前方文脈を考慮した冠詞の推定

竹内 裕己† 河合 敦夫† 細田 直見† 永田 亮‡

†三重大学大学院工学研究科

‡甲南大学知能情報学部

{takeuti, kawai, hosoda}@ai.info.mie-u.ac.jp

rnagata@limsi.fr

1 はじめに

近年，英語非母語話者による英文執筆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の話者においては，冠詞の誤用が多く報告されている．このような冠詞の誤用を自動訂正する手法では，英字新聞等のコーパスから冠詞の用法を機械学習し，正しい冠詞の推定を行っている．推定した冠詞を正解とし，訂正対象文に付与されている冠詞と照合することで，自動訂正を実現している．文献 [1] には，近年における種々の冠詞推定手法がまとめられているが，特に最大エントロピー（ME）分類器による冠詞推定手法が多く取り入れられ，高い推定性能を示している．

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 の決定要因の内，前方文脈に含まれる要因の考慮が不十分ということである．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 [2] において，特定・限定される要因は 4 分類され，その内 Coreferential（共参照）と Bridging は前方文脈に含まれる情報である．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.）．Bridging は，関連語により特定・限定する要因である（e.g. I caught a taxi. The driver was tired.）．これに対し多くの従来研究では，冠詞前後の数単語や，その品詞等の冠詞周辺情報のみを用いて推定を行うため，これらの要因は考慮されない．

研究 [3] では，文字列一致をベースとして Coreferential を考慮した冠詞推定を行っている．しかしこの手法では，文字列一致しない同義語での言い換えや上位語（e.g. I caught a taxi. The car was red.），Bridging などが考慮出来ない．したがって，これらを考慮することで，従来に比べより多く定冠詞 the を正確に推定出来ると考えられる．

そこで，本稿では，共起語を利用し，効果的に前方文脈を考慮する手法の提案を行う．提案手法では，3 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 の決定要因となる名詞を抽出する．これにより，Coreferential に加え Bridging も含めた，前方文脈中の the の決定要因を一括考慮する．この共起語を，ME 分類器による冠詞推定手法で利用することで，前方文脈考慮した冠詞推定を行う．

2 従来手法

本章では，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，冠詞周辺情報から得られる素性を用いる．また冠詞推定は，{a/an, the, ϕ（無冠詞）} に対して行われる．

本稿の目的である冠詞推定への前方文脈の考慮は，冠詞の用法上 {a/an, ϕ} の使い分けには影響を与えない．したがって，文献 [3] 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} を other と表記する．この 2 値分類を行うために，文献 [3] で用いられている素性を図 1 に示す．

3 提案手法

3.1 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，冠詞推定対象主名詞の冠詞が 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

(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

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，本手法で用いる共起確率と共起語の定義を行う．本手法で対象としている冠詞は，the と other 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 $C_p((\text{the}|x), y)$ は（1）式で求める．

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

ここで，$f(x \cap y)$ は，x, y が同時に出現する頻度を，$f((\text{the}|x) \cap y)$ 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．$C_p((\text{the}|x), y)$ が，推定対象主名詞 x の the の冠詞生起確率 $p(\text{the}|x)$ 以上となれば y を共起語とする．また，共起語の信頼性をある程度確保するため，$f(x \cap y) \geq 10$ とする．以上の条件をまとめた（2）式を共起語条件式として用いる．

$$Y = \{y | C_p((\text{the}|x), y) \geq p(\text{the}|x), f(x \cap y) \geq 10\} \quad (2)$$

3.2 共起語リストの作成

共起語の抽出は，学習用のコーパス中に含まれるすべ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．

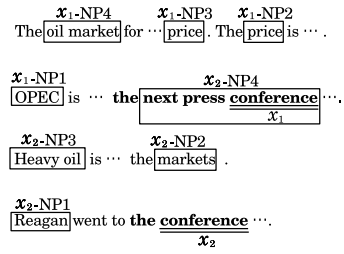

図 2: 名詞句抽出例

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より前に出現した名詞句（同一文内を含む）を，文章の先頭までの範囲から抽出する．具体的な抽出例を図 2 に示す．図 2 は同一文書中の連続した文であり，x_1-NP，x_2-NP は，対象主名詞 x_1，x_2 のそれぞれに対する抽出名詞句を示している．対象主名詞 x_2 において，x_2-NP4 中に対象主名詞 x_1 が含まれるため，それより前の名詞句は抽出しない．これは，対象主名詞間において，抽出する名詞句の重複を避けるためである．

抽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，さらに 2 語以上の名詞からなる場合は，別々に分けたものを共起語候補とする．また，“price”のように，同一の対象主名詞に対して共起語候補に重複して現れる名詞は 1 回のみ対象とする．図 2 で抽出した名詞句に含まれる共起語候補は表 1 のようになる．

表 1: 抽出された共起語候補

対象主名詞	共起語候補
x_1	OPEC, price, oil, market
x_2	Reagan, markets, oil, press, conference

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．表 2 に，“conference”を対象として場合の共起語候補集計結果の例を示す．

そして，この中から共起語条件式（2）を満たすものを抽出し，“conference”の共起語リストを作成する．$p(\text{the}|\text{conference}) = 75\%$ とすると，共起語条件式より，表 2 の中からは“conference”，“price”，“OPEC”の 3 語が共起語となる．

表 2: 共起語候補の集計結果 (x=conference)

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

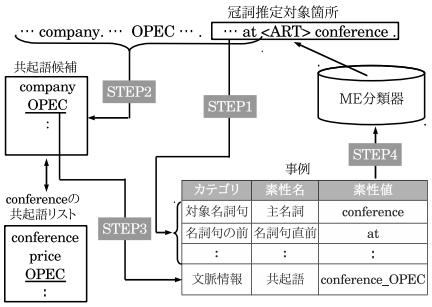

図 3: 共起語リストを利用した冠詞推定

3.3 共起語リストの適用

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．また，“ <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（2 章）と同様に，冠詞周辺情報の素性を事例に追加．

STEP2 対象名詞句前方の文章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 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“ OPEC ”を事例に追加する際に，“ conference_OPEC ”としている．

4 評価実験

本章では，2 章で述べた従来手法と，提案手法である共起語リストにより前方文脈を考慮した冠詞推定手法の性能評価実験について述べる．また，提案手法で共起語条件式を用いる有効性を示すため，4.3 節で説明する比較手法についても実験を行う．

4.1 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，Reuters が配信した新聞記事である Reuters-21578[4] を用いた．コーパス中には，507,576 の冠詞推定対象箇所（the, other）が含まれ，(the:26%, other:74%）となっている．また，10 分割交差検定を採用し，学習用コーパスと評価用コーパスを分割した．

本研究で用いる ME 分類器は，機械学習アルゴリズムの実装の一つである Classias[5] 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ツールは，OAK System[6] を用いている．

4.2 評価方法

本実験では，コーパス内に冠詞誤りはないと仮定し，冠詞推定結果とコーパス内の冠詞が，同一かどうか評価する．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，推定結果を信用するかを決定する指標として，Classias が出力する判定結果のスコアに対し，閾値 $\theta(\geq 0)$ を設定する．このスコアは，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．スコアの絶対値が，閾値 θ より小さければ，その推定結果は採用しない．

実験では評価指標として Recall と Precision を用いる．冠詞 ART（$\in \{\text{the}, \text{other}, \text{all}\}$）に対する Recall（$R_{ART}$）と Precision（$P_{ART}$）を (3) , (4) 式で定義する．ここで，$ART = \text{all}$ は，the と other の両方を対象にする場合を表す．

$$R_{ART} = \frac{\text{正しく } ART \text{ と推定した数}}{ART \text{ と推定すべき総数}} \quad (3)$$

$$P_{ART} = \frac{\text{正しく } ART \text{ と推定した数}}{ART \text{ と推定した数}} \quad (4)$$

4.3 比較手法

提案手法では，3.1 節で定義した共起語条件式を用いることにより，対象名詞句前方の文章に存在する全名詞から選択して前方文脈素性として考慮する．これに対する比較手法として，推定対象主名詞と文字列一致する主名詞のみを素性として用いる．すなわち，3.2

表 3: 手法別冠詞推定結果（$\theta = 1$）

評価値	従来手法	主名詞一致	提案手法
R_{the}(%)	48.6	51.2	55.3
P_{the}(%)	92.5	92.4	92.2
R_{other}(%)	85.6	86.5	87.4
P_{other}(%)	94.3	94.5	94.8
R_{all}(%)	76.0	77.4	79.1
P_{all}(%)	94.0	94.1	94.3

節と同様に共起語候補として抽出した名詞に対し，共起語条件式を適用する替わりに，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．そして，この共起語リストを 3.3 節と同様に冠詞推定に利用する．これを“主名詞一致”手法と呼ぶ．

4.4 結果

表 3 に，$\theta = 1$ における各手法の冠詞推定結果を示す．表 3 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，提案手法では，6.7 ポイント改善している．また，R_{other} は 1.8 ポイント向上し，R_{all} も 3.1 ポイント改善した．これらのことから，推定結果が the に偏ることなく効果的に R_{the} が向上していることが分かる．また，主名詞一致も，従来手法に比べ Recall が改善しているが，提案手法よりは低い．

次に，the の推定について Precision が同じ値の場合で Recall を比較する．冠詞誤り訂正を行う上で，高い Precision を設定し，どれだけ Recall を維持した推定をするかは重要な指標である．図 4 に比較結果を the に対する Precision-Recall 曲線で示す．一般的に Precision-Recall 曲線が，より右上に位置するとシステム全体の性能が高いと言える．したがって図 4 から，従来手法と主名詞一致に比べ提案手法のシステム性能が高いということが視覚的に理解できる．これらの結果から，提案手法は，the の推定精度向上に効果的であることが確認できた．

5 おわりに

本稿では，共起語リストを利用し，前方文脈を考慮した最大エントロピー分類器による定冠詞 the の推定手法を提案した．提案手法では，共起語リストにより，前方文脈中に存在する定冠詞 the 付与の要因となる名詞の一括した抽出を行った．実験の結果，冠詞周辺情報のみを用いた従来手法に比べ，the の推定性能が向上し，特に Recall が大幅に改善されることが確認された．また，前方文脈中から主名詞が一致する名詞のみを考慮した場合との比較を行い，実験結果より，提案手法が高い推定性能を示した．これらの結果から定冠詞 the の冠詞推定において，前方文脈考慮の必要性と，the に寄与する素性を選択して考慮することの有効性が示せた．

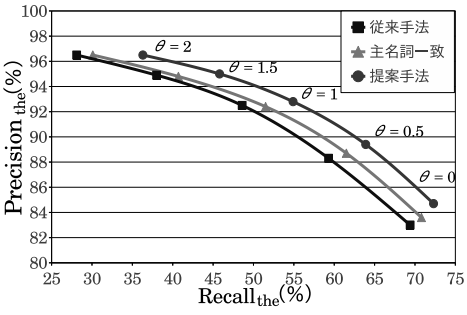

図 4: the に対する推定精度の Precision-Recall 曲線

提案手法により，the の Recall は改善したが，未だ改善の余地が残されている．さらなる精度改善には，前方文脈からより質の高い要因抽出を行う必要がある．改善案の一つとして，多くの研究がなされている名詞句照応解析手法 [7] 等を冠詞推定に応用することで，より正確に Coreferential の関係にある名詞句を抽出可能であると考えられる．具体的な冠詞推定への応用方法については，今後の課題としたい．

参考文献

[1] 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.

[2] F. Bond, Translating the Untranslatable: A Solution to the Problem of Generating English Determiners, CSLI Publications, Stanford, 2005.

[3] 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，no.10，pp.1–7，Nov. 2011.

[4] D. Lewis，“Reuters-21578 text categorization test collection，” 1997.

[5] N. Okazaki，“Classias: a collection of machine-learning algorithm for classification，” http://www.chokkan.org/software/classias/, Sept. 14,2011.

[6] S. Sekine，“Proteus Project: OAK System (English Sentence Analyzer)，” http://nlp.cs.nyu.edu/oak/, Feb. 29,2004.

[7] R. Iida, M. Yasuhara, and T. Tokunaga，“Multi-modal Reference Resolution in Situated Dialogue by Integrating Linguistic and Extra-Linguistic Clues，” The 5th International Joint Conference on Natural Language Processing, pp.84–92, 2011.