくいパラメータの除去ができる上に，冠詞決定の要素を付与箇所ごとに決定できることから性能向上につながり，効果的である．今後の課題として，重みの設定手法の改良と，複数の冠詞が正解である場合を考慮に入れた評価の必要がある．

参考文献

[1] 鈴木 英次，科学英語のセンスを磨く，化学同人，1999
[2] 井口 達也，永田 亮，河合 敦夫，英文アブストラクトを対象とした冠詞付与手法：電気関係学会当会支部連合大会，O-426，2004
[3] J. Lee，Automatic Article Restoration：Proc. HLT-NAACL2004，May 2004
[4] R. Han，Detecting Errors in English Article Usage with a Maximum Entropy Classifier Trained on a Large，Diverse Corpus，Proc. 4th International Conference on Language Resources and evaluation，May 2004
[5] Nagata，et al，Extracting Collocations for Determining Articles in English Writing: Proc. PACLING2005, Aug. 2005.
[6] 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
[7] Institute for Scientific Information, Journal of Citation Reports