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，正しい冠詞の選択は困難となる．その一方で，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（以後，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に基づいて，冠詞を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（以後，この手法を「イディオム手法」と呼ぶ）．

本論文では，文脈手法とイディオム手法を組み合わせて付与精度を向上させる手法を提案する．文脈手法とイディオム手法では，それぞれ付与できる冠詞の傾向が異なるため，両者をうまく組み合わせることで付与精度の向上が期待できる．組み合わせは，2 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ことで行う．本論文では，前述のとおり，専門用語の冠詞決定が特に難しいことを考慮して，対象文書を科学技術英文（計算機科学，材料科学，医化学の論文）とする．

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. で両手法を統合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．ここで文脈情報とは，文中においてその名詞の出現が何度目か，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（取得対象は“... ultimate goal of ...”の「goal」）. 例えば, コーパス中に出現する名詞「goal」のうち，名詞「goal」において無冠詞の生起確率は，無冠詞となって出現する割合で推定できる．同様に，名詞「goal」が文中で初めて出現する場合において，無冠詞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の生起確率が推定できる．形容詞「ultimate」が修飾する場合，前置詞「of」が後に接続する場合の冠詞生起確率も同様に推定できる．「a」「the」についても，同様に生起確率をコーパスから推定する．推定した確率と文脈情報を付与規則として辞書に登録し，冠詞付与の際に利用する．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 …

名詞「goal」のみの冠詞生起確率（φ）＝ 名詞「goal」の無冠詞での出現頻度 / 名詞「goal」の出現頻度

文中で初出の場合の冠詞生起確率（φ）＝ 文中で初出の場合で無冠詞となる出現頻度 / 文中で初出の場合の出現頻度

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

前置詞「of」が後に接続する場合の冠詞生起確率（φ）＝ 名詞「goal」が無冠詞で後に前置詞「of」が接続する場合の出現頻度 / 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 θf未満の場合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 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) = 「is altered as **a** result」の出現頻度 / 「is altered as 〈冠詞〉 result」の出現頻度

「as 〈冠詞〉 result of」での冠詞生起確率(a) = 「as **a** result of」の出現頻度 / 「as 〈冠詞〉 result of」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．冠詞全体の出現頻度がθ_f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．両手法それぞれで冠詞付与を行う場合に用いられる規則と，その生起確率を重み付きで混合し新たな冠詞生起確率としたものを用いる．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を，1 に近いほどイディオム手法を優先するとする．ただし，どちらかの手法で規則が辞書に登録されていない場合には，もう一方の手法での冠詞生起確率のみを用いる．また，この重みαを様々な値に変化させることで，最適な重みを調査した．

新たな冠詞生起確率 = (1－α)×文脈手法で用いた冠詞生起確率 ＋α×イディオム手法で用いた冠詞生起確率

文脈手法で規則が存在しない場合:

新たな冠詞生起確率 = イディオム手法で用いた冠詞生起確率

(イディオム手法で規則が存在しない場合も同様)

図 3:統合手法での冠詞生起確率の導出法

図 4 に，実際の提案手法での冠詞付与の例を示す．図 4 のように，名詞「function」に冠詞を付与する場合，文脈手法では，前に前置詞「as」後に前置詞「of」が接続するときの，不定冠詞の生起確率が 92.86%と最大となり，この規則と生起確率が用いられる．一方イディオム手法では単語列「as 〈冠詞〉 function of the」のときの不定冠詞の生起確率が 94.12%と最大で，この規則と生起確率が用いられる．したがって，α＝0.4 の場合，提案手法における新たな不定冠詞の生起確率は，両手法で用いられる規則と生起確率から導出して 93.364%となる．

冠詞付与箇所
…as 〈冠詞〉 function of the scattering vector …

付与規則:文脈手法

1	〈a〉 function 前に前置詞「as」後に前置詞「of」	92.86
2	〈a〉 function 前に前置詞「as」	92.21
3	〈a〉 function 後に前置詞「of」	75.51

付与規則:イディオム手法		生起確率
1	as 〈a〉 function of the	94.12
2	as 〈a〉 function of	93.75
3	〈a〉 function of the	93.18

α＝0.4の場合

a	不定冠詞の生起確率＝0.6×92.86+0.4×94.12=	93.364
the	定冠詞の生起確率＝0.6×0＋0.4×0＝	0
ϕ	無冠詞の生起確率＝	0

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，医化学，材料科学（セラミック），計算機科学（人工知能）の 3 つの分野を選んだ．それぞれの分野から 1 つの論文誌を選択し，200 論文の英文（平均 80 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（冠詞付与箇所は平均 5600 箇所）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．引用数の多い論文誌であるため，冠詞の用法などの文法的な誤りは，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）[6] 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，冠詞を付与する箇所のうち，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精度（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から1までの0.1刻みで変化させた．さらに，提案手法で用いる重みαの値を 0.1, 0.3, 0.5, 0.7，0.9 と変化させた．また，α＝0 の場合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ので，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，医化学，材料科学，計算機科学の順に図 5 から図 7 に示す．Recall は，閾値が大きい場合（θ ≦0.3）にはα＝0.5 の時に高い値を示す．また，閾値が上がるにつれ，αが大きいほうが Recall は高くなることもわかる．しかしながら，ベースラインとしたイディオム手

法のみでの付与率をほとんど下回る結果となった．

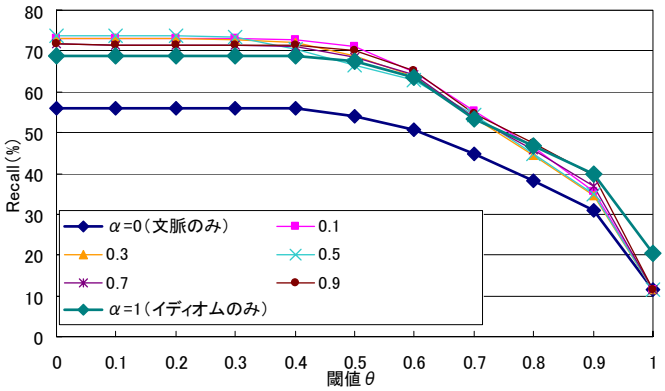

図 5：Recall（医化学分野）

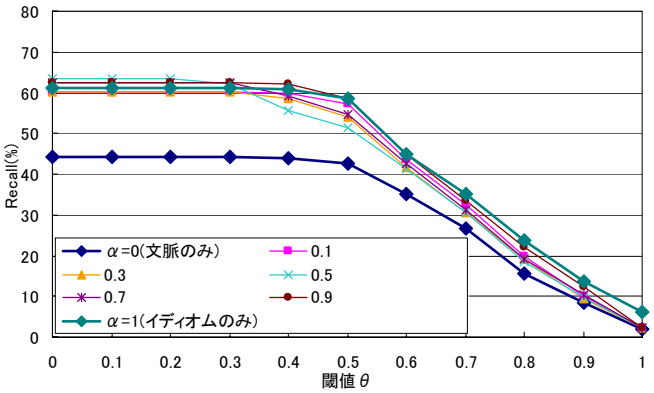

図 6：Recall（材料科学分野）

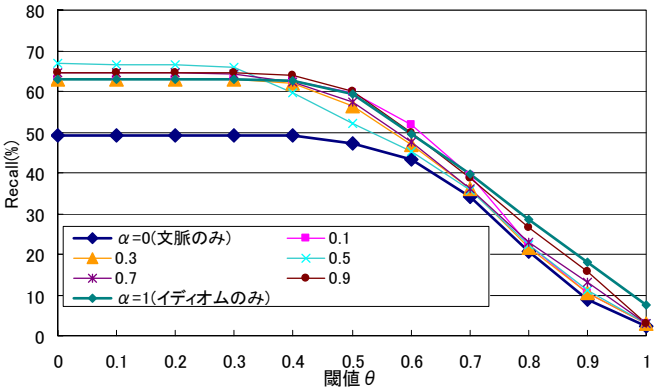

図 7：Recall（計算機科学分野）

同様に，冠詞付与精度（Precision）を図 8 から図 10 に示す．閾値が大きい場合には全体的に精度が高く，閾値が小さい場合でも，文脈手法のみでの精度を下回ることはほとんど無かった．全体的には α =0.5 から 0.7 である場合に最も精度が高くなった．

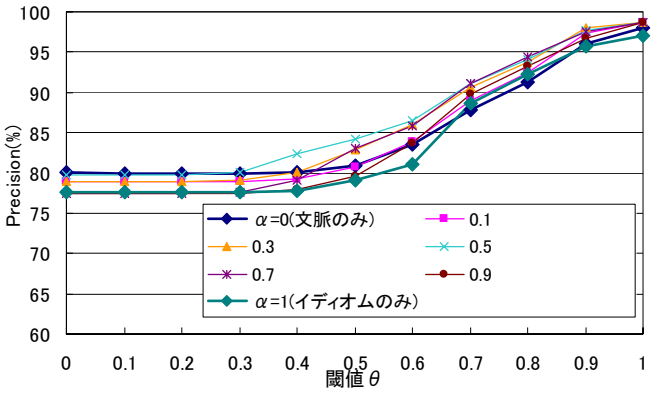

図 8：Precision（医化学分野）

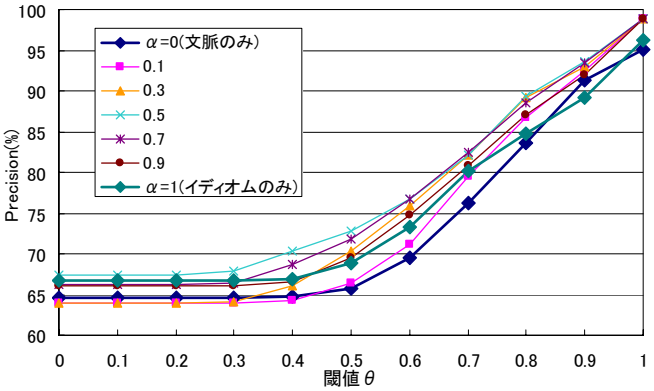

図 9：Precision（材料科学分野）

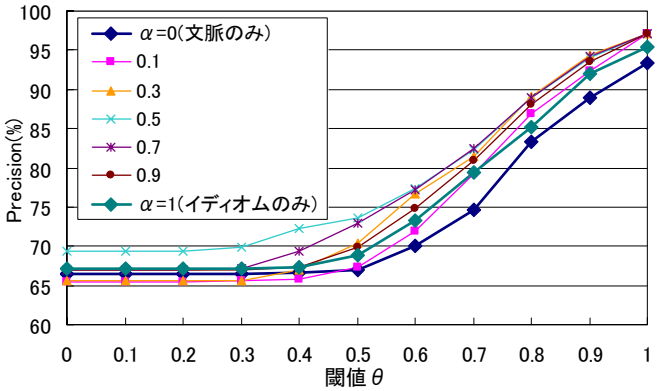

図 10：Precision（計算機科学分野）

最後に，総合的な性能値を示す F 値を図 11 から図 13 に示す．F 値も Recall の場合と同様に，閾値が小さい場合には重み α =0.5 の場合がベースラインであるイディオム手法のみでの F 値を上回る結果となる．一方で，閾値が大きくなると α =0.7 から 0.9 の場合にわずかに上回るものの，全体的にはイディオム手法のみの性能を下回る結果となった．

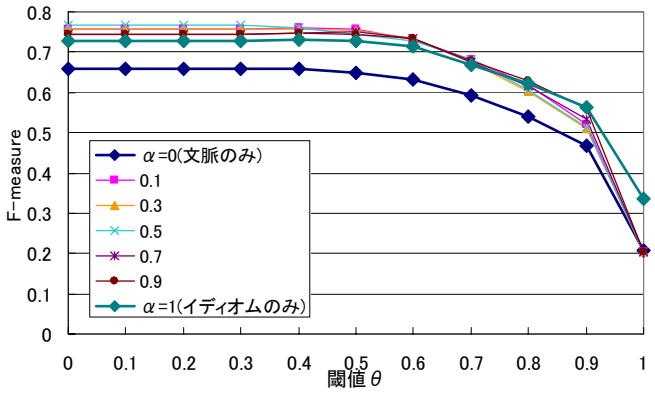

図 11:F 値(医化学分野)

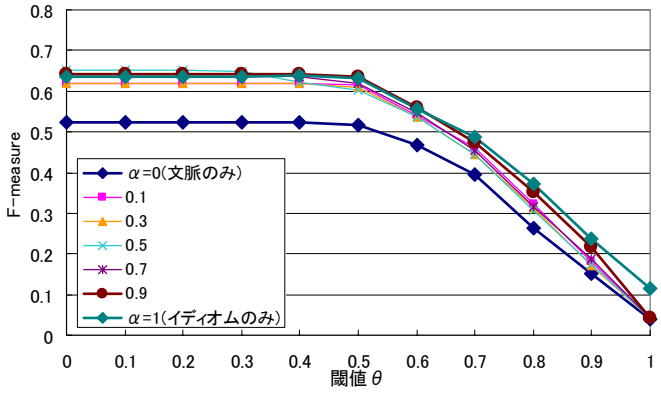

図 12:F 値(材料科学分野)

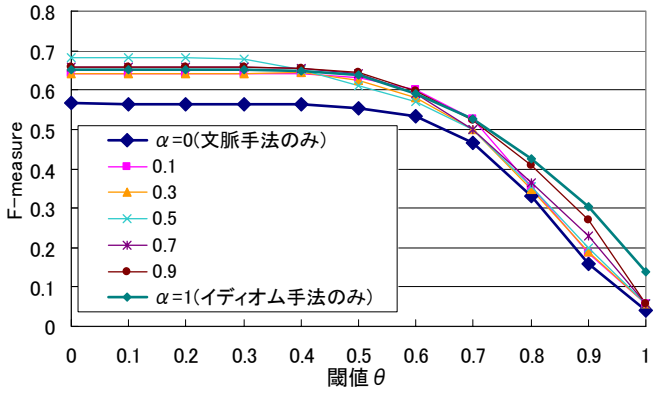

図 13:F 値(計算機科学分野)

5. 考察

提案した統合手法は，生起確率の閾値 θ が低い場合には，重み α =0.5 のときに F 値が最大となり，文脈手法，及び，イディオム手法を上回った．すなわち文脈手法，イディオム手法に等しい重みを与えたときに性能が最大となった．この理由として，文脈手法とイディオム手法共に，生起確率に偏りがなく，各手法単体ではうまく冠詞付与できない場合でも，両者の情報を混合して用いることで，より確実に冠詞が決定できるためであると考えられる．この傾向は，医化学分野，材料科学分野，計算機科学分野，全ての分野に共通して見られる．したがって，分野を問わず，文脈手法とイディオム手法を混合したほうが冠詞付与の性能が高くなるといえる．

一方で，生起確率の閾値 θ が高い場合には，イディオム手法に，より重みを与えたときに性能が比較的良くなった．イディオムでは，慣用的に冠詞が一意に決定される（例えば，「as a result」など）ことが多い．すなわち，イディオムでは，ある冠詞の生起確率が非常に高くなることが多い．特に，専門英語では，分野に依存した慣用的な冠詞の決定が多く見られる（例えば，化学分野での「the chemistry of」など）．そのため，イディオム手法では，生起確率の閾値 θ が高いときに，付与性能が高くなる．ただし，慣用的な冠詞の決定は，冠詞の全用法の一部であるため，Recall が低くなることに注意しなければならない．

以上を考慮すると，ある程度高い付与精度で網羅的に冠詞を付与するときには，文脈手法とイディオム手法を同じ重みで混合するべきであるといえる．また，非常に高い付与精度で一部の名詞にのみ冠詞を付与する場合は，α を大きくしてイディオム手法に依存した付与を行うべきであるといえる．

また，本実験では，前述のとおり文脈手法とイディオム手法から最大でもそれぞれ 1 つずつしか規則を用いていない．また，2 手法で付与しようとする冠詞が食い違う場合にはどちらの冠詞生起確率も重み α によって値が小さくなってしまう．これを解決するために，文脈手法，イディオム手法から，最大の生起確率となる付与規則だけでなく，「a」「the」「ϕ」それぞれを付与する規則のうち最大の生起確率を持つものを取り出し，その冠詞生起確率を重み付きで混合する方法が考えられる．

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手法を提案した．文脈手法，イディオム手法を同じ重みで混合することで，性能は単独で用いるよりも良くなることが示せた．今後の課題として，この統合手法に用いる付与規則を増やすことがあげられる．

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：電気関係学会東海支部連合大会, 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] Institute for Scientific Information, Journal of Citation Reports