社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2005—NL—167（20）
2005—SLP— 56（20）
2005／5／27

語彙的連鎖からの名詞照応連鎖の抽出

杉原　大悟†　　増市　　博†　　大熊　智子†　　吉村　宏樹†

† 富士ゼロックス (株) 研究本部
〒 259-0157 足柄上郡中井町境 430 グリーンテクなかい
E-mail: †{daigo.sugihara,hiroshi.masuichi,ohkuma.tomoko,hiroki.yoshimura}@fujixerox.co.jp

本稿では、テキスト中の語彙的連鎖から名詞照応連鎖を分離する実験について述べる。名詞照応解析において最も重要な手掛かりは名詞の形態的な類似である。我々は名詞照応現象を「語彙的連鎖に含まれる照応連鎖」と「語彙的連鎖に含まれない照応連鎖」とに分けて捉え、語彙的連鎖から照応連鎖を得るという方針で実験を行った。既存の名詞照応解析手法で用いられてきた素性セットに加えて、解析対象の名詞の修飾句および係り先の名詞から「文脈全体に関する情報」についての素性セットを取得し、語彙的連鎖中の名詞ペアが同じ参照先を持つか否かを分類する分類器を SVM で構築した。実験結果を分析したところ、新聞報道記事では、約 8 割の名詞照応連鎖が語彙的連鎖に包含されていることが分かった。また、文脈に関する情報は、個々の名詞に関する情報よりも、その名詞の修飾句や係り先の名詞から得られた情報が、照応解析においてより有効に働くという結果が得られた。

Extraction of the coreference chains from the lexical chains

Daigo SUGIHARA†, Hiroshi MASUICHI†, Tomoko OHKUMA†, and Hiroki YOSHIMURA†

† Corporate Research Group Fuji Xerox Co., Ltd.
430 Sakai, Nakai-machi, Ashigarakami-gun, Kanagawa,Japan
E-mail: †{daigo.sugihara,hiroshi.masuichi,ohkuma.tomoko,hiroki.yoshimura}@fujixerox.co.jp

We report experimental results on the extraction of the coreference chains from the lexical chains. The morphological similarity is an important clue for noun phrase anaphora resolution. We divide noun phrase anaphora resolution process into two parts: One is ”the process to detect the coreference chain completely contained in the lexical chain” and the other is ”the process to detect coreference chain which is not contaied in the lexical chain”. We take a course of the extraction of the coreference chain from the lexical chain. We constructed the classifier which judges whether the pair of the noun phrases in the lexical chain is coreferential or not by using the feature set concerning the entire context of the text concerning the nouns that modify the target noun and the noun on which the target noun depends and by using the existing basic feature set proposed by the previous works. The results suggest that approximately 80% of the coreference chains is contained in the lexical chain. The results also suggest that the information concerning the entire context of the text concerning the nouns that modify the target noun or the noun on which the target noun depends is more useful than the information of that from the target noun itself in noun phrase anaphora resolution.

1. はじめに

本稿では、テキスト中の語彙的連鎖から名詞照応連鎖を分離する実験について述べる。

ある言語表現が、別の場所に現れている言語表現と同一の対象を参照する場合、これらの表現は照応関係にあるという。照応関係にある 2 つの表現のうち一方に注目した場合、その言語表現を照応詞と呼び、もう一方の表現を先行詞と呼ぶ。照応解析とは、テキスト中の照応関係を特定する処理であり、高品質な翻訳システムや質問応答システムなどの高度な自然言語処理アプリケーションの実現に不可欠な処理である。照応詞は、大別すると名詞、代名詞およびゼロ代名詞に分けることができ、それぞれの照応現象を名詞照応、代名詞照応、ゼロ照応という。本稿では、日本語における名詞照応を対象とする。

また、テキストの文脈を扱う概念には、結束性 (cohesion) と呼ばれる概念がある。結束性とは言語表現間の表層的な結びつきを表す言語的な性質である。Halliday と Hasan は、結束性を代名詞 (pronoun)、省略 (ellipse)、接続 (conjunction)、代入 (substitution)、語彙的結束性 (lexical cohesion) の 5 つに分類している [2]。接続とは、接続詞によって媒介される文や語の間の関係のことをいう。代入とは、既出の語を別の語で言い換えることである。語彙的結束性はテキスト中の表層語間に何らかの意味的な関係があることを言う。照応解析という側面から、これらの結束性を分類すると、代名詞と省略はそれぞれ代名詞照応とゼロ照応が扱う領域となる。代入と語彙的結束性は、テキスト中の表層間の関係に現れる結束性であり、名詞照応が扱う領域となる。特に名詞照応解析では、語の表層上の類似性が重要な手掛かりとなるため、語彙的結束性に深い関わりを持つといえる。テキスト中における語彙（特に形態）的な繋がりを、語彙的連鎖（lexical chain）[4] という。

名詞照応解析は、テキスト中に出現した名詞間の照応関係を特定する処理である。この処理において最も重要な手掛かりは表現の形態的な類似であり、名詞照応解析の先行研究の多くで利用されている。村田ら [15] によるルールベースの日本語名詞照応解析手法では、「自分」などの特定の表現を解析対象とする場合を除いて、解析対象の名詞よりも前に出現した同一名詞か、解析対象の名詞を末尾に含むような名詞に対して、照応関係の評価を行うようにシステムが構築されている。機械学習を用いた名詞照応解析手法においても、表現の形態的な類似は照応解析の重要な手掛かりであるとされている。Soon ら [5] は、照応詞とその先行詞候補が照応関係にあるかないかを判定する決定木を学習データから構築し、得られた決定木を用いて照応詞とその先行詞候補を、照応詞と近いものから順に評価していくことで解析を行うシステムを構築した。Soon らのシステムが MUC-6 のデータに対して 12 種の素性を用いて学習を行った実験結果の精度は F 値=0.624 であり、照応詞と先行詞の文字列の一致の素性（STR_MATCH）のみを用いて学習をした場合の精度は F=0.539 であった。この結果は、文字列の一致素性がシステムの性能の大きな部分を占めることを示唆している。また、Soon らはシステムの性能を左右する最も重要な素性として「文字列素性（STR_MATCH）」、「別名 (ALIAS)」、「同格の関係にあるか (APPOSITIVE)」を挙げている。Strube ら [6] は照応詞と先行詞の文字列編集距離に関する素性を機械学習に組み入れることでシステムの精度が向上したと報告している。さらに、Yang ら [8] は、修飾節を含む文字列の類似度を素性に用いて学習を行うことにより、名詞照応解析の精度を向上させることができると報告している。

しかし、現実のテキストには、同一の文字列からなる二つの名詞が異なる参照先を有する場合も存在する。Soon らや飯田ら [10] は、文字列素性が強く働きすぎたため照応解析を誤る場合があることを報告している。先行詞と照応詞の文字列の類似性は名詞照応解析において有効な手掛かりであるが、類似した文字列を持つ名詞が全て照応連鎖を構成するわけではない。したがって、同一の文字列からなる名詞が別々の照応連鎖に含まれている場合や、あるいは照応連鎖の要素にならない場合を区別しなければならない。すなわち、テキスト中の表現の表層上の類似性から得られる語彙的連鎖は照応連鎖と強い相関関係を持つが必ずしも一致するとは限らないことを十分に考慮する必要がある。

そこで、我々は名詞照応現象を「語彙的連鎖に含まれる照応連鎖」と「語彙的連鎖に含まれない照応連鎖」とに分けることによって語彙的連鎖から照応連鎖を得るという方針で実験を行った。

以下に本稿の構成を述べる。2 章では、語彙的連鎖から名

詞照応連鎖を抽出するシステムの概要を述べる。3 章では、実験のために構築した正解コーパスの仕様について述べる。4 章では、実験の結果と考察を述べる。5 章ではまとめと今後の展開について述べる。

2. システム概要

前章で述べた通り、我々は名詞照応現象を「語彙的連鎖に含まれる照応連鎖」と「語彙的連鎖に含まれない照応連鎖」とから成立していると考え、「語彙的連鎖中に含まれる照応連鎖」と「語彙的連鎖に含まれない照応連鎖」をそれぞれ別の処理によってテキスト中から得る手法の実験を行った。本稿では、語彙的連鎖を、以下の条件を満たす名詞をテキスト中から集めて纏めたものであると定める。

条件 1　同一の形態素を共有する名詞
条件 2　前方マッチまたは後方マッチする名詞
条件 3　分類語彙表 [12] で同じ意味番号を持つ名詞

条件 1 と条件 2 は、形態的に類似した語の連鎖であり、条件 3 は意味的に類似した語の連鎖を得ることを意図している。図 1 は、システムの処理フローの概要である。以下に説明を行う。

まず、入力テキストから語彙的連鎖を生成する。次いで、各語彙的連鎖中の全ての名詞ペアを評価し、語彙的連鎖中のペアのうち照応連鎖として相応しいとされたペアを抽出し、さらに同じ照応関係を持つと判断されたペアを纏めることによって、新たな語彙的連鎖に分割する。次に、得られた語彙的連鎖中の語に、元の語彙的連鎖外の語との間に予め設定した関係(注1)が認められるならば、その語を語彙的連鎖に追加する。システムは、以上の処理によって得られた語彙的連鎖を照応連鎖として出力する。今回の実験では、語彙的連鎖の生成のために、入力テキストを「LFG に基づく日本語解析システム [14]」と「Cabocha [13]」を併用して解析し、解析結果から照応解析の対象となる名詞部分と照応解析のための各種素性を得た。

以下に、語彙的連鎖中のペアの評価、および語彙的連鎖に含まれない照応連鎖に対する処理についての説明を行う。

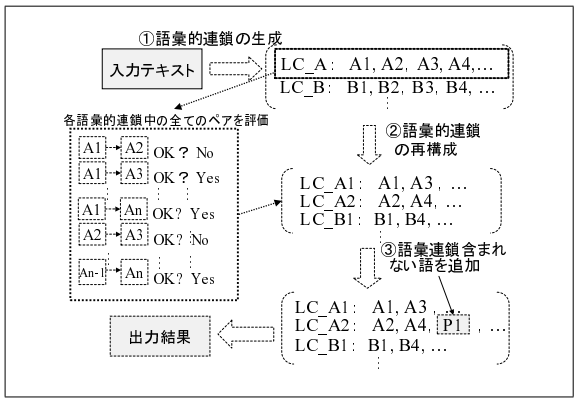

図 1　システムの概要

2.1 語彙的連鎖に含まれる照応連鎖に対する処理

本稿による名詞照応解析システムは、語彙的連鎖中の全ての名詞ペア間の関係について照応連鎖であるか否かを評価し、同じ照応関係があると判断したペアを集め、語彙的連鎖を分割する。本稿では、語彙的連鎖中のペアが照応連鎖として相応しいか否かの評価に、名詞ペアが文字を共有しているか否かのヒュ-リスティクスと、機械学習に基づく 2 値分類器を用いている。学習の手順は以下のように行った。照応関係の正解を付与したコーパス中から語彙的連鎖を抽出し、語彙連鎖中の名詞ペアのうち文字を共有している名詞ペア集合を生成する。ついで、名詞ペア集合の各ペアのうち、同一の参照先を持つならば正例、そうでないならば負例として学習データを作成した。学習には、SVM [7] を、SVM 学習パッケージには TinySVM を用いた。本稿による解析システムの語彙的連鎖に対する処理を図 2 に示す。

テキスト中の名詞が照応連鎖の要素となるか否かの評価には、二つの側面からの視点が必要となる。二つの名詞が同一の参照先を持つか否かという視点と、名詞が文脈上で言及されている要素か否かという視点である。前者は例えば、「モナコの外交官」と「イギリスの外交官」は同じく「外交官」という意味を共通して持つが、それぞれ別々の参照先を持つことを判断する視点である。後者は、名詞指示性 [15] [16] とも呼ばれているものである。

飯田ら [10] は、この 2 つの視点を統合したモデルを構築している。また、飯田ら [11] は、先行文脈情報を利用することで照応性判定の精度が大幅に向上したことを報告している。さらに、Yang ら [9] は、英語名詞照応解析における先行文脈の利用が、英語名詞照応解析の精度向上に有益であると報告

(注1)：本稿では、括弧内の名詞とその直前の名詞についてルールを設定した。本稿の 2.2 に記述した内容である。

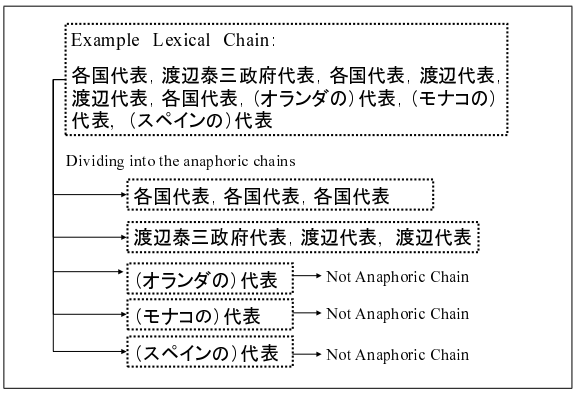

図 2　語彙的連鎖に対する処理の例

している。Ge ら [1] は、１度先行詞となった名詞は、繰り返し先行詞になりやすいと述べている。

先行研究における先行文脈とは、照応詞からテキストの先頭へと連なる先行詞候補に関する情報であり、照応解析をテキストの先頭から行う過程で蓄積される情報である。しかし、我々の試みた手法では、テキストから一度に語彙的連鎖を生成し、その語彙的連鎖を分割する過程から照応連鎖を得るため、そのような情報を用いることはできない。本稿で用いる文脈に関する情報とは、テキスト全体が持つ処理過程に依存しない情報である。

名詞ペアの同一性を判定するための基本素性の設計は、特に飯田ら [10] を参考とした。また文脈全体に関する素性の設計には Yang ら [9] および Ge ら [1] の研究を参考にした。

2.2　語彙的連鎖に含まれない照応連鎖に対する処理

今回の実験では、語彙的連鎖に含まれない照応連鎖に関しては、各名詞ごとに個別に対応することにした。Soon らによる「別名（ALIAS)」素性が有効であるという指摘を受け、今回は括弧に関する以下のようなルールを用い、語彙的連鎖を分割した後に、連鎖中の名詞に対して語彙的連鎖に含まれない照応連鎖を加えることを試みた。

@括弧に関する前提 1　　分割処理後の語彙的連鎖連鎖に含まれる名詞 A に続く括弧内の名詞 B がある時、または、

@括弧に関する前提 2　　分割処理後の語彙的連鎖連鎖に含まれる名詞 A が括弧に囲まれ、かつ直前の名詞 B がある時、

@前提を受けての処理　　A と B の品詞が一致するか、どちらかの品詞が「未知語」の場合には、B を A が属する連鎖に追加する

以下に上記ルールに関する例を挙げる。ある分割された語彙的連鎖を $C\{NIE, NIE\}$ とする。その時、テキスト中に「博覧会国際事務局 (NIE)」という表現があった場合は上記条件に合致し、「博覧会国際事務局」を連鎖に加え、$C\{$ 博覧会国際事務局$, NIE, NIE\}$ となる。

3.　照応連鎖正解付けコーパス

読売新聞報道記事中に現れる名詞について、同一の参照先を持つ表現に対して同一の参照を表す ID を付与し、これを正解として学習と実験を行った。正解付けの対象としたのは、新聞記事中の以下の基準を満たす名詞である。

@　記事中の文脈中に現れる物事を指し示す名詞で、同一参照先を持つ名詞同士

@　数字表現が含まれる名詞については「日時」のみを正解付けの対象とする

飯田ら [10] は名詞照応の正解付けについて、テキスト中の照応詞タグに対し、その照応詞の先行詞が持つ ID を付与することを行っている。その際、飯田らは総称名詞と不定名詞は正解付けの対象から除外し、外界照応を除外し、照応詞は文節の主辞を対象としてタグを付与している。本稿でも飯田らと同様に、外界照応を対象としない。しかし、名詞の指示性については、飯田らの正解付け基準よりも緩く、総称名詞、不定名詞、定名詞の区別を特に意識していない。これは、このような名詞の区別に関わらず、文脈上で言及されている名詞は、他のアプリケーションへの応用上、取り扱うことができることが望ましいと考えたからである。

また、複合名詞の構成要素については、今回は学習および実験の対象とはしなかった。飯田らは複合名詞について、複合名詞を構成する要素の境界を問題としている。本稿ではさらに、文脈上の要素として認められるか否かに関して曖昧性が高いことも勘案し、実験の対象から除外することにした。以下の例文を用いて説明を行う。

例文　尾道 市民$_1$ プラザは、尾道 市民$_2$ の社会活動のために開放されている。

上記の「尾道 市民$_1$ プラザ」の「市民$_1$」部分には、文脈の上での意味的な曖昧性が存在する。一つの解釈としては、この「市民$_1$」は、地方自治体「尾道市」の市民権を持つ「尾道市民」であり、上記文の「尾道 市民$_2$」の先行詞と解釈するこ

とが可能である。しかし、同時に「市民プラザ」という公共施設一般の名称の一部であると解釈することもできる。この場合は「尾道 市民$_2$」の先行詞とはいえない。本稿では、複合名詞の構成要素についての照応解析は、別のフレームワークで行うべきだという立場をとり、今回の実験には用いないことにした。

また、数量表現は「日時」を除いて正解付けの対象から除外した。MUC [3] の照応正解付けコーパスでは、名詞とその名詞が持つ値との間に照応関係があるとしてタグ付けを行っている。例えば、「The temperature rose to 90 degrees before dropping to 70 degrees.」という文では、「90 degrees」の先行詞として「The temperature」の ID を付与するようタグを付けている。ただし、「70 degrees」も「The temperature」を先行詞となるようにタグを付けてしまうと「90 degrees」と「70 degrees」が「同一」であるという矛盾が生じることになる。そのため、「The temperature」と「90 degrees」のみが照応関係にあるとしてタグを付け、「70 degrees」に関しては無視する方針をとっている。しかし、本研究においては、同一の参照先を持つ名詞を同一 ID で纏めるため、上記のような数量表現に対して矛盾なくタグを付与することはできない。よって、数量表現に関しては正解付けの対象から除外した。ただし、「日時」はテキスト中において、何かの値として用いられるよりもむしろ「日時」単体で現れる場合が多いため、正解付けの対象に含めた。

以下、本稿では、照応連鎖中の連鎖をカウントする名称として、ペア数とリンク数という名称を用いる。ペア数は照応連鎖中の名詞の可能な組み合わせであり、リンク数は照応連鎖の先頭から現れる先行詞-照応詞のペアの数である。例えば、４つの名詞から構成される名詞照応連鎖 $C\{e1, e2, e3, e4\}$ があった場合に、照応解析対象数 (連鎖に含まれる名詞の数) は $\{e1, e2, e3, e4\}$ の4、ペア数は $\{e1-e2, e1-e3, e1-e4, e2-e3, e2-e4, e3-e4\}$ の 6、リンク数は $\{e1-e2, e2-e3, e3-e4\}$ の 3 であり、連鎖数は 1 とカウントする。

4. 実験結果

実験には 98 年と 99 年の読売新聞報道記事 82 記事のうち、60 記事を学習に、22 記事を評価に用いた。精度は以下を用いた。AC は照応連鎖の略称である。DE は照応解析対象を意味し、照応連鎖に含まれる名詞を表す。また下記の「$(C \mid P \mid L)$ 数」における、「C」、「P」、「L」は、それぞれ「連鎖」、「ペア」、「リンク」に対応している。それぞれの出力数、正解出力数、および正解コーパス中の数を用いて、「連鎖」、「ペア」、「リンク」ごとの精度を算出したということである。「正しく出力できた連鎖数」は、出力した照応連鎖の要素が正解の照応連鎖と比較して欠損や過剰出力がない場合にカウントする。「正しく出力できたペア数」および [正しく出力できたリンク数」とは、出力された照応連鎖中のペアおよびリンクが正解コーパス中に存在している場合にカウントする。

$$
\begin{aligned}
NounCoverRate &= \frac{\text{出力中の } DE \text{ 数}}{\text{正解中の } DE \text{ 数}} \\
ChainCoverRate &= \frac{\text{出力 } AC \text{ に正解 } AC \text{ が包含される数}}{\text{正解 } AC \text{ 数}} \\
Recall &= \frac{\text{正しく出力できた } (C \mid P \mid L) \text{ 数}}{\text{正解中の } (C \mid P \mid L) \text{ 数}} \\
Precision &= \frac{\text{正しく出力できた } (C \mid P \mid L) \text{ 数}}{\text{システムが出力した } (C \mid P \mid L) \text{ 数}} \\
F \text{ 値} &= \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}
\end{aligned}
$$

上記の NounCoverRate は、語彙的連鎖を軸にした我々のシステムがどの程度真の照応連鎖を評価の対象とできているかを表す。また、ChainCoverRate は、出力された連鎖が真の照応連鎖をどの程度包含しているかを表す。実験の設定を表 1 に、実験の結果を表 2 に示す。

表 2 の LC_ONLY、COMP_STR および HEAD_STR の実験設定では、分類器を用いず、単純なルールによって語彙的連鎖を分割する。BASIC_FEATURE では、語彙的連鎖中の名詞ペアを、基本素性を元に学習を行った分類器によって評価し、語彙的連鎖を分割する。+GLOBAL では、名詞ペアの２つの名詞について、基本素性と文脈全体に関する素性を元に学習を行った分類器によって語彙的連鎖を分割する。+GLOBAL_MOD では、名詞ペアの２つの名詞については基本素性を用い、名詞ペアの各名詞ごとに「名詞に係る修飾句、あるいは係り先の名詞」について文脈全体に関する素性を抽出し、それらの素性セットを用いて学習を行った分類器によって語彙的連鎖を分割した。+GLOBAL_MOD は、「文脈全体で言及されている名詞によって修飾されている名詞」や「文脈全体で言及されている名詞を修飾する名詞」は、

表 1　実験の設定

設定の略称	説明
LC_ONLY	生成した語彙的連鎖を分割後の語彙的連鎖とする
COMP_STR	完全に文字列の一致する名詞集合を分割後の語彙的連鎖とする
HEDA_STR	主辞の一致する名詞集合を分割後の語彙的連鎖とする
BASIC_FEATURE	語彙的連鎖を基本素性による分類器で分割する
+GLOBAL	基本素性と名詞ペアの各名詞についての文脈全体に関する素性を用いて学習した分類器で語彙的連鎖を分割
+GLOBAL_MOD	基本素性と名詞ペアにおける各名詞の「修飾句と係り先の名詞」についての文脈全体に関する素性を用いて学習した分類器で語彙的連鎖を分割

表 2　実験結果：R は Recall、P は Precision、F は F 値をそれぞれ略している

	Noun	Chain	Chain			Pair			Link		
Experiments	CoverRate	CoverRate	R	P	F	R	P	F	R	P	F
LC_ONLY	0.948	0.834	0.247	0.176	0.204	0.876	0.097	0.174	0.471	0.198	0.279
COMP_STR	0.715	0.491	0.402	0.312	0.351	0.568	0.619	0.593	0.580	0.534	0.556
HEAD_STR	0.861	0.722	0.467	0.325	0.383	0.784	0.493	0.606	0.721	0.487	0.581
BASIC_FEATURE	0.715	0.515	0.385	0.357	0.370	0.629	0.632	0.631	0.596	0.581	0.589
+GLOBAL	0.688	0.497	0.355	0.373	0.364	0.649	0.658	0.653	0.593	0.617	0.604
+GLOBAL_MOD	0.732	0.55	0.414	0.402	0.408	0.664	0.652	0.658	0.631	0.619	0.625

同様に文脈上の要素になりやすいであろうという仮定に基づく設定である。いずれの設定においても、2.2 で述べた括弧に関するルールを用いて処理した結果を出力した。以下に考察を述べる。

LC_ONLY の NounCoverRate および ChainCoverRate の値から、名詞照応連鎖中の名詞の 94% は「語彙的連鎖+括弧の形で別名を表現している名詞ペア」に含まれ、名詞照応連鎖の 83% が「語彙的連鎖+括弧の形で別名を表現している名詞ペア」に完全に包含されていることが分かる。ペアに関する Recall も 0.876 と高い値になっている。しかし、語彙的連鎖には複数の照応連鎖が含まれることになり、LC_ONLY の設定ではそれらは 1 つの照応連鎖として出力されるため、Precision が 0.097 と極めて悪い。

次に、COMP_STR と HEAD_STR の比較について述べる。NounCoverRate および ChainCoverRate の値は HEAD_STR の方がよい。また、ペアの精度に関して両者は対照的である。COMP_STR は Precision が高く Recall が低いが、HEAD_STR は Recall が高く、Precision が低い。そして、全体の精度としては連鎖、ペア、リンクの全ての場合において HEAD_STR が COMP_STR を上回っている。これは、日本語において、特に新聞報道記事では、複合名詞などの要素の省略が頻繁に起こり、同一の照応連鎖内に同一主辞の名詞が含まれるからだと考えられる。COMP_STR の設定ではそのような照応連鎖の要素を取りこぼしてしまい、Recall が低下するものと考えられる。また、同一の主辞であっても別の参照先を持つ場合が頻繁に生じ、HEAD_STR で Precision が低くなると考えられる。

BASIC_FEATURE では、HEAD_STR と COMP_STR の場合に見られた Recall と Precision の差がなくなり、ペアとリンクの精度は COMP_STR と HEAD_STR を上回る。しかし、リンクについては大きな差はなく、連鎖についての精度は HEAD_STR に劣る。＋ GLOBAL に関しても同様の傾向がある。＋ GLOBAL では、ペアは BASIC_FEATURE を上回るが、リンクに関しては BASIC_FEATURE と大差なく、連鎖全体に関しては BASIC_FEATURE および HEAD_STR に劣っている。＋ GLOBAL_MOD では、連鎖に関しては F 値 0.408、連鎖内のペアに関しては F 値 0.658、連鎖のリンクに関しては F 値 0.625 を得た。これらの結果は、今回の実験設定の中で最良のものである。個々の名詞における素性セットに基づく BASIC_FEATURE および＋ GLOBAL よりも、修飾句や名詞の係り先の文脈情報を用いた学習を行った＋ GLOBAL_MOD がより高い名詞照応解析精度を達成していることから、文脈に関する情報の利用は、解析対象としている名詞自体からではなく、その名詞の近傍の名詞を介して用

いるほうが有効であると考える。

最後に、語彙的連鎖に含まれなかった照応連鎖について述べる。LC_ONLY において、語彙的連鎖に完全に包含される名詞照応連鎖は全名詞照応連鎖 169 中 141 であった。語彙的連鎖に包含されない名詞照応連鎖は 28 存在しており、そのうち語彙的連鎖のペアを構成できなかったために連鎖の要素が欠けてしまった場合は 20 事例、1 つの名詞照応連鎖が別々の語彙的連鎖に分割されてしまった場合が 8 事例あった。これらは、本稿におけるシステムのフレームワーク内で処理のできなかったものである。その内訳を表 3 に示す。表中の「同格」などの例を以下に列挙する。

表 3　語彙的連鎖に含まれなかった照応連鎖の内訳

連鎖形成不可	同格	2
	類義語	5
	省略	1
	括弧認識誤り	4
	数量認識誤り	4
	文字列認識誤り	4
本来の連鎖が分離	複数語彙的連鎖間で言い換え	8

「同格」は、語と語の間に生じた言い換えの関係を正しく認識できなかった場合である。我々のシステムは、「土佐日記は、紀貫之が日記風に記した紀行文である。」の「土佐日記」と「紀行文」の間の関係を認識できなかった。また、「類義語」とは、語と語の間に意味的な結束性はあるが分類語彙体系では認識できなかった場合であり、「装い」と「ファッション」などの事例があった。「省略」は、名詞に形態的な省略が生じた結果、文字列マッチングに失敗した場合である。「高槻」と「同市」などの場合である。「括弧認識誤り」は、括弧外の語と括弧内の語の対応関係の抽出に失敗した場合である。「私塾「知新館」(後の知来館)」における「私塾」と「知新館」と「知来館」など括弧構造が複雑な場合に連鎖を構築できなかった。「数量認識誤り」は、語と語の包含関係について認識できなかった場合であり、「豚 2000 頭」と「子豚 100 頭と親豚 1900 頭」のなどの場合である。「文字列認識誤り」とは、文字列のマッチングに失敗した場合である。「米テレビ各局」と「CBS 各テレビ局」については、形態素解析結果の単位が異なり、かつ、互いが前方マッチも後方マッチもしないため、語彙的連鎖として認識できなかった。

5. まとめと今後の課題

新聞報道記事 22 記事を対象に、語彙的連鎖から名詞照応連鎖を抽出する実験を行った。既存の名詞照応解析手法で用いられてきた素性セットに加えて、名詞の修飾句および係り先の名詞から「文脈全体に関する情報」についての素性を新聞報道記事 60 記事から取得し、語彙的連鎖中の名詞ペアが同じ参照先を持つか否かの問題を解く分類器を SVM を用いて構築した。その分類器で語彙的連鎖を分割したところ、連鎖に関しては F 値 0.408、連鎖内のペアに関しては F 値 0.658、連鎖のリンクに関しては F 値 0.625 を得た。また、実験結果から、日本語新聞報道記事では、複合名詞の省略が多く、それによって連鎖全体に対する評価とリンクに対する評価が悪くなっていることが分かった。また、文脈に関する情報は、個々の名詞に関する情報よりも、その名詞の修飾句や係り先の名詞から得られた情報が、より有効に働くという結果が得られた。また、名詞照応連鎖中の名詞の 94% は「語彙的連鎖+括弧の形で別名を表現している名詞ペア」に含まれ、名詞照応連鎖の 83% が「語彙的連鎖+括弧の形で別名を表現している名詞ペア」に完全に包含されているという結果が得られた。

以上から、報道記事に関しては「語彙的連鎖に含まれる照応連鎖」と「語彙的連鎖に含まれない照応連鎖」に対する処理を分ける手法は有効であると考えられる。

今後、形態的な語彙的連鎖では捉えきれない名詞照応連鎖を以下の方策で取り扱い、より精度の高い名詞照応システムの構築を目指す。

@　複合名詞の省略の過程を考慮した連鎖全体のモデルを構築する

@　語彙的連鎖上では捉えきれない名詞照応連鎖ペアに対しては、言い換え事例の収集や括弧認識ルールの精査など個別対応を行う

文　献

[1] N. Ge, J. Hale, E. Charniak. A Statistical Approach to Anaphora Resolution *Proceedings of the Sixth Workshop on Very Large Corpora*, pp.161-170, 1998.

[2] H.A.K. Halliday and R. Hasan. Cohesion in English *Longman*, 1976.

[3] L. Hirschman and N. Chinchor. MUC-7 Coreference Task Definition, Version 3.0. *Proceedings of MUC-7*, http://www.itl.nist.gov/iaui/894.02/related _projects/muc/proceedings/muc _toc.html, 1998.

[4] J. Morris and G. Hirst. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of

Text. *Computational Linguistics*, Vol.17, No.1, pp.21-48, 1991.

[5] W. M. Soon, H. T. Ng, and D. C. Y. Lim. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, Vol.27, No.4, pp.521-544, 2001.

[6] M. Strube, S. Rapp, C. Muller. The Influence of Minimum Edit Distance on Reference Resolution. *In the Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP)*, pp.312-319, 2002.

[7] V. N. Vapnik. Statistical Learning Theory Adaptive and Learning Systems for Signal Processing Communications and contorol, John Wiley and Sons, 1998.

[8] X. Yang, G. Zhou, J. Su and C. L. Tan. Improving Noun Phrase Coreference Resolution by Matching Strings. *In Proceedings of 1st International Joint Conference of Natural Language Processing (IJCNLP04)*, pp.326-333, 2004.

[9] X. Yang, J. Su, G. Zhou and C. L. Tan. An NP-Cluster approach to coreference resolution. *In Proceedings of 20th International Conference on Computational Linguistics (COLING04)*, 2004.

[10] 飯田龍, 乾健太郎, 松本裕治, 関根聡. 最尤先行詞候補を用いた日本語名詞同一指示解析. 情報処理学会論文誌,Vol.46, No.3 2005.

[11] 飯田龍, 乾健太郎, 松本裕治. 先行文脈と局所文脈を併用した照応性判定モデルの学習. 言語処理学会第 11 回年次大会, 2005.

[12] 国立国語研究所：分類語彙表. Vol 国立国語研究所資料集. **秀英出版**, 1993.

[13] 工藤拓, 松本祐治. Support Vector Machine を用いた Chunk 同定. 自然言語処理, Vol.9, No.5, pp.3-21, 2002.

[14] 増市博, 大熊智子. Lexical Functional Grammar に基づく実用的な日本語解析システムの構築. 自然言語処理, Vol.10, No.2, pp. 79-109, 2003.

[15] 村田真樹, 黒橋禎夫, 長尾真. 名詞の指示性を利用した日本語文章における名詞の指示対象の推定. 自然言語処理, Vol.3, No.1, pp.67-81, 1996.

[16] 村田真樹, 黒橋禎夫, 長尾真. 表層表現を手がかりとした日本語名詞句の指示性と数の推定. 自然言語処理, Vol.3, No.4, pp.31-48, 1996.