修士論文

参照を考慮した技術英文への冠詞付与システムの研究

平成　22 年度修了

三重大学大学院工学研究科

博士前期課程情報工学専攻

小島　鉄也

目　次

1. はじめに 1
2. 研究背景 2
2.1. 冠詞の用法 2
2.2. 可算/不可算情報を利用した手法 4
2.3. 慣用句による冠詞の学習手法 6
2.4. 前置詞・形容詞・文脈を考慮した冠詞付与手法 8
2.5. 従来手法の問題点 10
3. 提案手法 11
3.1. 前方参照の調査 12
3.2. 規則の学習方法 14
3.3. 規則の適用 17
4. 評価実験 18
4.1. 実験手順 18
4.2. 評価方法 19
4.3. 実験結果 20
5. 考察 24
5.1. 定冠詞の付与精度 24
5.2 冠詞全体の精度 25
謝辞 26
参考文献 27

1. はじめに

近年，日本人が英語に触れる機会が多くなっており，英文添削の重要性は増している．特に理系の研究者にとって，英作文をする機会は多い．論文を日本語で書く場合であっても，アブストラクトは英語で記述するように求められることが多いからである．このことから，投稿する前に英文の添削を依頼する研究者も少なくない．しかし，英文の添削は時間と費用がかかる．そのため，英文の自動添削システムの需要が高まっている．特に，日本語には冠詞の概念が存在しないため，日本人は冠詞の用法を誤りやすいと言われている．河合らは，実際に日本人の書いた英文には冠詞の誤りが多いことを確認している[1]．科学技術分野においては，可算/不可算のいずれでも使用される名詞が多い[2]．そのうえ，技術文書向けの豊富な用例を含む冠詞の文法書は存在しないため，冠詞の判断はさらに困難となる．

文章中の冠詞付与誤りを自動的に検出し利用者に提示する構成ツールが実現されれば，このツールは英作文を行う場合に有用である．この冠詞の誤りを検出するために，いくつかの手法[3][4][5]が提案されている．たとえば井口らは，電子化された英字新聞などから統計量を抽出し，その統計量を用いた2つの冠詞付与手法を提案している[3]．また若菜らは，可算/不可算情報を利用して冠詞誤りを検出する手法を提案している[4]．しかし，従来提案されている手法は基本的に一文内の処理を中心におこなっているため，冠詞の参照を考慮出来ていない．また，名詞句が Japanese wordnet のように二語以上の単語の場合，後ろの wordnet を冠詞付与の対象としている．

本研究では Japanese wordnet のような二語以上連続する名詞列を複合名詞，一語からなる名詞を単一名詞と定義し，これらの名詞が一度目に出現する場合と二度目以降に出現する場合にわけて規則を学習する．参照の the を考慮した冠詞付与をすることで，定冠詞付与の精度向上を目指す．本手法の適用範囲は，同じ名詞が複数回出現する機会が多い文章全般だが，本論文では科学技術分野に限定して実験・評価を行う．

以下，2 章で従来研究，3 章で提案手法，4 章で評価実験について述べ，5 章で考察を行う．

2. 研究背景

本章では，冠詞の基本的な用法とこれまでに提案されてきた手法を紹介する．本章の流れは以下のとおりである．2.1節で冠詞の用法，2.2節で可算/不可算判定手法，2.3節で慣用句による冠詞付与手法，2.4節で前置詞・形容詞・文脈を考慮した冠詞付与手法，2.5節で従来研究の問題点について述べる．

2.1. 冠詞の用法

冠詞の用法は，主に対象となる名詞句の中心となる名詞が，可算/不可算のいずれか，読者にとって特定・限定できる対象か，の2つの要素によって決定される．

科学技術分野では可算/不可算のいずれでも用いられる名詞が多く，その名詞のみで可算/不可算を判断するのは困難である．若菜らは名詞の周囲にある単語から，可算/不可算の判定規則を学習する手法を提案している．詳しくは2.2節で述べる．

対象となる名詞句が特定・限定できる場合は，冠詞の「the」や指示代名詞，所有代名詞が使用される．ここで「the」の用法は，一般的な文法書によると，以下の7つに分けることができる．

1. 同一文内の関係代名詞等に修飾されることにより，限定特定される場合．
2. 形容詞の最上級で名詞句が修飾される場合．
3. 慣用句の一部にtheが出現する場合．
4. 同一文書中で二回目に出現する場合（前方参照）．
5. 名詞そのものに関する知識を読者がもっていることにより，特定限定できる場合（外界照応）．
6. 発話時の周囲の状況から，何を指しているか分かる場合（指さしなどの言語外情報のがかり）．
7. 総称的に使われる場合（The ostrich cannot fly.のように種族全体を表す場合）．

従来の冠詞誤り検出システムでは，処理時に同一文内の処理のみで済むことから，用法1，2，3を規則として学習している．これらの手法は英字新聞やエッセイ，Web文書といった冠詞が付与された一般文書の集合から，冠詞を付与する特徴的な規則を統計ベースで学習する．学習の詳細は2.3節で述べる．

英字新聞やエッセイは1トピックあたりの単語数が少ないため，同じ名詞が複数回出現する可能性が少ない．そのため，名詞句の前後4単語で冠詞が97%定まると言われており[6]，規則の学習は同一文内の処理に限定されてきた．唯一，井口らは参照のtheに対処するために，複数の文を解析して規則の学習を行なっている．この手法は2.4節で詳しく述べる．

科学技術論文で出現する複合語の多くは，一定の概念を表す技術用語として使われることが多い．定着の度合いの深い技術用語は，いわゆる技術用語辞書やウィキペディアに記載されている．しかしながら，これらのみで多くの技術用語をカバーしているとは言い難い．そのため，分野内の知識をデータベースから学習することは容易ではない．平野らは

検索エンジンを利用することで，単語自体の出現回数を増やし，作成する規則数を増やしている[7]．確かにWEBには様々な知識が詰まっている．しかし，玉石混淆な文書内から有用な情報のみを取り出すのは困難である．特に，冠詞付与手法では，正しいと仮定した英文を学習している．そのため，英文が正しいという仮定が難しいWeb文書ではノイズが多く含まれてしまう．そのため2つの視点から，外界照応に対処することが考えられる．1つは，専門分野別論文集を利用であり，もう1つは，参考文献の利用である．例えば，情報工学の英文添削のためには，情報工学の論文集から収集した複合名詞を利用する．そうすることで，技術用語の収集を効率的に行える．さらに，添削対象論文の参考文献を利用すれば，その論文に最も内容が近い複合名詞の収集が期待できる．用法6は，主に話し言葉における問題であり，文書の添削では考慮する必要がない．用法7の解決は必要であり，若干のヒューリスティックスも提案はされている．しかし，出現頻度が1%程度[12]と多くない．

2.2. 可算/不可算情報を利用した手法

名詞の可算/不可算の判定は，正しい英文から実際の使用例を規則として学習し，その規則を評価用の英文に適用することで行う．以後，この規則を判定規則とする．

正しい英文中では，比較的容易に名詞の可算/不可算の判定が行える．例えば，

(A)This allows the user to implement these devices in power sensitive applications.

英文(A)の場合 device，application ともに複数形になっているため，これらの両方が可算名詞として使われていることがわかる．しかし，user に関しては定冠詞が付与されており，可算/不可算の判別が不可能である．これらのように，言語学の知見に基づいて作成された規則を使い，正しい英文中の名詞に対して可算/不可算の判定を行い，その情報を付与する．

冠詞誤りの検出は可算/不可算の判定結果と，現在付与されている冠詞や単数形/複数形に文法的な矛盾が発生していないかを判定することにより行う．例えば，不可算名詞が複数形で使用されていたり，不定冠詞が付与されていたりすれば，それは文法上の誤りとして検出できる．このような判定をまとめたものを表 1 に示す．表中の○はその使用法に誤りがないことを示し，×は誤りがあることを示す．

表 1. 可算/不可算に基づいた冠詞誤りの判定規則

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

また，表 1 に加えて，表 2 の誤りも検出可能である．表 2 の記号の意味は表 1 と同じである．例えば，可算名詞が much に修飾される場合は誤りであることがわかる．なお，不可算名詞を複数形で用いることは，基本的に誤りであるので，表 2 では省略した．

表 2. 可算/不可算に関連した誤り

	可算		不可算
	単数形	複数形	単数形
another, each, one	o	×	×
all, enough, sufficient	×	o	o
much	×	×	o
that, this	o	×	o
few, many, these, those	×	o	×
various, several, numerous	×	o	×
one 以外の数詞	×	o	×

表 1, 2 を用いて, 英文(B)に対し可算/不可算判定の誤り検出を行うと, 可算名詞の単数形 device が無冠詞で使用されていることから, 表 1 より, この用法を誤りとして検出できる.

(B)It is implemented by hardware device.

2.3. 慣用句による冠詞の学習手法

2.1 節で述べたように，冠詞は基本的には，冠詞付与の対象となる名詞が数えられるか，限定できるかで定まる．しかし，慣用句による例外が多数存在する．例えば，一般に可算名詞を無冠詞で使用することは誤りであるが，慣用句では許される(e.g. by car)．井口らは慣用句では可算名詞にも無冠詞が付与されるという規則を学習するために，冠詞を含む単語の並びを学習する冠詞付与手法を提案している[3]．まず，大規模な英語で記述された文書集合(以下，コーパスとする)から冠詞を含む単語列を抽出し，その単語列から冠詞の付与規則を学習する手法である．ここでは，頻出する単語列を慣用句とみなし，中でも冠詞の分布に偏りのあるものだけが冠詞付与規則となる．ここでは，真の意味の慣用句と区別するために，システムが抽出する単語列をイディオムと呼ぶ．

まず，KWIC[8]を応用して，冠詞を含むイディオムを抽出する．KWIC (KeyWord In Context) とは，あるキーワードを中心とした文字列で，ここでは冠詞をキーワードとしてコーパスから KWIC を作成し，単語列を得る．このように入力文中の全ての冠詞に対して KWIC を取得することで，冠詞を含む，イディオムを文書中から抽出できる．対象の冠詞から，前にn_a単語，後にn_b単語を取得し，得られた単語列を 1 つの KWIC とする．次に，得られた KWIC から冠詞を含む単語の並びを全て抽出し，データベースに辞書順に格納する．KWIC が“altered as a result of”のときにデータベースに格納される全ての組み合わせを表 3 に示す．1 つの KWIC から，$(n_a+1)\cdot(n_b+1)$個の組み合わせが抽出されるが，冠詞のみではイディオムとして扱えないので除外する．すべての KWIC から抽出を行うと，コーパス中でよく使われる表現は，複数回出現することになる．この単語列をイディオムとみなし，その出現回数を数え，頻度を取得する．

表 3. KWIC が“alterd as a result of”のときにデータベースに格納される組み合わせ

2 つ前	1 つ前	冠詞	1 つ後	2 つ後
altered	as	a	result	of
altered	as	a	result	
altered	as	a		
	as	a	result	of
	as	a	result	
	as	a		
		a	result	of
		a	result	

このように取得したイディオムの生起頻度から，イディオム中の冠詞の生起確率を求めることができる．冠詞をART，ARTを含むイディオムをIとすると，IにARTが含まれる生起確率は

$$\mathrm{p}(ART|I) = \frac{f(ART|I)}{f(I)}$$

で求めることができる．$\mathrm{p}(ART|I)$をイディオムの冠詞生起確率と定義し，これをコーパスから学習し，辞書に登録する．得られたイディオムの冠詞生起確率を用いて冠詞付与を行う．入力された英文について，冠詞を付与すべき個所（以下，冠詞付与個所と呼ぶ）を検索する．冠詞付与個所は，あらかじめ人手により記入しておく方法と，Chunker[9]を用いて名詞句を抽出し，名詞句の直前を冠詞付与個所とする方法が考えられる．この検索されたそれぞれの冠詞付与個所について，上記の方法で作成したイディオムの冠詞生起確率が登録された辞書を検索し，生起確率を取得する．

まず，検索に用いる単語列を取得する．検索された冠詞付与個所を基準に，前後の単語列を取得する．取得できるのは，対象の冠詞付与個所から，前後の冠詞付与個所の直前にある単語までとなる．さらに，得られた単語列から，冠詞を含む全ての組み合わせについて，イディオムの冠詞生起確率辞書を検索し，冠詞生起確率を取得する．なお，イディオムの単語数は考慮していない．また，検索された生起確率の中に，最も高い確率値が複数存在する場合は，辞書に最初に登録されているイディオムを用いる．

2.4. 前置詞・形容詞・文脈を考慮した冠詞付与手法

名詞に付与する冠詞を決定するには，付与対象の名詞の性質や，意味的に限定されているかを知る必要がある．本手法では，名詞がどの冠詞に修飾されやすいかをコーパスから学習する．さらに，文脈，前置詞，形容詞を考慮することで，名詞が限定的であるかを判断し，冠詞を付与する．本論文では，a，the，冠詞がない状態（ϕ）の3種類を冠詞とする．

N を名詞，ART を冠詞とすると，N がART で修飾される確率は

$$P(ART|N) = \frac{f(ART|N)}{f(N)}$$

と，条件付き確率で表すことができる．ここで，f はコーパス中の生起頻度を表す．例えば，“method”という名詞が100回出現し，そのうち60回「the」が修飾した場合，

$$\mathrm{f(method)} = 100,\ \mathrm{f(the|method)} = 60,\ \mathrm{P(the|method)} = {}^{60}/_{100} = 0.6$$

となり，この$P(ART|N)$ を冠詞生起確率と呼ぶ．この生起確率をコーパス中から学習し，辞書に登録する．

この冠詞生起確率を用いれば，ある程度の精度で冠詞付与を行うことができる．ただ，多くの名詞では，文脈によって冠詞が使い分けられるため，単純に冠詞生起確率が最も高い冠詞を付与するだけでは，精度の高い冠詞付与を行うことはできない．精度の高い冠詞付与を行うためには，名詞の周囲に出現する単語を的確に学習する必要がある．本手法では，文脈，前置詞，形容詞を考慮した冠詞生起確率をコーパスから学習し，それらを冠詞付与に用いることで状況に応じた冠詞付与を行う．

文脈によって冠詞が決定される場合，名詞が文章内で初出のときは，その名詞は非限定的になりやすいため，不定冠詞「a」が付与されることが多い．また，ある名詞が，同一文章内で既に出現していたとき，その名詞は限定されやすく，定冠詞「the」が付与されることが多い．この冠詞用法に対応するために，文脈を考慮した冠詞生起確率を定義する．いま，名詞をN ，冠詞をART ，N が前回出現したときに修飾していた冠詞をPRE_ART とし，N が前回PRE_ART に修飾されていた状態を${}^{\mathrm{PRE_ART}}N$ と表す．なお，N が文中で初出のときは，PRE_ART = first とする．N が前回出現したときの冠詞がPRE_ART で，同じ文中でN が再び出現したときART に修飾される確率は

$$\mathrm{P}\left(ART|{}^{PRE\_ART}N\right) = \frac{f\left(ART|{}^{PRE\_ART}N\right)}{f\left({}^{PRE\_ART}N\right)}$$

となる．この$\mathrm{P}\left(ART|{}^{PRE_ART}N\right)$ を，文脈を考慮した冠詞生起確率と呼び，これをコーパス中から学習し，辞書に登録する．

また，名詞句の前後の前置詞を考慮する場合，その前置詞によって，名詞の意味が限定的，または非限定的になることがある．例えば，文献[2]によると，mechanism の後に of があるとき，mechanism は一般的な意味で用いられていることが多く，意味が非限定的になるため，無冠詞「ϕ」になりやすい．したがって，名詞句の前後の前置詞を参照することで，冠詞が決定できる場合がある．このような前置詞の効果に対応するために，前置詞を考慮

した冠詞生起確率を定義する．先程の定義に加えて，N を含む名詞句をNP，NP の前にある前置詞をPP_b ，NP の後にある前置詞をPP_aとする．NP の前にPP_b ，NP の後にPP_b があるとき，N がART に修飾される確率は

$$\mathrm{P}(ART|PP_bNPP_a) = \frac{f(ART|PP_bNPP_a)}{f(PP_bNPP_a)}$$

となる．また，前置詞が前，あるいは後のみにあった場合も同様に考える．この$P(ART|PP_bNPP_a)$を，前置詞を考慮した冠詞生起確率と呼び，これをコーパス中から学習し，辞書に登録する．

名詞が形容詞で修飾されている場合，その形容詞によって名詞の意味が限定的または非限定的になることがある．例えば，文献[10]によると，original が名詞を修飾しているとき，その名詞は限定的になることが多く，定冠詞「the」が付与されやすい．したがって，名詞句内の形容詞を参照することで，冠詞が決定される場合がある．このような形容詞の効果を考慮するために，形容詞を考慮した冠詞生起確率を定義する．先程の定義に加えて，NP の主名詞を修飾している形容詞をADJ とする．ADJ が NP 内にあるとき（$ADJ \in NP$と表す），NP の主名詞がART に修飾される確率は

$$\mathrm{P}(ART|ADJ \in NP) = \frac{f(ART|ADJ \in NP)}{f(ADJ \in NP)}$$

となる．この$\mathrm{P}(ART|ADJ \in NP)$ を，形容詞を考慮した冠詞生起確率と呼び，これをコーパス中から学習し，辞書に登録する．

辞書に登録された４種類の冠詞生起確率を用いて，冠詞付与を行う．評価用コーパスから名詞句を抽出し，その名詞句から，名詞句の主となる名詞を抽出する．名詞句の主となる名詞とは，名詞句中で一番後ろに位置する名詞である．例えば，名詞句“ultimate goal”の場合，“goal”が主名詞として抽出される．また，名詞句の前後にある前置詞と，名詞句内にある形容詞も同時に抽出する．これら付与対象の名詞と，名詞句の前後の前置詞，名詞句内の形容詞を用いて，辞書から 4 種類の冠詞生起確率を取得する．辞書からは，複数の規則が得られることも考えられる．この得られた生起確率のうち最も確率値が高い冠詞を付与する．

2.5. 従来手法の問題点

2.3, 2.4節で紹介した手法はそれぞれ付与できる冠詞の傾向が異なる．2.4節の手法は，文脈を考慮しているために定冠詞の規則に強く，2.3節の手法は，単語の並びをそのまま学習しているため，冠詞の規則として例外的な慣用句に対処することができる．宮井らは，2つの手法それぞれから得られる生起確率を混合して統合する手法を提案している[11]．しかしこの手法は，規則の学習，生起確率の算出を手法毎に個別に行なっており，規則の作成方法を統合しているわけではない．

2.1節で述べた冠詞の用法に当てはめて考えると，可算か不可算かを判定したあとで，周囲の単語や文脈から限定されているかを判定，さらに例外的な場合に対処，と 3 つの手法を順序付けて統合することで，冠詞本来の用法をシステムが満たすことができる．つまり，それぞれの手法の精度を個別に上げることで，冠詞付与全体の精度向上を目指すことができる．本研究では，これまでにあまり議論されてこなかった文脈から限定される場合の冠詞付与を扱う．また，従来の研究では，冠詞 a，the，ϕ を同一で扱ってきた．つまり，学習の規則や冠詞付与のタイミングを同じ処理で行っている．しかし，上記の 3 つの手法を組み合わせることで冠詞付与全体の精度が見込めるため，本手法では，定冠詞 the についてのみ取り扱う．

3. 提案手法

名詞句が特定・限定できるのは大きく分けて，修飾語など周囲の単語による場合と，同一文書内で既に述べられている場合の二通りである．新聞記事やエッセイなど，1トピックあたりの単語数が少ない文書を対象とする手法の場合，同一文内の処理のみを行っており，文と文にわたる冠詞の参照は考慮されていない．しかし，単語数が多い文書では，同じ単語が複数回出現する可能性が高くなる．特に技術論文では，システムや手法の説明など，複数の文にわたって同一の名詞を用いて説明をするため，前方参照がよく用いられる．また，本来 the とするべき箇所に誤った定冠詞を付与すると，文の意味が大きく異なってしまう．そのため，定冠詞を正しく付与するためには冠詞の前方参照を考慮することが効果的である．参照を考慮する上で注意すべきことは，同じ意味を指している名詞句かどうかである．たとえ同一の表記であっても，異なる意味で用いられることがある．また，従来手法では冠詞付与の対象として，名詞句内の最後の名詞に注目している．例えば，solar system という名詞句の場合，system を冠詞付与の対象とする．これは規則を細かく分類することで，一規則あたりの出現頻度数が少なくなってしまう，いわゆるデータスパースネスの問題を避けるためである．しかし，前方参照を考慮する場合，二つの名詞句が同一の概念を指しているかが重要なポイントとなる．そのため，複数の単語から構成される名詞列（e.g. solar system）についても冠詞付与の対象とする必要がある．

本研究では，solar system のように二語以上連続する名詞列を複合名詞，system のように一語からなる名詞を単一名詞と定義し，これらの名詞が一度目に出現する場合と二度目以降に出現する場合にわけて規則を学習する．複合名詞を考慮して文と文にわたる解析を行うことで，定冠詞の参照を考慮した冠詞付与システムの構築を行う．そのための準備として，同じ名詞句が複数回出現したときに付与されている冠詞がどう変化するか調査した．

3.1. 前方参照の調査

定冠詞の前方照応的用法とは，同一文書内で同じ名詞が複数回出現したときに，後方の名詞が前方の名詞を受けることで特定され，the が付与されることを指す．そのため，同じ名詞が複数回出現した場合，一回目と二回目以降では冠詞の扱いが異なる．しかし，同じ名詞が複数回出現すれば二回目以降は必ず the になるわけではなく，むしろならない場合が多い．また，たとえ the になったとしても，その the が参照によるものか同一文内の周辺語から限定されたものか区別するのは難しい．

前方に出現している同一の名詞を参照・限定できるかどうかは，読者が同じ意味を指していると判断できる場合に限られる．すなわち，読者が判断できないような参照の仕方はされない．そのため，同じ名詞が前方に出現した 100 行後に再度出現したとしても，それらが同じ名詞を指している可能性は低いと考えられる．また，同じ名詞を直近で異なる意味で用いる可能性も低いと考えられる．これら 2 つの仮説を確かめるために，計算機科学分野の論文 200 編を対象に複数回同じ名詞が現れたときの名詞間の距離と参照の関係を調査した．

同じ名詞が 2 回以上出現したときに，名詞間の距離を次のように定義する．

「名詞間距離」 = 「次回出現したときの行番号」 − 「前回出現したときの行番号」

つまり，前回出現した行と次回出現した行が同一の行であったとき名詞間距離は 0 となり，次の文なら単語間距離は 1 となる．全ての名詞を同一で扱ったときの各冠詞の割合を図 1 に示す．φ は無冠詞，OTHER は my などの the 以外で後ろの名詞を限定する単語をまとめたものを表す．しかし，章を超えて出現する場合であれば参照は起きにくいが，OTHER は限定されているので the と同様に扱えば，10 行程度の距離ならば，同一文内以外では距離による差は見られなかった．

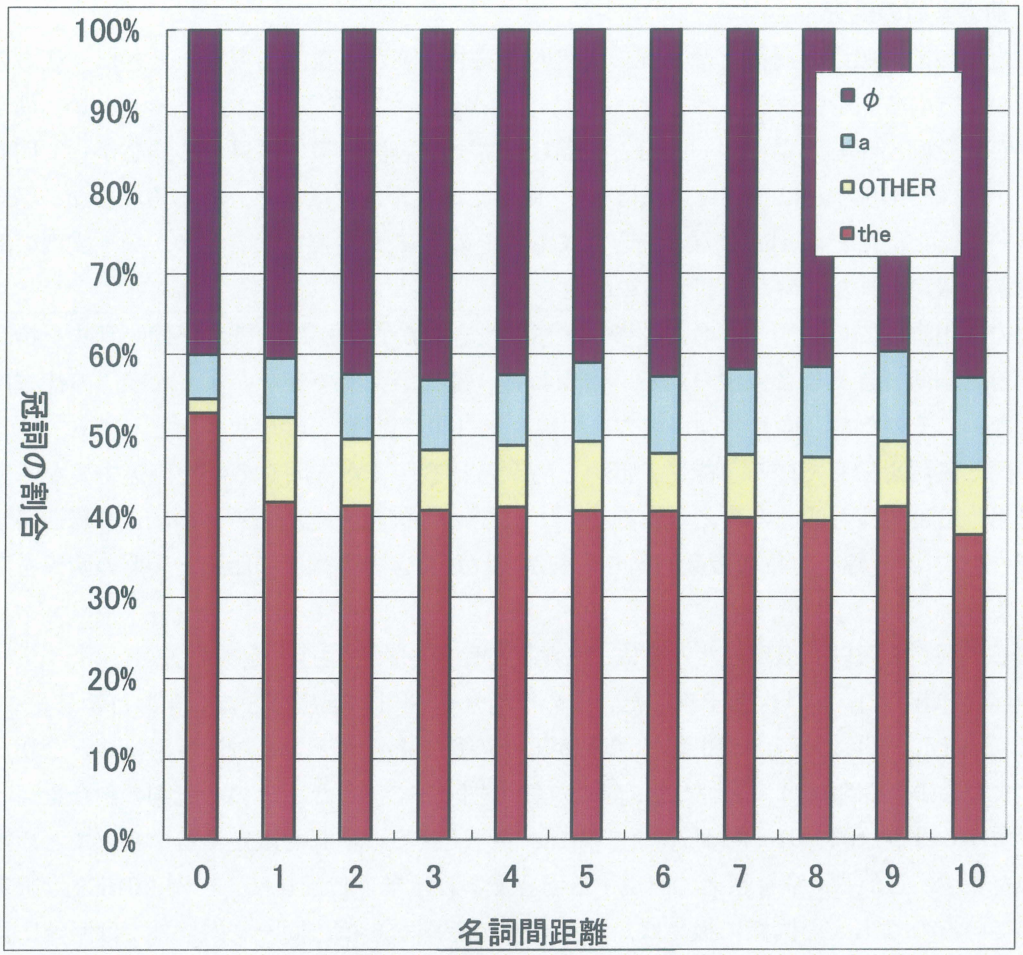

図 1：同一の名詞が複数回出現したときの名詞間距離と冠詞の関係

3.2. 規則の学習方法

冠詞付与システムは，論文などの文書群であるコーパスから規則を学習する学習フェーズと，学習した規則を適切な名詞句に適用して冠詞を付与する適用フェーズから成る．本節では，学習フェーズについて説明する．学習フェーズは次の三プロセスから成る．

① 名詞句の中で冠詞の付与対象となる単一名詞，または複合名詞を決定

② 名詞が初出か既出を判定

③ 名詞の出現状況と付与されている冠詞を学習

プロセス①では，名詞句の中で冠詞が指し示す対象となる名詞（以下，対象名詞）を判定する．例えば，

(C)We describe the construction of an illustrated Japanese Wordnet.

という文を処理する場合，名詞句は(a)the construction と(b)an illustrated Japanese Wordnet である．そして，(a)では[construction]という名詞が，(b)では[Japanese Wordnet]という名詞列がそれぞれ対象名詞となる．二語以上連続する名詞列を複合名詞，一語からなる名詞を単一名詞と定義する．

次にプロセス②では，対象名詞が初出か既出かを判定する．基本的には同一の単一名詞または複合名詞が同一文書内の前方に出現しているか否かで判断する．ただし，前方に複合名詞 Japanese Wordnet が出現し，後方に単一名詞 Wordnet が出現した場合，後ろの Wordnet は Japanese が省略されたと考え，既出と判断する．本研究では，同じ名詞が複数回出現したとき，規則の学習を一度目に出現した場合と，二度目以降に出現した場合で処理を分ける．そうすることにより，周辺単語による限定と前方参照による限定の双方を独立に規則として学習することができる．また，2.4節の手法では，前回出現したときに付与されていた冠詞によって，規則が分けられている．しかし，付与された規則を確認したところ，前回出現した名詞に the がつけられていたならば，次回出現した名詞にも the を付与する，といった同じ冠詞を付与する規則が基本的に適用されていた．そのため本研究では，この規則は用いない．

最後にプロセス③では，プロセス①②で得られた判定結果を元に付与規則を学習する．学習する方法は 3 パターンにわかれる．判定結果それぞれが行う学習方法を表 4 にまとめる．

表 4. 名詞の種類と学習方法の関係

	初出		既出	
	単一名詞	複合名詞	単一名詞	複合名詞
学習方法 α	○	○	○	○
学習方法 β	×	○	×	○
学習方法 γ	×	×	○	○

一文内における冠詞の規則は主に，冠詞付与の対象となる名詞の前後にある前置詞や形容詞によって定まる．学習方法αでは，単一名詞または複合名詞と，前後の前置詞，名詞を修飾する形容詞の組み合わせを学習する．例えば，英文(C)があった場合，名詞句[the solar system]は表5のように規則が学習される．

(C) Make a scale model of the solar system with this JavaScript enabled page.

表 5. 学習方法αによる形容詞，前置詞の学習

前置詞	冠詞	形容詞	名詞	前置詞
of	the	なし	solar system	with

学習方法βは，複合名詞にのみ適用される規則である．複合名詞は数が膨大であり，コーパス内に同じ複合名詞が存在しない可能性は多分にある．さらに，新しいシステムや手法を提案する場合，コーパス内に同じ複合名詞は存在しない．しかし，名詞句の最後にある名詞が同じであれば，その名詞句の概念はさほど違わない．つまり，冠詞の規則を一文内に限定すれば，付与される冠詞もさほど違わないと考えた．そのため，最後の名詞以外を省略して，規則を学習する．例えば，[Japanese Wordnet]と[English Wordnet]があれば，同じ[Wordnet]とまとめる．英文(C)の場合，表6のように規則が作られる．

表 6. 学習方法βによる形容詞，前置詞の学習

前置詞	冠詞	形容詞	名詞	前置詞
of	the	なし	system	with

参照を考慮する場合，主格か目的格かで限定される程度が大きく違う．つまり，theになる確率が異なる．そのため，単語自体を規則として学習するだけでなく，対象が主格か目的格か，つまり前後に動詞があるか，等の文の構造情報を利用する．具体的には，名詞の前後2単語の品詞を抽出し，冠詞を品詞の並びを学習する．英文(C)の場合，表7のように規則が作られる．

表 7. 学習方法γによる品詞情報の学習

二つ前	一つ前	冠詞	一つ後	二つ後
名詞	前置詞	the	前置詞	代名詞

しかし，学習を単語から品詞レベルに落としこむことで，意図しない規則を学習する可能性がある．また，表層的には同じでも意味が異なる名詞が複数回論文に出現すると，誤った規則を学習してしまう．

技術論文では，システムや手法，技術名が多く使われ，それらは複合名詞であることが多い．また，その論文特有，またはその論文で特徴的な表現は，参照が起きやすい．さらに，どの文書にも複数出現するような単語は複数の意味を内包することが多い．つまり，冠詞付与の対象と，学習する対象をその論文における重要語句に限定することで，規則の過学習を抑えることができるのではないかと考えた．

文書中の重要語句を抽出するために tf-idf を用いた．tf-idf は，単語の出現頻度である tf と出現頻度の逆数である idf の二つの指標で計算される．

$$\mathrm{tfidf} = \mathrm{tf} \cdot \mathrm{idf}$$

$$tf_i = \frac{n_i}{\sum_k n_k}$$

$$\mathrm{id}f_i = \log \frac{|D|}{|\{d: d \ni t_i\}|}$$

n_i は単語 i の出現頻度，$|D|$ は総ドキュメント数，$|\{d: d \ni t_i\}|$ は単語 i を含むドキュメント数である．そのため，idf は一種の一般語フィルタとして働き，多くのドキュメントに出現する一般的な語は重要度が下がり，特定のドキュメントにしか出現しない単語の重要度を上げる役割を果たす．今回は名詞句の中で重要な語句を抽出するために，n_i は名詞句中の単一名詞，または複合名詞 i の出現頻度，$|\{d: d \ni t_i\}|$ は名詞句中の単一名詞，または複合名詞 i を含むドキュメント数とした．また，tf-idf 値が上位 5%の名詞を重要語として抽出した．

3.3. 規則の適用

3.2 節で作成された規則を基に，適用フェーズでは対象名詞に付与すべき冠詞を決定する．処理の流れは，学習フェーズと同様にプロセス①，②を行い，次に，

③´ the の生起確率が the 以外の生起確率を上回るときに，the を付与する．

つまり，たとえ既出だと判断しても，the 以外の生起確率の方が高いときは the を付与しない．冠詞付与規則の適用例を図 2 に示す．

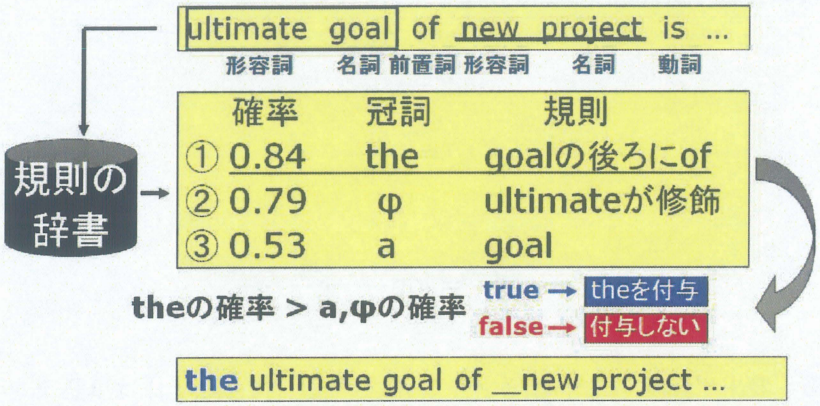

図 2. 冠詞付与規則の適用例

4. 評価実験

本章では，精度評価のための実験について述べる．4.1 節で実験の手順と条件について，4.2 節で評価の方法について，最後に 4.3 節で評価結果について述べる．

4.1. 実験手順

本実験は，以下の 3 つの手順で行う．

(1) 学習用コーパスから規則を学習する．

(2) (1)で作成した規則を使い，評価用コーパスに冠詞を付与する．

(3) 論文に付与されている冠詞を正解とし，(2)の評価を行う．

本研究では，論文に付与されている冠詞を正しいと仮定して，規則の学習と評価を行う．つまり，英語として信頼のできる論文を選択する必要がある．内容の質が高い論文誌ほど，査読時のチェックが厳しいために注意深く英文が記述されていると考え，インパクトファクター値の高い論文誌を選択した．インパクトファクター値とは，ある雑誌における 1 論文あたりの被引用回数の平均値を示し，その値が高いほど，影響力の高い論文を収録している論文誌であるといえる．もちろん，論文誌の質や影響力は，英文そのものの質に直接関連するものではない．しかし，論文の査読時に英文の質が悪いことを理由に拒否されることも考えられるため，投稿者は英文の質にも注意を払うことが推測できる．つまり，無作為に論文を選び出すよりは，質の良い学習データであると考えられる．

本実験では学習用の論文誌として，計算機科学分野の論文誌“Pattern Recognition”を用いた．学習用コーパスに 180 論文用い，この学習用コーパスに出現する単一名詞，または複合名詞について規則と冠詞の生起確率を学習した．学習するに当たり， Chunker である“OAK System[9]”を用いて，冠詞を付与すべき名詞句と単語ごとの品詞情報を取得した．続いて，評価用のコーパス 20 論文に Chunker を用いて名詞句，品詞情報を抽出し，学習した規則を用いて冠詞を付与した．the を付与すべき総名詞句数は 3，495 箇所である．

2.4 節で紹介した形容詞・前置詞・文脈を考慮した冠詞付与手法の内，定冠詞の付与結果のみをベースラインとした．

4.2. 評価方法

本実験では，3種類の尺度を用いて冠詞付与の性能を評価する．特にtheの付与性能を評価するために，theがどれだけ適切に付与できているかを比較する．評価用論文にtheが付与されている箇所の数をM，本手法によって theが付与された箇所の数をM_S，本手法によってtheが付与された箇所のうち評価用論文もtheが付与されていた数をM_Cとする．以下，これらの記号を用いて，3種類の尺度を定式化する．

1. theの再現率(Recall)

$$\text{Recall} = \frac{M_C}{M}$$

2. theの適合率 (Precision)

$$\text{Precision} = \frac{M_C}{M_S}$$

3. RecallとPrecisionの両方を考慮して，冠詞付与の総合的な性能を評価する．ここでは，文書検索システムなどでシステム評価に一般的に用いられている調和平均(F_measure)を使用する．

$$\text{F\_measure} = \frac{1}{\frac{1}{2}\left(\frac{1}{Recall} + \frac{1}{Precision}\right)}$$

4.3. 実験結果

学習によって作成された規則は冠詞それぞれについて生起確率が算出されている．この生起確率を目安に，適用する規則を選定する．そのため，生起確率が高い規則ほど付与する冠詞に偏りがあるため，信頼できる規則だと考えられる．本実験では，規則を適用する際に 0.1 刻みで閾値を設け，生起確率が閾値以上である規則のみを適用する．つまり閾値が 0.3 のとき，生起確率が 30%以上の規則を適用することになるため，作成された全ての規則を適用することができる．閾値が 0.9 のときは，生起確率が 90%以上の規則のみを適用するため，規則の適用数は下がることになる．評価結果の Recall, Precision, F-measure を図 4，図 5，図 6 に示す．表の横軸は閾値を表す．

閾値が 0.6 以下，つまり生起確率が 60%以下の低い規則も使用した場合，ベースラインに比べ precision を下げずに recall を 5〜8%上げることができた．これは，名詞をそのまま規則に用いるのではなく，品詞レベルに落としこむことで適用できる規則数が増えたことによる．また， f-value も 3.6%上昇していることから the の付与性能自体も向上したと言える．閾値が 0.8，0.9 の生起確率が高い規則のみを使用した場合も同様に，recall と f-value が共に向上した．

しかし，閾値が 0.7 のとき，適合率は 5.7%上昇しているが，再現率が 4.8%も低下してしまった．これは提案手法では，閾値が上がったことにより，付与できる冠詞が閾値 0.6 のときよりも大幅に減ったのに対し，従来手法ではそれほど減らなかったことを意味している．実際に規則を確認したところ，従来規則では，「同じ名詞で前回 the が付与されていた場合，次も the になる」という規則が多く適用されていた．つまり，この名詞は the になりやすいから前後の単語に関わらず the を付与する，という規則になる．そのため，precision が低い代わりに recall が高くなっていた．もちろん閾値が 0.6 以下でもこの規則は適用されているが，その場合，提案手法では他の規則により the を付与しているため再現率に差は見られなかった．

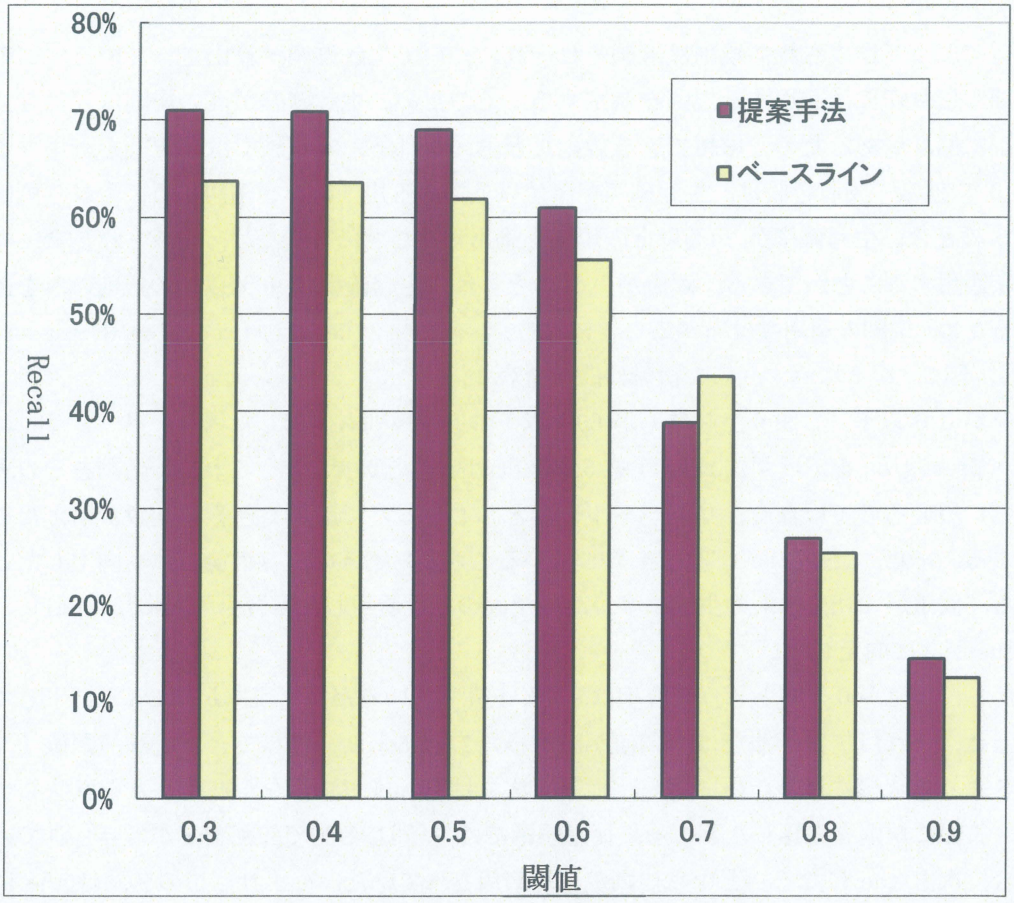

図 4. 提案手法とベースラインの Recall

閾値	proposed	baseline
0.3	71.0	63.7
0.4	70.8	63.6
0.5	69.0	61.9
0.6	60.9	55.5
0.7	38.8	43.5
0.8	26.9	25.3
0.9	14.4	12.4

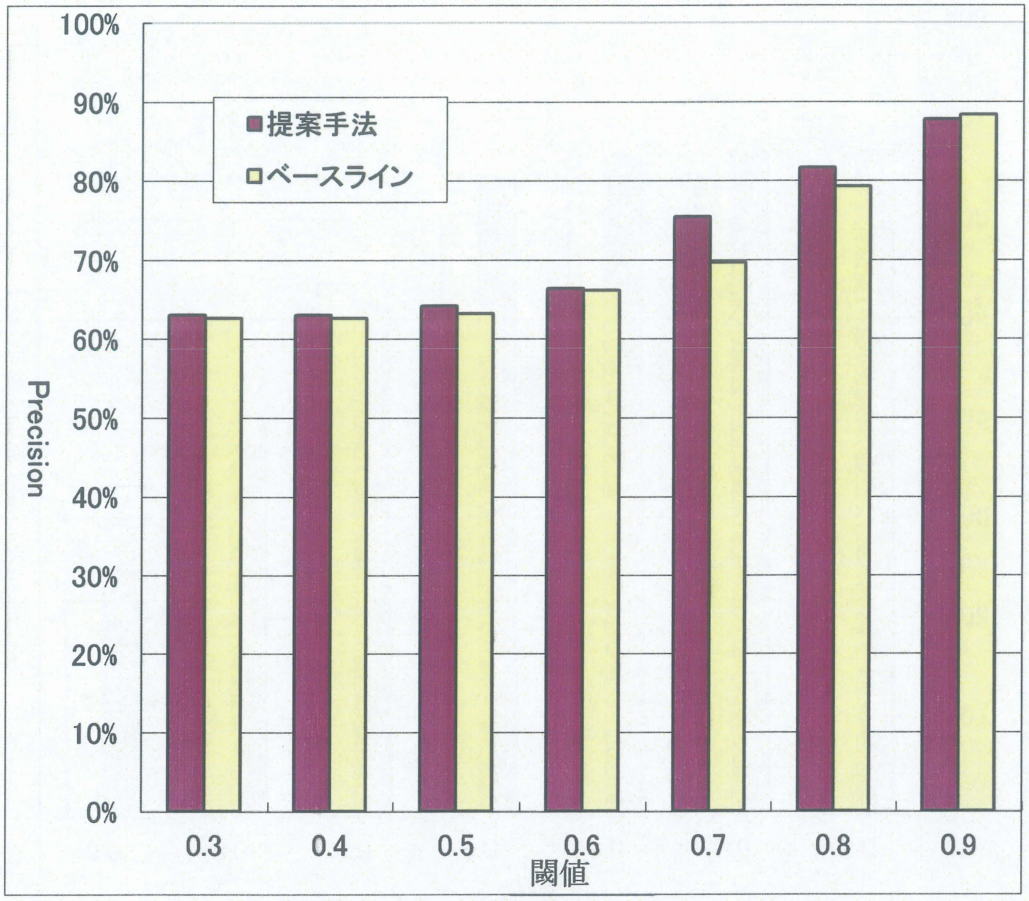

図 5. 提案手法とベースラインの Precision

閾値	proposed	baseline
0.3	63.0	62.6
0.4	63.0	62.6
0.5	64.2	63.3
0.6	66.4	66.1
0.7	75.4	69.7
0.8	81.6	79.3
0.9	87.8	88.4

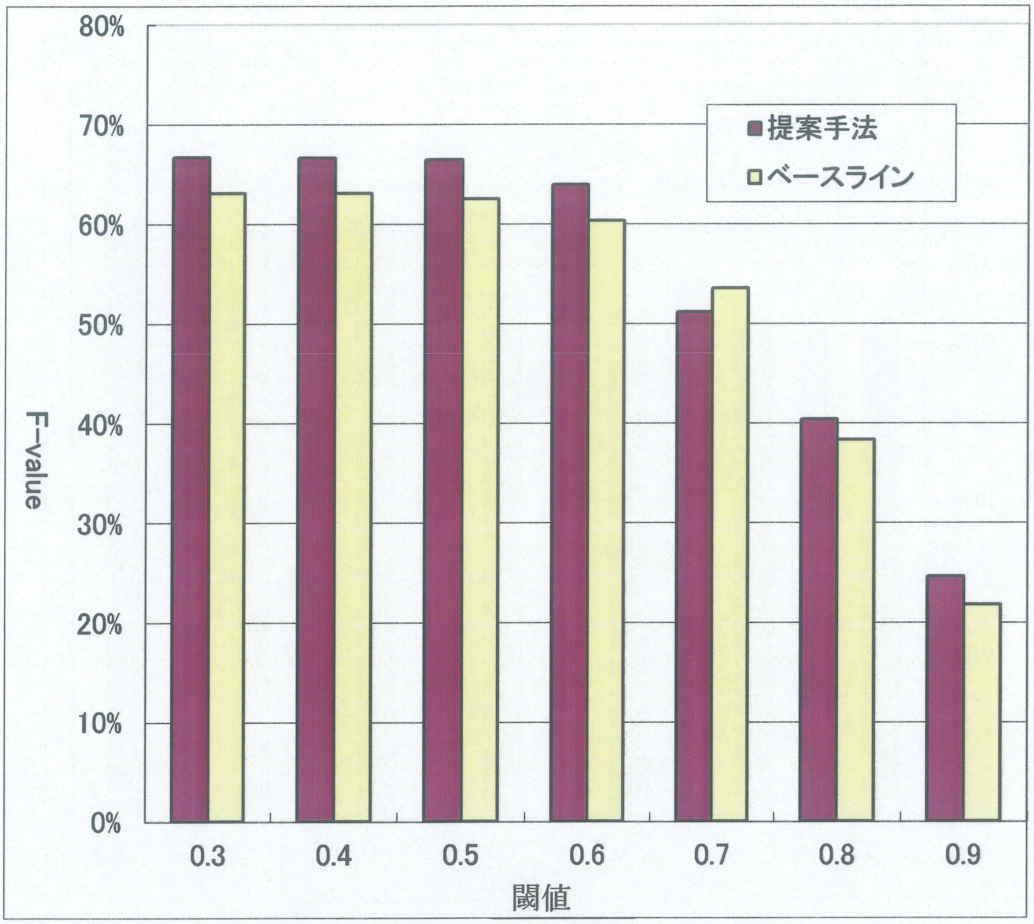

図 6. 提案手法とベースラインの F-value

閾値	proposed	baseline
0.3	66.7	63.1
0.4	66.7	63.1
0.5	66.5	62.6
0.6	64.0	60.4
0.7	51.2	53.6
0.8	40.4	38.4
0.9	24.7	21.8

5. 考察

5.1. 定冠詞の付与精度

全体的な評価としては，従来手法よりも定冠詞についての精度は向上したといえる．しかし，precision はほとんど上昇しなかった．これは，規則を抽象化する際に，過学習をしてしまっていることが考えられる．特に，本研究では既出の規則として品詞を利用した．その際に，全ての名詞を同一のものとして扱っている．しかし，名詞には固有名詞や一般名詞，また専門用語や一般用語といった種類が存在する．これらを同一で扱ったことによって，本来参照が起こらない名詞に対しても the を付与してしまっている可能性がある．また 3.1 節の調査でも，同様に全ての名詞を同一のものとして扱っている．また，今回の調査では名詞間の距離を行数で定義したため，同一の節内にあるか，同一の章内にあるか，等の論文の構成を無視してしまっている．

今後の課題としては，固有名詞と一般名詞，専門用語と一般用語，単一名詞と複合名詞，といった名詞の種類と参照の関係を人手で分析する必要がある．そして，同一の規則で扱える名詞の分類規則を見つける必要がある．また，論文はある程度形式が決まっているため，例えば提案手法を説明する章では，従来手法の説明で使われた単語が多いなどの知識を学習することで，より参照を意識した冠詞の付与ができるようになる．そこで，論文の構成と参照の関係を人手で分析する必要がある．また，実際に技術論文を分析したところ，１つの概念を複数の言い回しで表現していることがわかる．例えば，WordNet project at Princeton と Princeton Wordnet は同一の概念を指している．このような表現は技術名や特定の概念を指すときに使われるが，他の表現方法についてはさらなる分析が必要となる．さらに，Bridging reference という参照方法も存在する．意味的には同一の名詞を指しているが，参照元の名詞が省略されることによって，参照先と参照元の名詞が異なる場合である(e.g. I got into a taxi. The driver was a woman.)．これは，照応の分野でもあまり有効な手法が確立されていない．上位概念や関連辞書を利用することで，参照先と参照元の意味的な同一性判定を行う手法も研究されてはいるが，まだ実用的なところまできていない．

5.2 冠詞全体の精度

本論文では，定冠詞に注目し，付与精度の向上を図った．そのため，本手法のみでは冠詞の添削は成り立たない．例えば，2章で紹介した従来手法はそれぞれ別の観点から冠詞の付与，または誤り検出を行っている．特にイディオム手法は語句の並びをそのまま学習するため，冠詞の生起確率が高い規則のみを用いれば，通常の規則からは外れた慣用句のみを学習することも可能だろう．つまり，可算/不可算情報を使って，a かφか判断し，その上で本手法で定冠詞の付与，さらにイディオム手法で慣用句などの例外に対処することで，文法書に載っているような冠詞付与を順序立てて行うことができる．さらに，手法自体は独立な学習を行っているため，それぞれの手法を改良することで，全体の冠詞付与精度も同時に向上させることができる．

また，本論文においては，あらかじめ論文に付与されている冠詞のみを正しい冠詞であるとみなしている．しかし，「the」の省略など，複数の冠詞が正しいとされる場合も存在する．それにより，評価だけではなく学習の面でも適切な冠詞を選択できているとは言えない．また，冠詞の省略を考慮することで，性能の向上を図るだけでなく，より母語話者が行っている冠詞選択に近づけることができる．よって冠詞の「揺れ」を考慮した冠詞付与と評価方法の構築が今後の冠詞付与精度の向上には必要である．

謝辞

本研究を進め，論文を完成させるにあたり，日頃から多くのご指導，ご鞭撻を賜りました河合 敦夫准教授，井須尚紀教授，桝井文人准教授に深く感謝いたします．また，お忙しい中，副査をお引き受け頂いた大山航助教，事務関連で大変お世話になりました田中みゆき事務官，吉永みゆき事務官に深く感謝いたします．最後に，様々な助言をいただきました人工知能研究室の皆様に感謝いたします．

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] 鈴木英次, 科学英語のセンスを磨く, 化学同人, 京都, 1999.

[3] 井口達也,"統計モデルに基づく英文への冠詞付与手法に関する研究,"三重大学大学院修士論文, 2005.

[4] 若菜崇宏,"名詞の加算/不可算性を利用した英文の冠詞誤り検出に関する研究," 三重大学大学院 修士論文, 2007.

[5] 乙武 北斗, 荒木 健治, "単語出現状況の帰納的学習による英文冠詞誤りの検出及び自動校正手法", 電子情報通信学会論文誌 D, J90-D, 6, 1592--1601, 2007.6

[6] Knight, K. et al, "Automated postediting of documents". In Proceedings of the National Conference on Artificial Intelligence (AAAI), 1994.

[7] 平野孝佳, 平手勇宇, "検索エンジンを用いた英文冠詞誤りの検出"

[8] 長尾 真, 言語工学, 昭晃堂, 東京, 1983.

[9] "OAK System" Web Site, http://nlp.cs.nyu.edu/oak/.

[10] 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,"電子情報通信学会論文誌, Vol. J87-D-I No. 1, pp. 60-68, Jan. 2004.

[11] 宮井俊也 他, "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 14 回年次大会, 2008.03, 東京.

[12] Bond Francis, "Translating the Untranslatable A solution to the problem of generating English Determiners.", CLSI publications, Stanford, 2005.

[13] 迫村純男, James Raeside, "英語論文に使う表現文例集", 1996.