Vol.2017-NL-234 No.18
2017/12/21

前方文脈を考慮した冠詞推定における新聞記事コーパスの選択

宮野悠馬† 水井啓太† 河合敦夫†

概要：英文書中の冠詞推定手法として，正しく書かれた文書を学習させ，学習した冠詞周辺の情報を利用し，最大エントロピー分類器を用いて冠詞を推定する手法が高い精度を示している．しかしながら，このような従来手法では，定冠詞 the の推定精度が低いという問題点が未だ残されている．言語学的な観点から，冠詞推定に前方の文脈が考慮されていない事がひとつの原因である．そこで，我々は，従来手法に加え，定冠詞が付与された名詞の前方文脈に出現しやすい名詞（共起語）を the の決定要因として考慮する事で推定精度を向上させた．しかし，この共起語の中には，冠詞付与対象名詞と同一の主名詞，類義語など以外に，対象名詞からの関連語，連想語も含まれる．対象名詞からどのような単語を連想するかは，（書き手と読み手が属する）分野，時代背景，コミュニティにより異なる．これを近似的に実証するために，学習コーパスから，前後 1 ヶ月，半年，1 年，8 年の年月差のある記事群を各評価コーパスとし，推定精度の変化を検証した．この結果，推定精度の向上幅が，冠詞付与対象名詞により変化し，直近の前後 1 ヶ月が最も高くなる傾向が確認された．

キーワード：機械学習，冠詞推定，コーパス，英語

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, その文書内には誤りが含まれる事が多い. 特に，日本語のように冠詞の概念がない言語の話者においては冠詞の誤用が多く報告されている[1][2].

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[3][4]. それらの手法の中でも，特に冠詞周辺の情報を用いた最大エントロピー分類器による冠詞推定手法が高い性能を示している[5].

冠詞推定において，残された大きな問題の一つは，定冠詞 the の推定精度の低さである．定冠詞 the は，対象となる名詞句が，受け手にとって特定・限定される場合に付与される冠詞である．文献[6]において，この特定・限定される要因は 4 つに分類されている．また，その半分以上の割合を占める 2 つの要因は，前方文脈に含まれる情報である（2 章参照）．ここでいう前方文脈には，対象の名詞句が含まれる同一文内に加え，それより前の文章も含まれる．これに対し，多くの従来研究では，推定対象となる冠詞の周辺情報のみを用いて推定を行っている．冠詞の周辺情報とは，冠詞前後の数単語や，その品詞などであり，複数文に渡る文章は含まれない．したがって，複数文に渡る前方文脈を考慮していないことが, the の推定精度の低さの大きな要因である

従来研究[7]では，冠詞推定に前方文脈に含まれる特定・限定要因の一つである Coreferential（共参照）の考慮を試みている．Coreferential とは，同一の実体を表す名詞句が，前方文脈に出現する場合である（e.g. I caught a taxi. The taxi was red.）．しかし，結果として十分な the の推定精度向上には至っていない．その原因の一つとして，Coreferential の関係にある名詞句の決定において，文字列一致のみを用いていることが挙げられる．それにより，文字列一致しない同義語での言い換えや上位語などが考慮出来ない（e.g. I caught a taxi. The car was red.）．さらに，もう一つの前方文脈に含まれる要因である Bridging は，関連語により特定・限定するため，文字列一致では考慮出来ない（e.g. I caught a taxi. The driver was tired.）．

そこで竹内らは共起語を利用することで，前方文脈を効果的に考慮する手法の提案を行った[8]．この手法では 4 章で説明する共起語を利用することで，文字列一致に依存することなく the の決定要因となり得る名詞を機械的に学習し，冠詞推定する際の特徴量として扱う．これにより，Coreferential に加え，Bridging も含めた，前方文脈に存在する the の決定要因を一括して考慮でき，この共起語を，最大エントロピー分類器による冠詞推定手法で利用することで，前方文脈を考慮した冠詞推定を行えるようになった．

この手法で用いられる共起語の中には，冠詞付与対象名詞と同一の単語, 類義語などが, まず含まれる. (一般語の) 類似語や上位後への対処は，既存のシソーラス等を利用することができる．また，それを冠詞推定に利用した研究も存在する. また, ConceptNet 5 等の汎用的な知識を用いて，Bridging の関連・連想語を捉える試みもある．しかし，それらの効果は十分ではないとの報告もある[9]．この理由として考えられるのが, the が付与される名詞をどのような単語から連想するかは，文書が属する分野，時代背景，（書き手，読み手が属する）コミュニティにより異なることである．言い換えれば，その連想範囲は，汎用的共通知識ではカバーできない．そのため，共起語を用いた手法では，どのような分野，時代背景，コミュニティの文書を学習させるかによって, 推定精度が異なってくる場合と予想される．

そこで，本稿ではこれらの中から，まずは，実現の容易

† 三重大学大学院工学研究科,津市
Graduate School of Engineering, Mie University, 1577 Kurimamachiya, Tsu-shi,514-8507 Japan

性から，時代背景による推定精度の変化を近似的に実証した．具体的には，同じ学習コーパスに対して，年月の異なる評価コーパスを用いることで，推定精度がどのように変化するかの検証実験を行った．

2. 定冠詞 the の用法

定冠詞 the は，対象の名詞句が受け手にとって特定・限定される場合に付与される冠詞である．文献[6]では，名詞句が特定・限定される定冠詞 the の用法は，大きく分けて 4 つに分類されると述べられている．この 4 つの分類とは，Coreferential（44%），Bridging（8.5%），Larger situation（23.5%），Unfamiliar（21%）である．ここで，それぞれの分類の後ろの括弧は，the Wall Street Journal 誌において，人手により集計された各用法の頻度を表しており，残りの 3%は明確に分類出来なかったものである．

このうち，Larger situation は，文書内で初めて出てきた名詞句であっても，多くの受け手にとって既知の実体を指し，the が付与されるものである（e.g. the earth, the sun）．この分類には，固有名詞など，常にその名詞句が表す実体を一意に特定できるものが属する．

Unfamiliar は，同一文内での修飾等により特定・限定される場合である（e.g. The taxi I caught yesterday was red.）．これら 2 分類は，いずれも冠詞周辺情報のみで特定・限定出来る用法である．したがって，従来の冠詞推定手法で既に考慮されている用法である．

一方，Coreferential は，推定対象の名詞句と同一の実体を表す名詞句が，前方文脈に存在する場合に the が付与されるものである（e.g. I caught a taxi. The taxi was red.）．同一の実体を表す名詞句には，文字列が完全に一致する場合を含め，一部省略や，文字列が一致しない同義語，または上位語での言い換えが存在する．英文書においては，同じ名詞句の繰り返しは避けられる傾向にあるため，一部省略（e.g. I caught a privately-owned taxi. The taxi was red.）や，言い換え（e.g. I caught a taxi. The car was red.）が多いと考えられる．

Bridging は，名詞句が指す実体を容易に特定できるような関連語が，前方文脈に存在する場合であり，Associative とも呼ばれる（e.g. I caught a taxi. The driver was tired.）．例文の”driver”は，前方文脈に存在する”taxi”の”driver”であることが容易に特定できるため，the が付与されている．このような関連語は，単純な文字列一致では得られないため，従来の冠詞推定手法では全く考慮出来ない分類である．これら 2 分類を考慮するには，前方文脈の情報が不可欠であり，共起語を用いた手法ではこれらの考慮を行うことが可能である．

3. 最大エントロピー分類器による冠詞推定

本章では，最大エントロピー（ME）分類器による冠詞推定手法について述べる．この手法では，英語新聞などの文法的に正しく書かれた文書に含まれる冠詞の用法を学習する．学習には，名詞句や名詞句周辺の文脈情報から得られる素性ベクトルを用いる．また，冠詞推定は，{a/an, the, ϕ} に対して行われる．ここで ϕ は無冠詞を表している．

本稿の目的は，共起語を用いた手法の際に学習データの時代による推定精度の変化を検証することであり，冠詞の用法上，共起語を用いたことによる {a/an, ϕ} の使い分けには影響を与えない．したがって，文献[8]と同様に，ME 分類器を用いて the と {a/an, ϕ} の 2 値分類を行う．以降，{a/an, ϕ} は，まとめて other と表記とする．この 2 値分類を行うために用いられる素性を図 1 に示す．図 1 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．

(例文)

…. This is the current_oil_prices in Japan.

カテゴリ	素性名	素性値
対象名詞句	主名詞	prices
	主名詞 POS	NNS
	名詞句 部分文字列	current_oil_prices oil_prices
	主名詞以外の名詞	oil
	修飾語	current
	修飾語 POS	JJ
	所有格を含む	No
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
	単名詞+前置詞	–
名詞句の後	句の分類	PP
	名詞句直後の語	in
	上の POS	IN
修飾句	句の分類	NP
	主名詞	Japan
	主名詞以外の名詞	–
	修飾語	–
	修飾語 POS	–

図 1 冠詞周辺情報の素性リスト

4. 共起語手法

本章では，従来の ME 分類器による冠詞推定に加え，前方文脈を考慮するための手法について述べる．まず，4.1 節

で，共起語手法の手法についての基本的な方針を述べ，4.2 節で，共起語手法で用いられる共起語リストの概要について述べる．そして，4.3 節では共起語リストの作成方法について述べ，4.4 節で従来手法への適用方法について述べる．

4.1 基本方針

共起語手法では，冠詞が the である名詞句の前方文脈中から，the の要因となる可能性が高い名詞を抽出する．抽出した名詞を the の要因候補とし，前方文脈の考慮を行う．候補の出現条件は必要最低限にとどめ，より多くの候補を抽出する．これにより，出現パターンの異なる the の決定要因を一括して抽出する．

4.2 共起語リスト

前方文脈に存在する the の決定要因候補を効率的に抽出するために，共起語リストを作成する．共起語リストとは，冠詞推定対象名詞の冠詞が the の場合に，共起確率が高い名詞からなるリストである．共起確率を利用することで，推定対象名詞句の前方文脈中に出現する語の集合から，the の決定に寄与する可能性がある名詞を抽出する．

ここで，共起語手法で用いる共起確率と共起語の定義を行う．この手法で対象としている冠詞は，the と other (a/an, ϕ)である．冠詞推定対象主名詞 x の冠詞が the である場合と，x の前方文脈中に出現する名詞 y との共起確率 $C_p(the|x,y)$は(1)式で求める．

$$C_p(the|x,y) = \frac{f((the|x) \cap y)}{f(x \cap y)} \qquad (1)$$

ここで，f（$x \cap y$）は，x,y が同時に出現する頻度を表し，$f((the|x) \cap y)$ は，x の冠詞が the である場合に y と同時に出現する頻度を表す．

$C_p(the|x,y)$ が，推定対象主名詞 x の the の冠詞生起確率 $p(the|x)$ 以上となれば y を共起語とする． $p(the|x)$ は(2)式で求められ，other の冠詞生起確率 $p(other|x)$ もまた(3)式で求められる．

$$p(the|x) = \frac{f(the|x)}{f(x)} \qquad (2)$$

$$p(other|x) = \frac{f(other|x)}{f(x)} \qquad (3)$$

また，共起語の信頼性をある程度確保するため，$f(x \cap y) \geqq 10$ とする．以上の条件をまとめた(4)式を共起語条件式として用いる．

$$\{ C_p(the|x,y) \geqq p(the|x) \} \wedge \{ f(x \cap y) \geqq 10 \} \qquad (4)$$

以降，共起語と表記した場合は，すべてこの the との共起語を表す．

4.3 共起語リストの作成

共起語の抽出は，学習用コーパスに含まれる，全ての冠詞推定対象名詞句の主名詞に対して行う．共起語リストは，主名詞ごとに作成し，冠詞推定に利用する．以降，冠詞推定対象名詞句を対象名詞句，その主名詞を対象主名詞と呼ぶ．

共起語リストの作成過程を対象主名詞 x が，“conference”の場合を例にして説明する．まず，学習コーパス中において，“conference”を主名詞とする対象名詞句から前方 i （≧0） 文前までを抜き出す．この i を抽出範囲と呼び，抜き出した文章を対象抽出文章と呼ぶ．特に，$i = 0$ の場合は，対象名詞句が含まれる文内（但し，対象名詞句より前のみ）を対象抽出文章とする．また，対象抽出文章中に対象主名詞が存在する場合は，その文中の対象主名詞より前の単語と，それより前の文は抜き出さない．

例として，$i = 2$ の場合の対象抽出文 $d_{1,2,3}$ を図 2 に示す．図 2 の d_2 と d_3 は連続した文章であり，d_3 では，1 文前で d_2 の対象主名詞と一致するため，それより前の“OPEC”や 2 文前は抜き出さない

次に，図 2 の対象抽出文章中に含まれる名詞句をすべて抽出する．但し，図 2 中の抽出文章 d_1 の“price”ように，同一文章内に 2 回以上現れる名詞は 1 回目のみを対象とする．

図 2 の $d_{1,2,3}$ から共起語候補を抽出すると，表 1 になる．ここでは，抜き出した名詞句中に含まれる名詞以外（形容詞など）の語を除去し，共起語候補となる名詞を抽出している．名詞句が 2 語以上の名詞からなる場合は，すべての名詞を結合し 1 つとする場合，すべてを別々にする場合の両方で抽出する．したがって，図 2 の d_3 で抜き出した名詞句“next press conference”は，表 1 では“press_conference”，“press”，“conference”となる．

得られた共起語候補を集計し，それぞれの出現頻度，対象主名詞“conference”の冠詞が the の場合との共起回数，その共起確率を計算する．“conference”が学習コーパス内に 100 回出現時の結果を表 2 に示す．

そして，この中から共起語条件式(4)を満たすものを抽出し，“conference”の共起語リストを作成する．作成された共起語リストを表 3 に示す．$p(the|\text{conference}) = 75\%$とすると，共起語条件式より，表 3 の中からは 3 語が共起語となる．

以上の工程を，学習コーパス内の全対象主名詞について行い，それぞれの共起語リストを作成する．

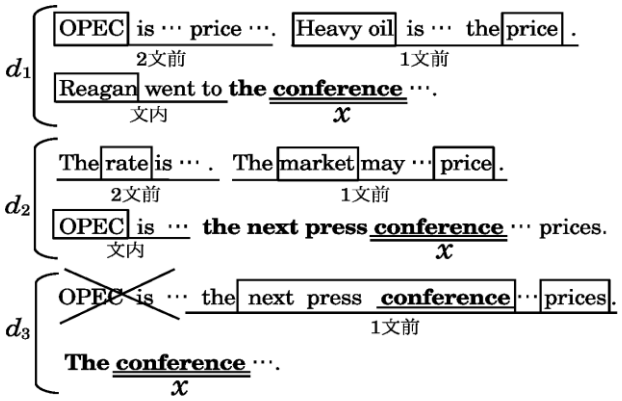


図 2 対象抽出文章の抜き出し例（i = 2）

表 1 抽出された共起語候補

対象抽出文章	共起語候補
d_1	Reagan, oil, price, OPEC
d_2	OPEC, price, market, rate
d_3	prices, press_conference, press, conference

表 2 共起語候補の集計結果（x=conference）

共起語候補 y	$f\,(x \cap y)$	$f\,((the \mid x) \cap y)$	$C_p(the\|x,y)$
oil	62 回	22 回	35%
market	50 回	20 回	40%
price	29 回	26 回	90%
prices	10 回	6 回	60%
OPEC	45 回	34 回	76%
Reagan	2 回	1 回	50%
press_conference	5 回	5 回	100%
press	14 回	8 回	57%
conference	42 回	41 回	98%
⋮	⋮	⋮	⋮

表 3 conference の共起語リスト

共起語 y	$f\,(x \cap y)$	$C_p(the\|x,y)$
conference	42 回	98%
price	29 回	90%
OPEC	45 回	76%
⋮	⋮	⋮

4.4 共起語リストの作成

前節で作成した共起語リストを，ME 分類器による冠詞推定に適用する方法について述べる．図 3 に，冠詞推定対象名詞句が“conference”である時の冠詞推定の流れを示す．図 3 において，“<ART>”は冠詞推定対象箇所である．冠詞推定は，以下の STEP1～4 で行う．

STEP1 従来の冠詞推定手法（3 章）と同様に，冠詞周辺情報の素性を事例に追加．

STEP2 対象名詞句前方の抽出範囲 i から共起語候補名詞を抽出．

STEP3 対象主名詞の共起語リストを参照し，得られた共起語候補名詞と一致すれば，その名詞を素性として事例に追加．

STEP4 作成事例を ME 分類器に与え，冠詞推定．

STEP3 で共起語を素性として事例に追加する際は，他の主名詞の共起語と区別するため，対象主名詞と共起語を連結して素性として与える．図 3 の例では，対象主名詞“conference”の共起語“price”を事例に追加する際に，“conference_price”としている

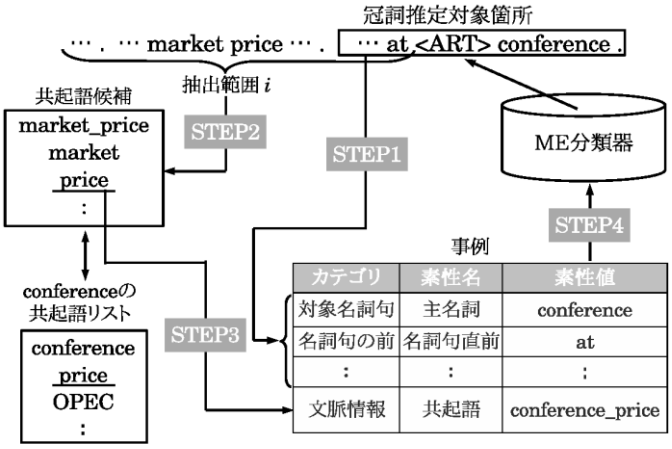


図 3 共起語リストを利用した ME 分類器による冠詞推定

5. 評価実験

本章では，文内情報のみを用いる従来手法と，共起語リストにより前方文脈を考慮した共起語手法を用いて各年月別評価コーパスの推定性能にどのような変化が生じるかの検証実験について述べる．

5.1 実験データ

本実験では，学習用および評価用コーパスとして，アメリカの AP 通信社が配信しているニュース記事（以降 Apw と呼ぶ）を用いた．Apw には，1994 年 11 月から 2010 年 12 月の期間の新聞記事が存在するが，その中から，2002 年 11 月の記事群を学習用コーパスとして，そこから前後 1 ヶ月（2002 年 10 月,2002 年 12 月），前後半年（2002 年 5 月,2003 年 5 月），前後 1 年（2001 年 11 月,2003 年 11 月），前後 8 年（1994 年 11 月,2010 年 11 月）の記事群をそれぞれ評価用コーパスとして用いた．また，前方文脈のつながりを確保するために，それらの記事の内，表や単語の羅列など本文以外の本文以外の要素が含まれている記事を除外した．

実際に使用する各コーパス内に含まれる記事数と冠詞推定対象箇所（a/an, the, ϕ）数を表 4 に示す.

5.3,5.4 節では，事前実験として，共起語手法の有効性を示すために, 従来手法と共起語手法の他に 2 手法を用いて，どの手法がより冠詞推定に効果的であるかの比較実験を行った．また，この実験では表 4 の評価コーパスを全て合わせた結果を用いて，比較検証を行った.

5.5,5.6 節では，事前実験で推定精度向上により効果的であると確認できた手法と従来手法を用いて，各年月別評価コーパスの推定性能にどのような変化が生じるかの検証実験を行った.

本研究で用いる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 1 つである Classias[12]の L2 正則化ロジスティク回帰モデルを用いた．また，チャンキングと素性値として利用する品詞タグ付けを行うソフトとして，OAK System[13]を用いている.

表 4 コーパス情報

	年'月	記事数	冠詞推定箇所数
学習	2002'11	16,883	735,867
評価	1994'11	8,731	368,703
	2001'11	18,236	833,960
	2002'05	20,164	945,338
	2002'10	18,467	826,748
	2002'12	13,198	569,775
	2003'05	17,787	811,170
	2003'11	17,687	803,884
	2010'11	17,789	764,636

5.2 評価方法

本実験では，各コーパス内に文法的誤りは含まれないと仮定し，推定された冠詞とコーパス内の冠詞が，一致しているかどうかを評価する．冠詞推定においては，推定数を減らしてでも誤った推定を減らしたい場合が多い. そこで, 推定結果が信頼できるものであるかどうかの決定を行う指標として，Classias から出力される判定結果のスコアに対し，閾値 θ（≧0）を設定する．Classias から出力されるスコアは, 絶対値が高いほど推定の信頼度が高いことを示し，正の数であれば the，負の数であれば other の推定結果を表す．ここで，スコアの絶対値が設定された閾値 θ より小さいものであれば，その推定結果を採用せず，未推定ということにする.

本実験では評価指標として，Recall ，Precision ，F-measure を用いる．冠詞 *ART*(∈｛the,other,all｝)に対する Recall(R_{ART}), Precision(P_{ART}), F-measure(F-measure$_{ART}$)を(5), (6),(7) 式で定義する．ここで，*ART*=all は，the と other の両方を対象にする場合を表す.

$$R_{ART} = \frac{\text{正しく}ART\text{と推定された数}}{ART\text{と推定すべき総数}} \quad (5)$$

$$P_{ART} = \frac{\text{正しく}ART\text{と推定された数}}{ART\text{と推定した数}} \quad (6)$$

$$F\text{-}measure_{ART} = \frac{2*R_{ART}*P_{ART}}{R_{ART}+P_{ART}} \quad (7)$$

5.3 事前実験

共起語手法では，4 章で定義した共起語条件（4）を用いることにより，対象名詞句前方の文章に存在する名詞を前方文脈素性として考慮する．この時に，4 節で定義した共起語条件を用いた手法（共起語手法），対象名詞句と一致した単語のみを共起語として適用した手法（主名詞一致手法），対象名詞句の前方文脈に存在する全ての名詞句を共起語として適用する手法（全前方名詞手法）の 3 つの共起語の適用方法が存在する.

主名詞一致手法とは，前方文脈において，推定対象主名詞と文字列一致する主名詞のみを冠詞推定の素性として用いる手法である．前方文章中の名詞を抽出する方法は，4 章の共起語候補の抽出方法と同様である．抽出した共起語候補から，推定対象主名詞と文字列一致する主名詞のみを共起語リストに追加する．これにより 2 章で説明した Coreferential の可能性がある名詞のみに的を絞った考慮を行うことが出来る.

全前方名詞手法とは，対象主名詞の前方文脈中に存在する名詞を全て素性として用いる手法である．4 章の共起語候補の抽出と同様に名詞を抽出し，全て無条件に共起語リストに追加し，適用する手法である.

これら 2 手法と従来手法，共起語手法を用いて推定精度を比較し，どの手法が本研究で用いる評価コーパスにより有効であるかを比較検証する.

また，この比較実験の評価指標の値は全評価コーパスに対する値であり，より推定精度の高い手法と従来手法を 6.5 節以降の比較実験で用いる手法とする.

5.4 事前実験結果

表 5 に，θ＝1 における従来手法と提案手法の冠詞推定結果を示す．表 5 より，従来手法で顕著なのが R_{the} の低さである．その R_{the} に対して，共起語手法では従来手法と比較して 5.97 ポイント改善されている．また，P_{the}，R_{other}，R_{all} も改善されており，推定結果が the に偏り過ぎることなく効果的に the の推定精度が向上していることが分かる．比較手法である主名詞一致，全前方名詞の 2 手法については従来手法と比較すると R_{all} は向上しているが，共起語手法よりも向上幅は低く, P_{all} は共起語手法よりも低下している.

次に，the の推定について Precision が同じ値の場合の Recall を比較する．冠詞誤り訂正を行う上で，高い

Precision を設定して，どれだけ Recall を維持できるかは重要な指標となる．図 4 に各手法の the に対する Precision-Recall 曲線の比較結果を示す．閾値 θ を，0≦θ≦1 の範囲で 0.2 刻みに変化させ，各結果を比較した．一般的に Precision-Recall 曲線がより右上に位置するほどシステム全体の精度が高いと言える．したがって図 4 より，推定精度は，

従来手法 ≒ 全前方名詞 ＜ 主名詞一致 ＜ 共起語手法

の順に高いことが視覚的に理解できる．

これらの結果から，共起語手法は，the の推定精度向上に効果的であることが確認できた．

表 5 手法別冠詞推定結果（θ = 1）

評価値	従来手法	主名詞一致	全前方名詞	共起語手法
R_{the}(%)	48.8	51.0	49.9	54.7
P_{the}(%)	79.0	79.8	78.8	79.5
R_{other}(%)	82.8	83.0	83.2	83.8
P_{other}(%)	97.3	96.1	96.1	96.4
R_{all}(%)	73.4	74.1	74.0	75.7
P_{all}(%)	93.3	92.1	91.9	92.5

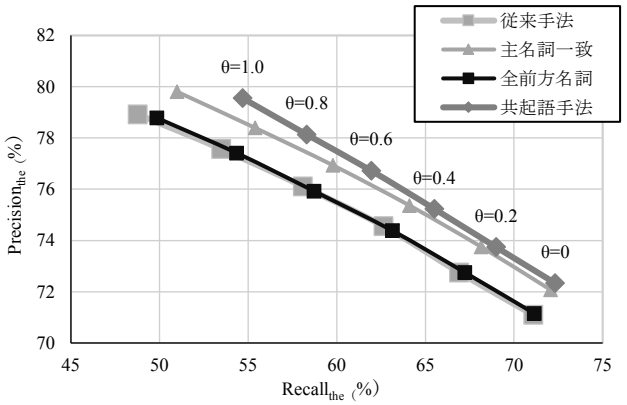


図 4 the に対する推定性能の Precision-Recall 曲線

5.5 実験

事前実験では共起語手法が冠詞推定精度の向上に効果的であることが確認できた．しかし，共起語の中には，冠詞付与対象名詞と同一の単語，類義語など以外に，対象名詞からの関連語，連想語も含まれる．対象名詞からどのような単語を連想するかは，文書が属する分野，時代背景，コミュニティにより異なる．これを近似的に実証するために，学習コーパスから，前後 1 ヶ月，半年，1 年，8 年の年月差のある記事群を各評価コーパスとし，推定精度の変化を検証した．学習コーパスと各評価コーパスの情報は表 4 に示す．

5.6 結果

図 5 に θ = 1 とした時の従来手法，共起語手法をそれぞれ用いた際の各評価コーパスの $F\text{-measure}_{all}$ を示す．図 5 より，どちらの手法においても，学習用コーパスから年月が最も離れた評価コーパスの推定精度は低く出ているが，近いからといって推定精度が高いとは限らないことが分かった．しかし，従来手法と共起語手法の $F\text{-measure}_{all}$ を比較すると，どの年月においても共起語手法の方が，精度が高い事が分かる．

次に，図 5 の共起語手法を用いたことによる $F\text{-measure}_{all}$ の向上幅に着目し，共起語手法 $F\text{-measure}_{all}$ と従来手法 $F\text{-measure}_{all}$ の差を図 6 に示した．図 6 より，学習コーパスから年月が近い評価コーパスほど，共起語手法を用いたことによる推定精度の向上幅が大きくなることが確認できた．

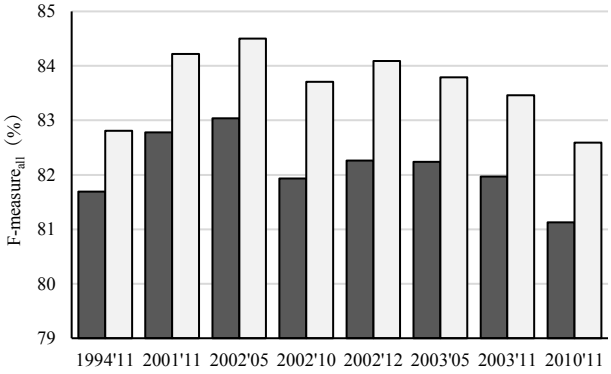


図 5 従来・共起語手法の $F\text{-measure}_{all}$

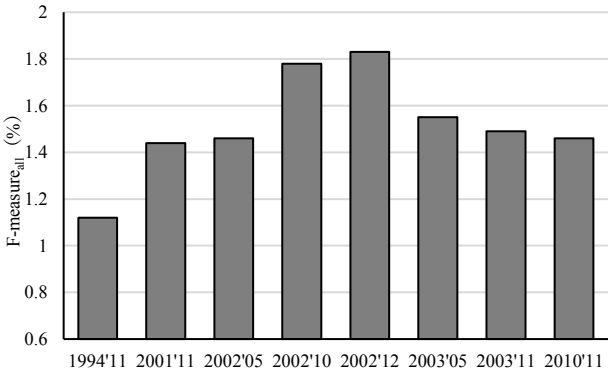


図 6 共起語手法 $F\text{-measure}_{all}$ と従来手法 $F\text{-measure}_{all}$ の差

6. 考察

本章では，5. 4 節または 5. 6 節の実験結果について考察を行う．

6.1 最大冠詞生起確率

共起語手法の効果を，4 章で定義した冠詞推定対象主名詞の最大冠詞生起確率 p（x_{max}）の観点から考える．この p（x_{max}）には，the の冠詞生起確率 $p(the|x)$，または，other の冠詞生起確率 $p(other|x)$ の内，どちらか高い方が採用さ

れる．閾値を θ = 1 とした時において，従来手法と比較し，推定正解数が増加，不変，減少した対象主名詞を p（x_{max}）ごとに集計した結果を図 7 に示す．増減は，主名詞の異なり数で表しており，共起語リストが作成された主名詞のみを対象としている．図 7 から，共起語手法は，p(x_{max})が小さい単語，つまり冠詞生起確率の偏りが小さい主名詞に対して，特に有効であることが分かる．それに対して，p(x_{max})が大きい，つまり冠詞生起確率の偏りが大きい主名詞ほど，正解増加数の割合が小さくなっていく．

この結果を推定精度について検証するために，p(x_{max})が 80%以上を偏り大，80%より小さければ偏り小として主名詞を分類し，それぞれを従来手法と比較した結果を表 6 に示す．表 6 から，従来手法において，偏り小と偏り大を比較すると，偏り大と比べ，偏り小の推定精度が明らかに劣っていることが確認できる．また，共起語手法による効果に注目すると，偏り大はわずかながら R_{all}，P_{all} が低下するのに対して，偏り小は P_{all} が 1.8 ポイント程低下したが，R_{all} は 7.9 ポイントと大幅に改善された．

次に，従来手法と共起語手法を用いた際の偏り小，偏り大の Precision-Recall 曲線の比較結果をそれぞれ図 8,図 9 に示す．閾値 θ を，0≦θ≦1 の範囲で 0.2 刻みに変化させ，各結果を比較した．図 8 より，偏り小のみを対象とした場合，共起語手法を用いたほうが推定精度は高くなり，図 9 より，偏り大のみを対象とした場合は，従来手法を用いたほうが推定精度は高くなることが分かる．

これらの結果より，冠詞生起確率の偏りが小さい主名詞の冠詞決定には，共起語リストの適用が有効であり，前方文脈の考慮の必要性が高いと言える．偏りが大きい主名詞は the の推定精度については向上しているが，other の推定精度と全体的な推定精度はわずかながら低下している．この原因としては，other に偏りが大きい主名詞はほとんど冠詞が the にならないため，共起語リストがない，または少数しか存在せず，あまり有効的に共起語リストを活用できなかったためだと考えられる．

表 6 主名詞における p(x_{max})の偏り別の
従来手法と共起語手法の性能比較（θ = 1）

	偏り小		偏り大	
	従来手法	共起語手法	従来手法	共起語手法
R_{the}(%)	36.8	46.9	66.7	67.9
P_{the}(%)	80.4	80.7	81.0	81.6
R_{other}(%)	44.3	50.0	94.2	93.8
P_{other}(%)	87.3	83.8	98.6	98.5
R_{all}(%)	40.7	48.5	88.8	88.8
P_{all}(%)	84.2	82.3	95.6	95.5

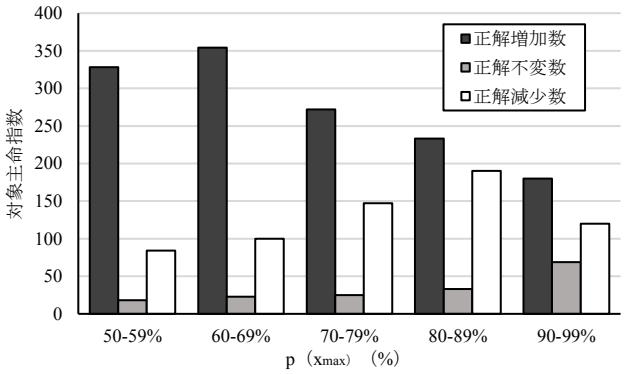


図 7 p(x_{max})ごとの主名詞別正解数の増減（θ = 1）

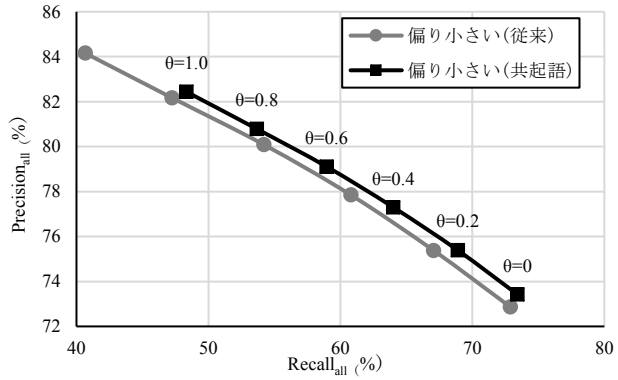


図 8 主名詞偏り小さい場合の
従来手法と共起語手法の推定性能 Precision-Recall 曲線

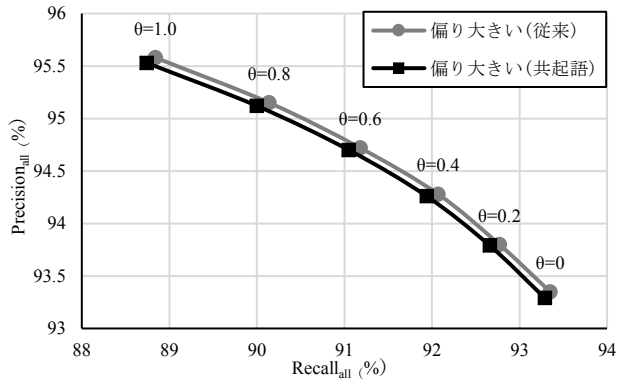


図 9 主名詞偏り大きい場合の
従来手法と共起語手法の推定性能 Precision-Recall 曲線

6.2 共起語一致率

表 7 は 5.5 節の実験で作成された学習コーパスの共起語リスト内の全共起語の数と学習コーパス同様に学習した場合に作成される各評価コーパスの共起語リスト内の全共起語の数を示している．学習コーパスの共起語と各評価コーパスの共起語がどれ位一致しているかを指す評価指標として，この一致率を共起語一致率として，(8)式で定義する．

$$\text{共起語一致率} = \frac{\Sigma(2\ \text{コーパス間で一致したある単語の共起語数})}{\Sigma(2\ \text{コーパス間のある単語の全共起語数})} \quad (8)$$

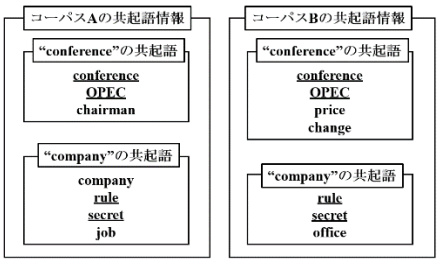
共起語一致率の具体例として図 10 のように算出される．図 10 では，コーパス A,B 内での”conference;”company”の

共起語をそれぞれ示している．この時，A,B の”conference”の一致した共起語として下線の 2 単語，総共起語数として 5 種類の単語が取られている．”company”についても同様に一致した共起語として下線の 2 単語，総共起語数として 5 種類の単語が取られており，これらを(8)式に当てはめると共起語一致率=0.4 と算出される．

図 11 は学習コーパスの共起語と各評価コーパスの共起語一致率を示している．図 11 より，学習用記事から前後 1 年以内の記事ではあまり差は見られなかったが，前後 8 年の記事は他と比較して低くなっている．また，図 11 と 5.6 節の従来手法と共起語手法の $F\text{-measure}_{all}$ の差を示している図 6 を比較すると，ある程度の相関があることが考えられる．

表 7 各コーパスの総共起語数

	年’月	総共起語数
学習	2002’11	185,914
評価	1994’11	87,849
	2001’11	207,739
	2002’05	241,083
	2002’10	225,516
	2002’12	135,781
	2003’05	204,566
	2003’11	219,239
	2010’11	204,394



$$\text{共起語一致率} = \frac{(\text{A, Bで一致した”conference”の共起語数}) + (\text{A, Bで一致した”company”の共起語数})}{(\text{A, Bでの”conference”の総共起語数}) + (\text{A, Bでの”company”の総共起語数})} = \frac{2+2}{5+5} = 0.4$$

図 10 共起語一致率算出法

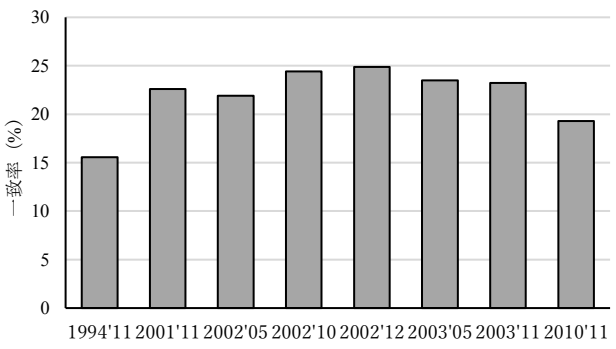


図 11 共起語一致率

6.3 共起語

本節では、作成された共起語リストのうち、冠詞 the，other 共に最低 100 回以上学習し，学習・評価コーパス共に 100 回以上出現する主名詞 56 単語を対象に考察を行う。この 56 単語のうち、45 単語は全てのコーパスで平均した推定精度が向上し、11 単語については推定精度が低下した。

向上した主名詞で、特徴のある推移を示した主名詞を図 12 に示す。図 12 の主名詞のうち，

”president”(p(the)=45%，学習回数 1638 回)

の共起語リストについて考察する．”president”を選択した理由としては，推定精度が向上しただけでなく，アメリカ大統領として

1993 年 1 月-2001 年 1 月までが”Clinton”
2001 年 1 月-2009 年 1 月までが”Bush”
2009 年 1 月-2017 年 1 月までが”Obama”

というように，年月により大統領が異なっており，本研究の趣旨にも適しているためである．

まず，”president”の推定精度が向上した要因として，p(other)=55%なのにも関わらず，学習された”president”の全共起語 1200 単語のうち，”president”の前方文脈に出現した時に，”president”に the が付与されていた確率が 100%の単語が 463 単語，90%-99%の単語が 558 単語と，the の決定要因となりやすい共起語を多く学習出来たことで推定精度が向上したと考えられる．

図 12 は，各評価コーパスの”president”の共起語手法による $F\text{-measure}_{all}$ の向上幅を示している．図 12 より，直近 1 年（2001 年 11 月-2003 年 11 月）と 8 年前（1994 年 11 月），8 年後（2010 年 11 月）の $F\text{-measure}_{all}$ を比較すると，

直近 1 年 ＞ 8 年前 ＞ 8 年後

の順に高くなっている．このようになった理由として，学習時点（2002 年 11 月）では，大統領は”Bush”であり，前方文脈に名詞”Bush”が出現した場合，”Clinton”が前方文脈に出現した場合よりも”president”に the が付与されやすい傾向があったためと考えられる．また，他の年月と比較して，2010 年 11 月の値が低くなっているのは，この時点で，大統領は”Obama”に変わり，学習時点ではまだ大統領になっていないため共起語として学習していなかった事が原因として考えられる．

”president”の共起語”Clinton”,”Bush”,”Obama”の 2002 年 11 月の学習コーパスでの重みを調査したところ，”Clinton”の重みが 0.13（学習回数 20 回），”Bush”が 0.97（学習回数 335 回），”Obama”が無しであり，”president”の前方に”Bush”が出現することで the を付与されやすくなっている事が分かった．

反対に推定精度が低下した主名詞として

”move”(p(the)=55%，学習回数 506 回)

が存在する．推定精度が低下した理由としては，学習された”move”の全共起語 538 単語のうち，the が付与されてい

た確率が 100%の単語が 3 単語，90%-99%の単語が 11 単語と, the の決定要因となりやすい共起語を今回の共起語手法では捉え切れなかった事が 1 番の原因であると考えられる。また，共起語を上手く学習できなかった要因として，”president”とは異なり, ”move”の意味には”動き”や”行動”といった，状況によって”move”が指す対象が異なることが多いためと考えられる．また，推定精度が低下した主名詞には従来手法で元から推定精度が高かった主名詞（F-measure_{all} が 80 以上）が多く，学習時点で最大冠詞生起確率 $p(x_{max})$が大きい, つまり冠詞が the や other に偏った主名詞が多く見られた．すなわち，6.1 節でも述べた通り、冠詞生起確率の偏りが大きい主名詞に対しては、共起語手法は有効ではないことが分かった。

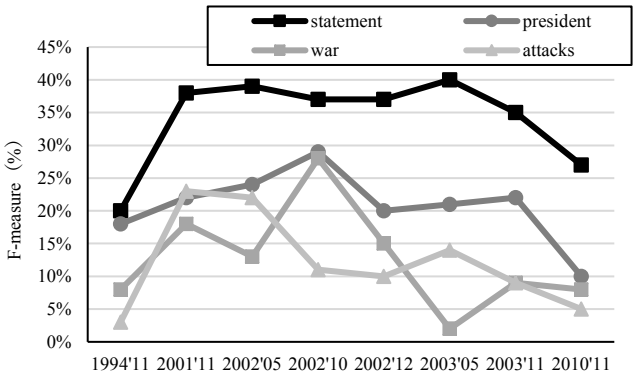


図 12 主名詞別の共起語手法による推定精度向上幅

6.4 共起語リスト

6.3 節と同様の理由で，“president”の共起語リストについて述べる．この節では，図 12 で推定精度の向上が低かった 2010 年 11 月,1994 年 11 月と一番向上幅の高かった 2002 年 10 月の各コーパスの共起語リストについて述べる．

従来手法に比べ，共起語手法では各コーパスの”president”の正解数は，1994 年 11 月は推定箇所数 762 箇所中 206 回（the:135, other:71）から 366 回（the:174, other:192），2002 年 10 月は推定箇所数 1670 箇所中 387 回（the:202, other:185）から 911 回（the:476, other:435），2010 年 11 月は推定箇所数 1796 箇所中 464 回（the:295, other:169）から 712 回（the:297, other:415）とそれぞれ増加している．

”president”に適用された共起語は，異なり語で 1500 語あるため，各評価コーパス中において，”president”への適用回数で降順に並べた上位 10 語をそれぞれ表 8.1~表 8.3 に示す．表中において，素性の重みが正であれば the に寄与する素性であり, 負であれば other に寄与する素性である．よって，”U.S.”や”United”のように，共起語リストに入っていても the の要因となっていない共起語が存在する．しかし，それらは必ずしも不必要な共起語とは言えない．なぜなら，冠詞周辺情報も含めた複数の素性との組み合わせにより，共起語が the の要因となる場合と，ならない場合をある程度区別出来ると考えられるからである．実際に，“president”の共起語は，全コーパスを平均すると平均 27 種類が同時に適用されている．精度面での検証を行うために，共起語リストの中から，負の素性値を持つ共起語をすべて除外し，再度実験を行った．結果としては，どの評価コーパスにおいても the の推定精度が向上したが，other の推定精度が下がり，全体的な推定精度も低下した．これらのことから, 負の重みを持つ共起語には, 推定結果が the に偏り過ぎないように抑制する効果があるため，むやみに除外すべきでないと言える．

表 8.1~表 8.3 より，冠詞推定対象主名詞と一致する共起語“president”が共起語として the に寄与していることがわかる．主名詞と一致する共起語は，2 章で述べた Coreferential に属すると考えられる．また, 表中の他の共起語を見てみると，“president”の人物（Clinton, Bush）を特定するような語が含まれている．これらの共起語は，2 章で述べた Bridging に属する語であると考えられる．また，”Bush”や”Clinton”のような，それぞれの年代異なる大統領は，その年代に対応した大統領名が共起語として多く適用され，このような要因が推定精度の向上幅に影響を与えていると考えられる．

表 8.1 “president”の共起語上位 10 語（1994 年 11 月）

共起語	適用回数	正解数	重み
President	353	163(the:111, other:52)	0.89
president	253	125(the:84, other:41)	0.51
government	222	85(the:36, other:49)	0.14
Clinton	203	103(the:95, other:8)	0.13
U.S.	187	99(the:60, other:39)	-0.13
Tuesday	186	85(the:49, other:36)	-0.28
years	168	88(the:30, other:58)	-0.71
United	159	65(the:41, other:24)	-0.49
Monday	144	75(the:49, other:26)	0.42
leaders	138	60(the:42, other:18)	0.06

表 8.2 “president”の共起語上位 10 語(2002 年 10 月)

共起語	適用回数	正解数	重み
President	669	402(the:299, other:103)	0.89
president	662	360(the:200, other:160)	0.51
United	416	272(the:197, other:75)	-0.49
U.S.	407	260(the:184, other:76)	-0.13
Bush	401	281(the:264, other:17)	0.97
officials	342	173(the:101, other:72)	-0.32
government	338	186(the:103, other:83)	0.14
Iraq	331	250(the:238, other:12)	1.16
States	323	204(the:143, other:61)	0.17
war	316	203(the:137, other:66)	-0.08

表 8.3 “president”の共起語上位 10 語（2010 年 11 月）

共起語	適用回数	正解数	重み
president	862	294(the:145, other:149)	0.51
President	650	248(the:164, other:84)	0.89
U.S.	582	240(the:107, other:133)	-0.13
government	468	197(the:100, other:97)	0.14
people	448	177(the:80, other:97)	-0.29
years	446	211(the:68, other:143)	-0.71
House	365	168(the:144, other:24)	0.2
time	365	167(the:48, other:119)	-0.51
year	336	178(the:41, other:137)	-0.48
country	334	129(the:66, other:63)	0.09

7. まとめ

本稿では，文内情報のみを用いる従来手法と，共起語リストにより前方文脈を考慮した共起語手法を用いて各年月別評価コーパスの推定性能にどのような変化が生じるかの検証実験を試みた．そのためにまず，共起語手法とは異なる手法との比較実験を行い，共起語手法の有効性を示した．次に，学習コーパスから，前後 1 ヶ月，半年，1 年，8 年の年月差のある記事群を各評価コーパスとし，従来手法と共起語手法における推定精度の変化を検証した．実験結果としては，どちらの手法でも，学習コーパスと年月が近いからといって必ずしも推定精度が高くなる訳ではなかった．しかし，共起語手法を用いたことによる推定精度の向上幅は，年月が近いほど高くなる傾向があることが分かった．また，それぞれのコーパスの共起語一致率と推定精度の向上幅にはある程度の相関があると考えられる．また，個々の主名詞ごとに分析すると，向上幅がより大きく現れる名詞が存在する．しかし、図 12 の一部のようにその変化が異なる名詞も存在する．本稿では、これについての十分な分析はまだできていない。

今後は，学習量を増やして更なる検証実験の実施や，学習用記事にクラスタリングを用いた文書分類を行い，評価用記事との文書間類似度がより高い学習用記事群を適用することで推定精度が向上するかの検証も行う．

また，共起語手法そのものにおける課題としては，共起語と冠詞付与対象名詞との距離を全く考慮していないという問題がある．文献[6]には，Bridging の現象は，Coreferential と比べ，より近い距離でのみ起こるとされているが，現在の共起語手法ではこうした事実を利用できていないという問題がある．

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] Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. Bulilding a large annolated corpus of learner English: The nus corpus of learner English. In *Proceedings of the eighth Workshop on Innovative Use of NLP for Building Educational Applications*, 2013
[3] C. Leacock, M. Chodorow, M. Gamon, J. Tetreault, Automated Grammatical Error Detection for Language Learners, G. Hirst, ed., Morgan and Claypool Publishers, La vergne, 2010
[4] Hwee Tou Ng, Siew Mei Wu et.al., The CoNLL-2014 Shared Task on Grammatical Error Correction: *Proceedings of the Eighteenth Conference on Computational Natural Language Learning,* Baltimore, Maryland, 2014.
[5] N.R. Han, M. Chodorow, and C. Leacock, “Detecting errors in English article usage by nonnative speakers”, Natural Language Engineering, vol.12, No.2, pp.115-129, 2006.
[6] F. Bond, Translating the Untranslatable: A Solution to the Problem of Generating English Determiners, CSLI Publications, Stanford, 2005.
[7] 竹内裕己，河合敦夫，永田亮，乙武北斗，“英文への自動冠詞付与における前方照応の考慮”，研究報告自然言語処理, vol.2011- NL-204，2011
[8] 竹内裕己，河合敦夫，細田直見，永田亮,”前方文脈を考慮した冠詞の推定", 言語処理学会第 19 回年次大会，2013
[9] 吉本一平，小町守，松本裕治,”定冠詞の前方照応用法を考慮した冠詞誤り訂正”, 言語処理学会第 20 回年次大会，2014
[10] 織田稔,”英語冠詞の世界:英語の「もの」の見方と示し方”, 研究社，2002.
[11] 原田豊太郎,”例文詳解 技術英語の冠詞活用入門”, 日刊工業新聞社，2000.
[12] http://www.chokkan.org/software/classias/
[13] http://nlp.cs.nyu.edu/oak/