Web 検索を用いた複合名詞同定

沢井 康孝 , 山本 和英
長岡技術科学大学 電気系
{sawai,ykaz}@nlp.nagaokaut.ac.jp

1　はじめに

複合名詞を扱う場合、辞書に登録されている固有表現以外は形態素単位で抽出される。しかし専門用語や製品名などを収集する際には複数の形態素から成る複合名詞をどのように扱うか問題がある。このような複合名詞を含んだ用語を網羅的に収集する際に、複合名詞としての妥当性を考慮した抽出が必要である。これはテキストマイニングなどにおいて、処理の単位をどのような長さにすれば良いのかという問題に相当する。このような問題は個人により抽出する範囲が異なることもあり、分野によっても最適な長さは異なる。例えば複合名詞を抽出する際に「低価格ノイズキャンセルヘッドホン」は「低価格ノイズキャンセルヘッドホン」、「低価格/ノイズキャンセルヘッドホン」、「低価格/ノイズキャンセル/ ヘッドホン」と複数の分割点が存在し、どこまでを複合名詞とするか判断が揺れる。

そこで我々は Web の検索エンジンを大規模なコーパスに見立てて、対象とする複合名詞が複合名詞として妥当であるかどうかを算出することにした。評価では評価者に複合名詞としての正解・不正解の例を提示した上で複合名詞として妥当である・妥当ではないの評価を行った。これにより一定の尺度で生成された複合名詞がどの程度妥当であるか調査しその結果を報告する。

以下、2 節では関連研究、3 節では対象とするコーパス、4 節では複合名詞候補の定義、4 節では複合名詞の同定方法、5 節では複合名詞同定の実験と評価について述べる。

2　関連研究

複合名詞は様々な所で考慮すべき問題として挙げられ多くの研究がされている。複合語の作成について中川ら [1] はコーパス中の語の連接頻度を用いた手法を提案している。また峠らは大規模テキストからの意見、評判情報の抽出手法 [3] のドメイン特徴語抽出において検索エンジンのヒット件数を用いた手法を用いている。佐々木ら [2] は Web を利用して専門用語辞書の構築を行っている。本研究では検索エンジンのヒット件数を用いた手法をもとに検索ヒット件数の割合を用いた方法を提案し検索エンジンを用いた 2 つの方法について結果を比較、考察を行う。

3　対象コーパス

複合名詞の網羅的収集を目的に、どのような方法が良いのか調査することにある。そのため複合名詞を抽出する対象として製品名や未知の用語が多く含まれていることが望ましい。よって抽出対象には Web 文書を用いることにした。本研究では「livedoorBlog[2)]」及び「価格コム、クチコミ掲示板、ユーザーレビュー [3)]」から複合名詞の候補を抽出し複合名詞の同定処理を行った。

4　複合名詞候補の定義

複合名詞として表現される語は「誕生日パーティー」や「ニューバージョン」、「人工衛星」、「エアダクト」、「入隊希望者」など複数の形態素で構成される。本節では複合名詞の構成形態素について述べる。

本研究では、入力されたテキストデータに対して形態素解析を行い、複合名詞候補を抽出する。本研究では形態素解析器には「茶筌 [1)]」を用い, 品詞体系は「IPA 品詞体系辞書 (ipadic)」に準ずる。形態素解析結果より, 以下に示す品詞が連続した部分を複合名詞の候補とする。

- 名詞-一般, 名詞-サ変接続, 名詞-固有名詞, 名詞-接尾, 接頭詞, 未知語, 記号（アルファベット）

但し、複合名詞候補として誤りが多く発生するパターンについては例外規則としてルールを作成し複合名詞の候補としない。

- 英数字以外の記号が含まれている複合名詞候補（♪、＄等）
- 名詞-数のみで構成される複合名詞候補（電話番号、IP アドレス他）

本研究では、本節で作成した複合名詞候補に対して複合名詞としての妥当性を評価して分割点を決定する (例 1)。

例 1) 複合名詞の同定
　　複合名詞候補：公開予定ベンチャー企業
　　⟶「公開予定」、「ベンチャー企業」

複合名詞の妥当性を判定するため、Web 検索エンジンの組み合わせた手法で複合名詞同定を行う。

5　複合名詞の同定方法

複合名詞候補には正しい結合でない場合が存在する。

例 2) ドライブレポ楽しみ, ＣＸ系最強, 再セットアップ完了, 連ドラＣＭカット

語として意味をなさない場合や表現として妥当ではない語については排除すべきである。本研究では Web 検索エンジンを利用した手法で複合名詞同定を行う。Web 検索エンジンを利用した複合名詞同定については峠ら [3] はドメイン特徴語抽出の際に検索エンジンのヒット件数の閾値によって行っている。本研究では峠らのヒット件数の閾値の他に検索ヒット件数の割合を用いた方法を提案し 2 種類の方法の比較を行う。

複合名詞の同定方法

- ヒット件数の閾値による手法
- ヒット件数の割合による手法

5.1　ヒット件数の閾値による手法

検索エンジンを用いたヒット件数の閾値による複合名詞同定方法について述べる。

STEP1　対象の複合名詞候補から同定候補を作成する
　　同定候補は複合名詞候補に含まれる形態素から作成する。形態素に分割した後、複合名詞を構成する全ての組み合わせを用いる、例えば複合名詞候補が「ＤＮＳサーバアドレス手動設定」の場合、複合名詞候補から作成された同定候補は 10 種類である (例 3)。

STEP2 同定候補の連接ヒット件数を求める

各同定候補の検索ヒット数を求める。対象とする同定候補を全て連接させた単語を含むテキストの検索を行う。検索ヒット数が閾値を超えるとき妥当性がある複合名詞として扱う。

同定候補が「ＤＮＳサーバアドレス」の場合、連接させた “ＤＮＳサーバアドレス” で検索を行う。検索対象 q の検索ヒット数を $HIT(\text{“}q\text{”})$ として、同定候補を W,W に含まれる形態素を $\{w_1, w_2 ... w_n\}$ として連接検索ヒット数を $HIT\_CO(W)$ として式 (1) に示す。

$$HIT\_CO(W) = HIT(\text{“}w_1 w_2 ... w_n\text{”}) \quad (1)$$

STEP3 最長一致法による複合名詞同定

検索ヒット数をもとに、複合名詞の同定を行う。妥当性がある複合名詞の先頭からの最長一致法により決定する。最長一致で同定候補が複数ある場合、連接検索ヒット数がより多い方を優先する。

例 3) 検索ヒット数
(ＤＮＳサーバアドレス手動設定:3), (ＤＮＳサーバアドレス手動:3), (サーバアドレス手動設定:3), (ＤＮＳサーバアドレス:49000), (サーバアドレス手動:3), (アドレス手動設定:2960), (ＤＮＳサーバ:201000) (サーバアドレス:31400), (アドレス手動:3630), (手動設定:106000) ＊数字は連接検索ヒット件数

閾値を 1000 件としたとき例 3 は「DNS サーバーアドレス」と「手動設定」の複合名詞が同定される。

5.2 ヒット件数の割合による手法

検索エンジンの連接検索と AND 検索のヒット件数の割合を用いた複合名詞同定方法について述べる。

STEP1 対象の複合名詞候補から同定候補を作成する

STEP2 同定候補の AND ヒット件数を求める

各同定候補を全て含むテキストを検索する。「ＤＮＳサーバアドレス手動設定」の場合、“ＤＮＳ AND サーバ AND ... AND 設定” の検索を行いそのヒット数を取得する。同定候補を $W = \{w_1, w_2 ... w_n\}$、W の AND 検索ヒット数を $HIT\_AND(W)$ として式 (2) に示す。

$$HIT\_AND(W) = HIT(\text{“}w_1 \, and \, w_2 \, and \, .. \, and \, w_n\text{”}) \quad (2)$$

STEP3 同定候補の連接ヒット件数を求める

各同定候補を全て連接した単語を含むテキストを検索する (式 (1))。

STEP4 最長一致法による複合名詞同定

連接検索と AND 検索ヒット数の割合 ($HIT\_CO \div HIT\_AND$) が閾値を超えるとき妥当性がある複合名詞として扱う。妥当性がある複合名詞の先頭からの最長一致法により決定する。最長一致で同定候補が複数ある場合検索ヒット数の割合が多い方を優先する。

例 4) 検索ヒット数の割合
(ＤＮＳサーバアドレス手動設定:0), (ＤＮＳサーバアドレス手動:0), (サーバアドレス手動設定:0), (ＤＮＳサーバアドレス:0.01), (サーバアドレス手動:0), (アドレス手動設定:0), (ＤＮＳサーバ:0.3) (サーバアドレス:0.04), (アドレス手動:0), (手動設定:0.1) ＊数字はヒット数の割合

閾値を 0.1 としたとき例 4 は「DNS サーバー」と「アドレス」と「手動設定」に分割される。

6 実験結果

6.1 複合名詞候補について

「livedoorBlog」及び「価格コム、クチコミ掲示板、ユーザーレビュー」の 2 種類のテキストから複合名詞の候補を抽出した結果を表 1 に示す。

表 1: 複合名詞候補抽出結果

種類	容量	異なり候補数
livedoorBlog	59M	353993
価格コム	60M	158056

但し、複合名詞候補のうち 2 形態素で構成され、接尾又は接頭が含まれている場合、容易に複合名詞と判断出来るため複合名詞候補から除いた。除かれた候補について無作為に 100 件サンプリングを行い複合名詞として意味をなさない場合や表現として妥当ではない語が含まれているかどうか評価を行った。複合名詞候補選別結果を表 2 に示す。妥当と判断された割合を複合名詞の精度として示す。

表 2: 複合名詞候補の選別

種類	除いた候補数	除いた候補の精度
livedoorBlog	59027	0.95
価格コム	25879	0.93

複合名詞の候補として 2 形態素で構成され、1 つの形態素が名詞-接尾・接頭詞である場合 0.95 程度で複合名詞として意味をなす語や表現として妥当である語という評価結果が得られた。

以降ではこの複合名詞を除いた複合名詞候補に対して評価、考察を行う。

6.2 複合名詞候補自体の評価

本研究では名詞連接を抽出して複合名詞の候補とした。複合名詞の候補自体の評価を示す。複合名詞候補を構成する形態素数別で評価を行った。評価は各形態素数ごとに 100 件を無作為に取得し、正解・不正解の例を提示した上で被験者に評価を行ってもらった。形態素数別の複合名詞の数を表 3 に示し、各形態素数別の複合名詞候補の評価を表 4 に示す。

表 3: 形態素数別の候補数

構成形態素数	livedoorBlog	価格コム
2	155585	60944
3	94116	47839
4	29983	16199
5	9533	4853
6 以上	5749	2342

複合名詞の抽出対象が Web の日記や掲示板を対象としたものであるため、名詞連接を用いると複合名詞として妥当ではないものが多く存在した。また構成する形態素数が増加すると複合名詞として扱うには妥当ではないと判断できるものが増加している。

評価者が正解、不正解とした例を例 5, 例 6、Web 特有の誤りについて例 7 に示す。例において「/」は分割点を示し「-」は形態素の区切りを示す。

例 5) 正解例：ひとくち-ビール, データ-ＢＯＸ, 独立-行政-法人, アナグロ-チューナ-内蔵-タイプ

例 6) 不正解例：ブラジル-ＶＳ-クロアチア, 今-イチ-応援, 低音-モコモコ, フジテレビ-系-月-９-ドラマ

表 4: 複合名詞候補の精度

構成形態素数	2	3	4	5
livedoorBlog	0.68	0.51	0.44	0.29
価格コム	0.69	0.57	0.47	0.38

例 7)Web 特有の誤り例：部長-キタ, ヤッパリ-うめ-ぇ, タイトル-うろ覚え-ビート

6.3 ヒット件数の閾値による複合名詞同定結果

ヒット件数の閾値を用いた複合名詞同定処理の結果を示す。検索エンジンには「Google[4)]」を使用し、検索エンジンによる閾値は 1000 件に設定して実験を行った。複合名詞に分割した後の複合名詞の形態素数別で評価を行った。各形態素数ごとに同定された複合名詞を無作為に取得し、その同定精度を表 5、実際の同定結果を例 8 に示す。形態素数 2,3 については 100 件のサンプル、形態素数 4,5 については 100 件以下のサンプルである。

表 5: ヒット件数：複合名詞の評価

構成形態素数	2	3	4	5
livedoorBlog	0.72	0.82	0.67	0.65
価格コム	0.81	0.77	0.70	0.72

例 8) 実際の同定結果：東急-電鉄/こどもの国-駅, 電源-用/ライン-ノイズ-フィルター, 全国-チェーン/某-電気-店, ＤＮＳ-サーバ-アドレス/手動-設定

検索ヒット数を考慮するという簡易な方法であるが、同定精度は比較的良好であった。検索ヒット数に閾値を設定する同定手法での同定間違いについては次の 2 つが挙げられる。

- 頻度の少ない語句
  固有名詞などは検索エンジンヒット件数が極端に少ない場合もあり複合名詞として判定できない
- 最長一致による誤り
  長い複合名詞候補「当サイト価格ドットコム」では「当サイト」と「価格ドットコム」と同定されてほしい。しかし「サイト価格ドットコム」が閾値を超えているため「当/サイト価格ドットコム」と分割される。

6.4 ヒット件数の割合による複合名詞同定精度

ヒット件数の割合を用いた複合名詞同定処理の結果を示す。検索エンジンには「Google」を使用した。本節では複合名詞同定の閾値を変化させた時の評価と複合名詞を構成する形態素数別の評価を示す。

複合名詞に分割する際の閾値を変化させ閾値別で分割の評価を行った。各閾値ごとに 100 件を無作為に取得した時の同定精度を図 1 に示す。

複合名詞に分割した後、複合名詞の形態素数別で評価した結果を示す。複合名詞同定には閾値 0.1 及び閾値 0.05 を採用した。各形態素数ごとに同定された複合名詞を無作為に取得し、その同定精度を表 6 及び表 7 に示す。実際の同定結果を例 9 に示す。形態素数 2,3 については 100 件のサンプル、形態素数 4,5 については 100 件以下のサンプルである。

表 6: ヒット割合：複合名詞の評価：閾値 0.1

構成形態素数	2	3	4	5
livedoorBlog	0.92	0.83	0.85	0.76
価格コム	0.84	0.85	0.75	0.81

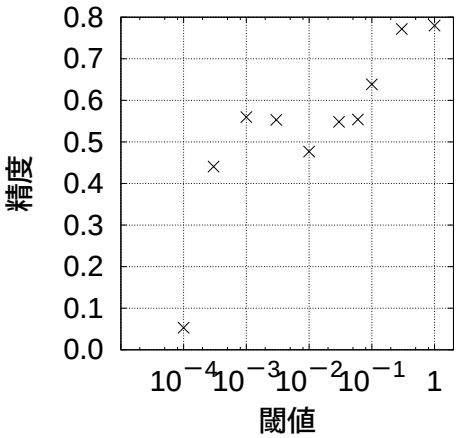

図 1: ヒット割合の閾値と分割精度

表 7: ヒット割合：複合名詞の評価：閾値 0.05

構成形態素数	2	3	4	5
livedoorBlog	0.95	0.90	0.80	0.82
価格コム	0.87	0.84	0.79	0.81

例 9) チューナー/無料/バラまき , ディスク/外周部/付近, 在庫処分/大特価

ヒット割合による方法はヒット件数による方法と比較して複合名詞の構成形態素数が短くなる傾向になった。実際に複合名詞候補の形態素数 2・3・4・5 からそれぞれ 100 件ずつ取り出し複合名詞同定処理を行った。同定処理を行って得ることが出来た複合名詞を構成形態素数別にカウントした。結果を表 8 に示す。

表 8: 構成形態素数の調査

構成形態素数	2	3	4	5
複合名詞候補	100	100	100	100
ヒット件数	392	197	55	20
ヒット割合	409	99	25	11

表 8 の結果より、作成される複合名詞の構成形態素数はヒット割合の方が長い形態素数の語句を許さない傾向にあり、ヒット割合を用いた方法ではヒット数を用いた方法より語句が細かく分割される。

検索ヒット割合を考慮するという方法では、検索ヒット件数を考慮した同定より良好な結果であった。検索ヒット割合に閾値を設定する同定手法での同定誤りについては次の 2 つが挙げられる。

- ページ内共起である問題
  良く使用される語句（リンク、両など）との同定では検索エンジンヒット件数が極端に多く相対的に割合が小さくなるため複合名詞として判定できない場合がある。その一方で検索数では分割できなかった「ＳＭＡＰ草なぎ剛」は本手法では「ＳＭＡＰ/草なぎ剛」に分割されている。
- 細かく分割される問題
  長い複合名詞候補「量的金融緩和政策」が本手法では「量的/金融緩和/政策」と同定される。固有名詞的であるため 1 つで扱いたい。しかし「量的金融緩和政策」を構成する形態素それぞれが金融の用語に属するため同一文書に表れやすく分割される結果となった。

表 9: 検索ヒット件数と検索ヒット割合の分割点の比較

検索ヒット数 閾値 1000	検索ヒット割合 閾値 0.05
身体-障害-者-手帳-等	身体-障害-者-手帳/等
住宅-用-アルミ/建材/カラー-サンプル	住宅-用/アルミ-建材/カラー-サンプル
ＵＳＢ-サウンド-デバイス/サウンド-カード	ＵＳＢ/サウンド-デバイス/サウンド-カード
液晶-プロジェクター/接続-用-アダプタ	液晶-プロジェクター/接続-用/アダプタ
マスメディア/元-社員/かつ-有名人	マスメディア/元-社員/かつ/有名人
外部-入力-音声/レベル-コントロール	外部-入力/音声/レベル/コントロール
データー/通信-用-カード/型	データー/通信-用/カード-型
等-速-ダビング/完了-時	等-速-ダビング/完了-時
勝ち-組-ｖｓ-負け-組	勝ち-組/ｖｓ/負け-組
円/洗濯-機-エアコン-サイクル-ドラム	円/洗濯-機/エアコン-サイクル-ドラム
密閉-型/機種-名/パナソニック-ＲＰ	密閉-型/機種-名/パナソニック-ＲＰ
シガー-ライター-ソケット-用/電源-コード	シガー-ライター-ソケット/用-電源/コード
ＤＶＤ-レコーダー/ク-チコ-ミ-掲示板	ＤＶＤ-レコーダー/ク-チコ-ミ-掲示板

6.5 複合名詞分割点の比較

本節では 2 つの同定方法について比較した結果を表 9 に示す。

検索ヒット割合では「等-速-ダビング/完了-時」の様に複合名詞候補を正しく分割することを維持しつつ、検索ヒット数では出来なかった「勝ち-組/ｖｓ/負け-組」の分割が可能となった。しかし「液晶-プロジェクター/接続-用/アダプタ」や「ＵＳＢ/サウンド-デバイス/サウンド-カード」の様に検索ヒット件数より細かい単位での複合名詞として分割される傾向にある (例 10)。

例 10) チケット先行発売/開始 ⟶ チケット先行/発売開始, 史上- 最年長/ワールド-シリーズ-出場/記録 ⟶ 史上-最年長/ワールド-シリーズ/出場-記録

2 つの手法の特徴をまとめて次に示す。

検索ヒット件数

- 形態素数が多い複合名詞を同定しやすい
- 「住宅-用-アルミ/建材」のようにある程度連接する語句に誤りが存在

検索ヒット割合

- 検索ヒット件数より短い形態素数で同定しやすい
- 共に出てくる文書数を考慮しているため検索ヒット件数より複合名詞として高い精度で同定できる
- 文書数が少ない固有表現を同定できる (「インディペンデント-ワールド-ジュニア-ヘビー-級」等)

6.6 Web 特有の語句

本研究では複合名詞同定を目的に行ったが、同定結果から得られた語句について述べる。本手法で得られた結果の一部を例 11 に示す。

例 11) 同定結果から得られた語句
ラ-ラ-バイ/ライブ, なめ-ら-かさ/等, こ-ー-ゆ-ー/大人, お-ｋ/カナ/野球-応援, めん-ど-くせ-ぇ/ポケモン, ハズ-カシ-ク-ナイ/程度

例 11 に示した、「なめらかさ」などの形態素解析誤りや「めんどくせぇ」などのブログや掲示板で使用される表現を収集出来る可能性がある。ブログや掲示板で使用される表現を収集することでブログや掲示板などの口語で書かれている文書に対する解析の精度向上やテキストマイニング等の解析の補助が可能ではないかと考えている。

6.7 複合名詞の単位

複合名詞の同定を行う 2 つ手法について実験及び評価、観察を行った。実際に必要とする複合名詞の長さ（構成形態素数）は対象の問題によって変化するものだと考えている。例えば「サーバー代稼ぎ」と「サーバー代/稼ぎ」の内どちらを採用するかの対象とする問題の処理内容によって選択されるべきである。従って実際に対象とする問題を想定した上で、2 種類の複合名詞の同定方法のうち問題に対してどちらの方法がより処理を効率的にできるか調査を行いたい。

7 おわりに

Web 検索を用いて複合名詞同定を行う手法について調査を行った。同定手法では検索ヒット数自体に閾値を用いる方法と検索ヒット数の割合を用いる方法の 2 種類を用いた。

その結果、ブログや掲示板から複合名詞を高い精度で同定できることがわかり、検索ヒット数の割合を用いた方法は検索ヒット数自体に閾値を用いる方法より短い構成で同定し、構成する形態素数が多い固有名詞を同定するには検索ヒット件数を用いた方が妥当な同定できると考えている。

両者共に異なる問題を抱えているため、どのような複合名詞を用いたいか処理に応じて使い分けを行いうまく活用することが必要となってくる。

使用したツール及び言語資源

1) 形態素解析器 ChaSen, Ver.2.3.3, 奈良先端科学技術大学院大学 松本研究室, http://chasen.naist.jp/hiki/ChaSen/.
2) livedoor Blog , http:/blog.livedoor.com//.
3) 価格.com クチコミ掲示板 ユーザーレビュー , http://bbs.kakaku.com/bbs/.
4) 検索エンジン Google , http://www.google.co.jp .

参考文献

[1] 中川裕志, 湯本紘彰, 森辰則. 出現頻度と連接頻度に基づく専門用語抽出. 自然言語処理, Vol. 10, No. 1, pp. 27–45, 2003.

[2] 佐々木靖弘, 佐藤理史, 宇津呂武仁. ウェブを利用した専門用語集の自動編集. 言語処理学会第 11 回年次大会発表論文集, pp. 895–898, 2005.

[3] 峠泰成, 山本和英. 大規模テキストからの意見・評判情報の抽出手法. Master's thesis, 長岡技術科学大学大学院, 2006.