名詞の上位概念を考慮した 英文への冠詞付与規則の拡張

Extension of Article Generation Rules using WordNet Thesaurus

三宅 裕文　　河合 敦夫　　井須 尚紀
三重大学工学研究科情報工学専攻

1　まえがき

近年, 急速なグローバル化に伴い国際コミュニケーション言語としての英語の比重が大変大きくなっており, 日本人が英語を記述する機会が増加している. しかしながら, 英語非母語話者である日本人が正しい文法で英語を記述することは非常に難しい. 特に日本語には英語における冠詞の概念が存在しない為, 冠詞の用法の誤り (以下, 冠詞誤り) を犯す確率が高い [1].

しかしながら英語における冠詞は, 対象名詞が意味的に限定されているか否か, 等の情報を持つなど, その役割は非常に重要である [2]. このため日本人が英語で記述した論文等を提出する際, 多くの人が英文の校正を希望している. これまで英文の校正は, 専門知識を持った英語母語話者によって行われてきた. しかし, 人手による校正は多くの時間と労力を必要とするため, 文法の誤りを検出, 訂正するシステムの需要が増えている. そこで本論文では日本人の記述した英文での誤りが多いとされる冠詞誤りを検出し, 正しい冠詞を付与する「冠詞誤り訂正システム」の実現を目指す.

ここで, 実際に正しい冠詞の付与を行うには冠詞の用例集, 辞書等を参照する方法が考えられるが, 紙面の情報だけでは限界がある. またイディオムなどの固定表現のように文法書に記述されていない例外的な用法も多い. 以上のような点から, 冠詞付与を行う際の規則を全て人手で抽出し, 構築するのは非常に困難である. そこで電子化された英字新聞などの文法的信頼性が高いコーパスから抽出した統計量に基づいて, 冠詞に関する規則を機械的に作成し, その規則を用いて冠詞誤りを検出し正しい冠詞を付与する研究等が行われている.

冠詞誤りを検出する手法として手法 [1, 3, 4, 5] 等が, 冠詞を付与する手法として手法 [6, 7] 等が提案されている. 手法 [1, 3, 4, 6, 7] はコーパスから抽出した規則に基づく冠詞誤りの検出や冠詞付与を行っている. しかしながらコーパスの有限性により, 添削対象となる文書とコーパスの相性によっては適用できる規則が少なくなる, いわゆるデータスパースネスの問題が発生している. 平野ら [5] は添削対象の冠詞を含む動詞句またはその類似フレーズをクエリとして,WWW の検索エンジンから得られる検索ヒット数に基づいた冠詞誤りの検出手法を提案している. 一般的なコーパスベースの手法と異なり WWW を用いている為, データスパースネスの問題にある程度対応しているが, 得られる検索結果であるスニペット内の英文の文法的信頼性の問題があげられる.

データスパースネスの問題を解決する一手法として, 規則中の名詞の上位概念を考慮する事により, その規則の適用範囲を拡張する手法が考えられる. また一般的な用例集には「前置詞 “by” と輸送手段を表す名詞の間に出現する冠詞は無冠詞である (e.g.: by ϕ bus)」[8] のように具体的な名詞が全て記述されているわけではなく, 個々の名詞の概念（シソーラス）で用例が記述されている場合も多い. 実際に人間が英語を記述する際にも, 具体的な名詞毎に用例を考えるのではなく, 名詞の概念により用例を考えるのが自然である. 例えば “bus” に冠詞を付与する場合と “taxi” に冠詞を付与する場合の考え方は似ている. このような観点から名詞の上位概念を用いた規則の拡張は妥当であると考えられる.

本研究では永田ら [7] の冠詞付与手法 (以下, 従来手法) をベースとし, 規則中の名詞の上位概念を考慮して規則を拡張する手法を提案する. これにより, データスパースネスの問題に対応した冠詞付与手法の実現を目指す.

2　従来の冠詞付与手法

2.1　従来の冠詞付与手順

従来手法ではコーパスから冠詞を中心とした単語列を抽出し, その単語列のうち冠詞の分布に偏りがあるものを規則として学習していた. また規則毎に冠詞生起確率を推定していた.

具体的には, はじめにコーパスから冠詞を中心として前に n_+ 単語, 後に n_- 単語の単語列を抽出する. 例えば,

... knowledge of the chemistry of alcohol ...

という文において,$n_+ \leq 2$ and $n_- \leq 2$ の場合,

knowledge	of	the	chemistry	of
knowledge	of	the	chemistry	
knowledge	of	the		
	of	the	chemistry	of
	of	the	chemistry	
	of	the		
		the		

という単語列が抽出される. 本研究では, イディオムの一般的な長さやデータ量を考慮して,$n_+ \leq 5$, $n_- \leq 5$ とする.
次に各単語列の生起頻度 f(x) を取得する.　また冠詞を ART={a,the,ϕ(無冠詞)}, 冠詞 α を含む単語列 B を (α $|B$) とし, 冠詞生起確率を以下の式で定義する.

$$p(\ \alpha\ |B) = \frac{f(\ \alpha\ |B)}{f(a|B) + f(the|B) + f(\phi|B)} \qquad (1)$$

このように取得した単語列, 生起頻度, 冠詞生起確率を冠詞付与規則として学習する.

最後に, 入力文中の冠詞付与対象箇所について適用可能な規則のうち, 冠詞生起確率が最大の規則を用いて冠詞付与を行う. ここで適用可能な規則の生起確率値の最大値が閾値 θ 以下であれば, 冠詞付与は行わない. またここで閾値 θ の値 $(0 \leq \theta \leq 1)$ は使用目的に応じてユーザ側で指定する.

2.2 冠詞付与規則の信頼性

(1) 式において, 規則の生起頻度自体が少なければ, その生起分布の偏りや生起確率値の信頼性は低い. 例えば, 冠詞 α とだけ 1 回共起している生起頻度 1 回の規則 I_a(生起確率 100%) より, 冠詞 α と 98 回共起している生起頻度 100 回の規則 I_b(生起確率 98%) の方がより確実に偏っており信頼できる, といえる. 本研究では統計的に生起確率の信頼性を評価するために,χ^2 検定を用いた. すなわち, 各冠詞付与規則について, 各冠詞に対する生起確率を標本値とし,「各冠詞の出現確率は冠詞付与規則に関らず等しい」を帰無仮説として χ^2 適合度検定を行う. 各冠詞の分布確率を理論確率 $p_x(x \in ART)$ とすると, 統計量 χ^2 は以下の式で与えられる.

$$\chi^2(y) = \begin{cases} \sum_{x \in ART} \frac{(f(x|y) - f(ART|y) \times p_x)^2}{f(ART|y) \times p_x} & (n_y > 15) \\ \sum_{x \in ART} \frac{(|f(x|y) - f(ART|y) \times p_x| - 0.5)^2}{f(ART|y) \times p_x} & (n_y \leq 15) \end{cases}$$

$\chi^2(y) > \chi^2_a$ であれば, 帰無仮説が有意水準 α で棄却される. 本研究では有意水準 1%で検定を行った. ここで棄却された規則が理論確率分布からのずれが大きく, 信頼できる規則として冠詞付与を行う.

3 提案手法

3.1 名詞の上位概念の考慮

本研究では手法 [7] におけるデータスパースネスの問題を解決する為に, 名詞の上位概念を考慮した規則の拡張を提案する. 以下, 単語列そのものから作成した従来の規則を “従来単語規則”, 名詞の上位概念を用いて作成した規則を “拡張規則” と定義する. ここで, ある名詞が別の名詞の上位概念であるとは, ある名詞がより一般的, より総称的, より抽象的なものを指すものである. 本研究では名詞の上位概念を取得する方法として, シソーラス辞書である WordNet[9] を用いた. 本研究では WordNet のシソーラスのうち上位概念を利用した. 図 1 に “apple” の WordNet の上位概念の構成例を示す. ここで図中の “[...#n#...]” は上位概念を意味する.

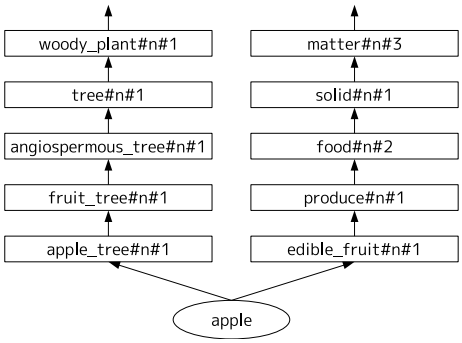

図 1: Example of Structure of WordNet Hypernyms

3.2 名詞の上位概念を考慮した冠詞付与手法

拡張規則を生成する手順として, 全ての従来単語規則の各々に対し, 規則中の名詞の上位概念を WordNet から抽出し本来の名詞と置換し拡張規則を生成する. 異なる従来単語規則から同一の拡張規則が生成される場合は生起頻度を足し合わせる事により一つの拡張規則とする. また一つの従来単語規則中に名詞が複数個含まれている場合は, データ量の問題から置換する名詞は一つずつとする. 従って, 一つの従来単語規則に名詞が m 個含まれているとき, 生成される拡張規則は m 個となる. 例えば,

ϕ further software testing

という従来単語規則からは以下の拡張規則が生成される.

ϕ	further	[code#n#3]	testing
ϕ	further	software	[investigation#n#2]

3.3 多義語の上位概念置換

多くの英語名詞は多義語であり, 多義語を上位概念で置換するには文脈を考慮し適切な上概念を選択する必要がある. しかし本研究では, 文の表層情報のみを考慮しているため, 文脈を考慮した上位概念の選択は不可能である. そこで「関連する語は同じ文章中や近い位置で共起する. 類似したコンテキストに現れる単語は, 類似した性質を持つ.[10]」というヒューリスティクスを利用する. すなわち, 同じ分野の文書では同じ上位概念をもつ単語が多い, ということを利用する.

本研究では同じ分野の基準として,Reuters-21578 コーパス [11] 中のトピックタグを利用した. すなわち, 同じトピックタグが付与されている文書を同じ分野の文書として扱った.

具体的には, はじめにコーパス中の全名詞を抽出し,WordNet からその名詞の上位概念を抽出する. 次に, 抽出した上位概念の頻度を取得し, トピック毎に上位概念の頻度分布表を作成する. このように作成した頻度分布表を用いて, ある名詞を上位概念で置換する際に, その名詞の複数の上位概念のうち, 名詞の所属するトピックと同じトピック名の頻度分布表の頻度が最上位の上位概念と置換する. ここで χ^2 適合度検定において, 上位の頻度に統計的な差が見られなかった場合は複数の上位概念と置換する.

例えばここで “wheat” という多義語が複数の上位概念 ([cereal#n#1],[grain#n#2],[yellow#n#1]) を持ち,grain トピックの頻度分布表が

上位概念	頻度
[cereal#n#1]	1405
[time_period#n#1]	1226
[grain#n#2]	1025
[metric_weight_unit#n#1]	956
[yellow#n#1]	826
:	:
:	:

という分布をしていた場合,grain トピック内から生成された

$$\phi \quad \text{wheat}$$

という従来単語規則からは, 以下の拡張規則のみが生成される.

$$\phi \quad [\text{cereal\#n\#1}]$$

但し,[cereal#n#1] と [grain#n#2] の頻度に統計的な差が見られない場合は以下の二つの拡張規則が生成される.

$$\phi \quad [\text{cereal\#n\#1}]$$
$$\phi \quad [\text{grain\#n\#2}]$$

なお今回はトピックタグが付与されている特殊なコーパスを利用した. このようなタグがない文書に本手法を適用する場合, 論文などの比較的長い文書の場合はその文書中の全名詞から頻度分布表を作成する. また今回の新聞記事のような比較的短い文章の場合は文書クラスタリング [12] などを行い, 同一クラスタ内の全名詞から頻度分布表を作成し, 利用することで応用が可能であると考えられる.

4 評価実験

4.1 評価方法

本実験では,Recall(冠詞付与率),Precision(冠詞付与精度),F 値の 3 種類の尺度に対し,10 分割交差法を用いた性能評価を行った. 各値は以下の式で定義できる.

$$Recall = \frac{\text{システムが正しい冠詞を付与した数}}{\text{入力文中の全冠詞付与箇所数}}$$

$$Precision = \frac{\text{システムが正しい冠詞を付与した数}}{\text{システムによる冠詞付与数}}$$

$$F\ 値 = \frac{(1+b^2) \times Precision \times Recall}{b^2 \times Precision + Recall}$$

ここで $|b|$ の値が 0 に近い程 Precision が重視された値が求められる. 本研究では $b = 1$(調和平均) として評価を行った.

実験手順として, はじめに, 評価用コーパス中の冠詞を全て除去する. 次に除去した冠詞を正しい冠詞とし, 除去した箇所に対して冠詞付与実験及び評価を行った.

総記事数	総単語数	総冠詞箇所数 (a, the ϕ)
4828	811473	180587

表 1: Corpus scale

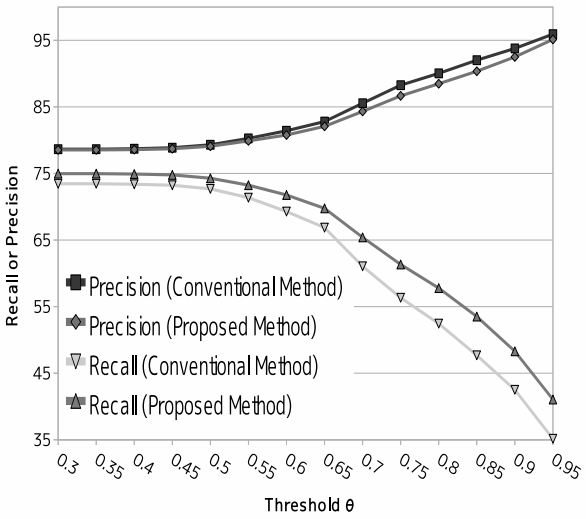

図 2: Recall and Precision

本実験では学習用および評価用コーパスとして,Reuters-21578 コーパス [11] を使用した. Reuters-21578 コーパスは Reuters が配信した記事に対して人手でトピックを付与したものである. 本実験では主要 9 トピックが付与されている記事を使用した. 使用したコーパスの規模は表 1 の通りである.

4.2 実験結果及び考察

はじめに Recall と Precision の実験結果を図 2 に示す. 従来手法に比べて Recall の値は, 全ての閾値において上昇している. これは上位概念を用いて規則の拡張を行った為, 適用できる規則が増えた為と考えられる. 一方,Precision の値は全ての閾値においてわずかに低下している. これは上位概念を用いて生成された拡張規則の曖昧性が高いことに起因していると考えられる.

次に Recall と Precision の調和平均である F 値を図 3 に示す. F 値は全ての閾値において上昇している. これは, わずかな Precision の減少は見られるがシステム全体のパフォーマンスは上昇している事を示している. 特に従来手法において,Precision は高い値を示すが Recall の値が低かった高閾値において, 例えば $\theta = 0.95$ においては 6%の顕著な上昇が見られた.

またここで横軸に Recall, 縦軸に Precision を取る,PR 曲線を図 4 に描く. 図 4 を見ると, 提案手法の PR 曲線が全体的に右上に位置していることからシステム全体のパフォーマンス向上が見られる.

F 値,PR 曲線の両観点から, 本提案手法は有効である事がわかった. また本手法を用いることにより, 一般的なイディオム集と違い, シソーラスを用いたイディオム集の作成が可能であり, 様々な学習支援にも応用できると考えられる.

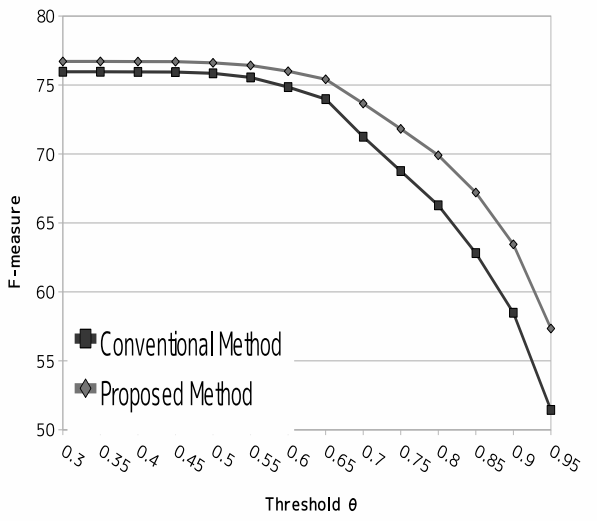

図 3: F-measure

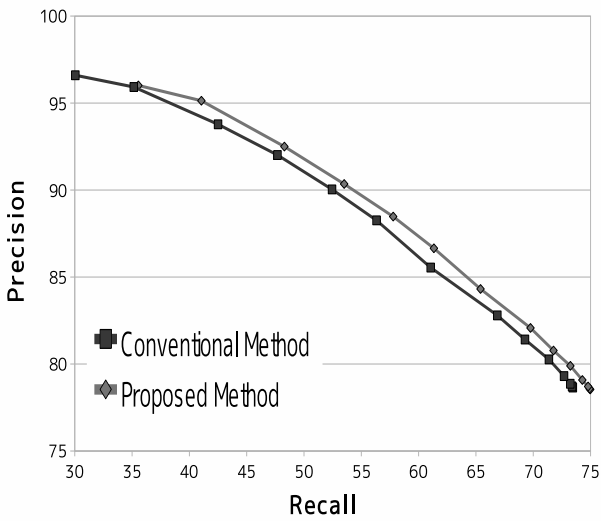

図 4: Precision-Recall Curve

5 まとめ

本論文では, 以前に提案されていた冠詞付与手法において名詞の上位概念を用いる事により冠詞付与規則を拡張し, データスパースネスの問題に対応した冠詞付与手法について述べた. また多義語の存在により, 名詞の上位概念が複数個存在する場合において, 適切な上位概念を選択する手法の提案を行った. 実験の結果, 名詞の上位概念を考慮した冠詞付与規則を用いることにより,Precision を下げずに,Recall を向上させられることが確認された. 特に従来手法において,Precision は高い値を示すが Recall の値が低かった高閾値において, 例えば $\theta = 0.95$ の F 値において 6%の顕著な上昇が見られた.

今後の課題としては, さらなる Recall の向上が挙げられる. 現在では名詞の一段階上の上位概念のみを考慮している. ここで更に複数段階上の上位概念を考慮することで, より多くのデータスパースネスの問題を解決することができると考えられる. しかしながら複数段階上の上位概念を考慮することにより, 適用できる規則が増す一方で, その規則の曖昧性が高くなり Precision の低下が引き起こされると予測される. したがって, 何段階上の上位概念までを考慮するか, という選定方法が重要になってくる. この選定方法の検討については今後の課題としたい. また WordNet に含まれていない人名, 組織名, 地名などの単語の対処については,WordNet の下層部を Wikipedia で補間したシステム「YAGO[13]」などの利用を検討している.

参考文献

[1] A. Kawai et al., “ASPEC-I : An error detection system for English composition”, Proc. IPSJ Journal, 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, 1999

[3] R. Nagata et al., “Recognizing article errors in the writing of Japanese learners of English”, Proc. IEICE, Vol. J87-D-I, no. 1,pp. 60-68, Jan. 2004

[4] H. Ototake et al., “Correcting and detecting article errors in English using conditions of word appearance”, Proc. IPSJ 2006-NL-171, pp. 25-30, Jan. 2006

[5] T. Hirano et al., “Detecting Article Errors in English using Search Engines”, Proc. DBSJ Letters, vol. 6, No. 3, pp. 1-4, Dec. 2007

[6] G. Minnen et al., “Memory-based learning for article generation”, Proc. CoNLL-2000 and LLL-2000, pp. 43-48, Mar. 2000

[7] R. Nagata et al., “Extracting Collocations for Determining Articles in English Writing”, Proc. PACLING2005, Aug. 2005

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

[9] C. Fellbaum, “WordNet : An Electronic Lexical database”, The MIT Press, 1998

[10] 関根聡 et al., “提唱「コーパスベース知識工学”, 言語処理学会第 13 回年次大会ワークショップ「大規模 Web 研究基盤上での自然言語処理・情報検索研究」論文集, pp. 12-15, March. 2007

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

[12] N. Uramoto et al”A Method for Relating Multiple Newspaper Articles by Using Graphs, and Its Application to Webcasting”, Proc. COLING-ACL’98, pp. 1307-1313, Aug. 1998.

[13] F. M. Suchanek et al., “Yago: A core of semantic knowledge”, Proc. 16th international conference on World Wide Web, May. 2007