B1IM2020

修士論文

形態素解析器の新規ドメインへの適応方法に関する研究

神谷　一輝

2013 年 1 月 16 日

東北大学 大学院
情報科学研究科 システム情報科学専攻

本論文は東北大学大学院情報科学研究科に
修士 (工学) 授与の要件として提出した修士論文である。

神谷　一輝

審査委員：

乾　健太郎 教授　　（主指導教員）
田中　和之 教授　　（副指導教員）
篠原　歩 准教授　　（副指導教員）
岡崎　直観 准教授　（副指導教員）

形態素解析器の新規ドメインへの適応方法に関する研究*

神谷　一輝

内容梗概

本論文では、形態素解析器内の辞書に新規ドメインの未知語を追加することで形態素解析器の精度向上を目指した。しかし、未知語抽出の研究の数は多く、同じ条件で比較している研究が少ない。そのため、未知語抽出の先行研究の比較・分析する。未知語抽出を複合語抽出と単名詞抽出にわけて、複合語抽出では複合語の専用用語の持つ性質を考慮して用語性を重視した TF-IDF、単位性を重視した C-value を使用して実験を行った。単名詞抽出では、単語の境界を判定する事で未知語を判別するをおこなう手法が多く、その中からに隣接する文字の異なり数をエントロピー使用する手法と、隣接文字自体の性質を利用した手法の２つを使用して実験を行った。４つの手法に対して比較分析を行い各手法の比較・分析し、手法の問題点を明らかにした。

キーワード

形態素解析、未知語抽出、新規ドメイン

*東北大学 大学院 情報科学研究科システム情報科学専攻 修士論文, B1IM2020, 2013 年 1 月 16 日.

目 次

図 目 次

表 目 次

1　はじめに

コンピュータで文章を処理するとき、日本語には英語の空白のような単語間の明確な区切りが無いため入力文を形態素（これ以上小さくする事ができない最小単位の単語) に分割し、品詞を付与する形態素解析という処理が必要となる。

形態素解析の技術は、固有表現抽出、構文解析などの言語処理の入力となるだけでなく、情報検索やテキストマイニング等にも使用される自然言語処理の基盤となる研究である。そのため、形態素解析器の精度が研究に与える影響は大きく、高い精度の形態素解析の実現が望まれている。

現在、形態素解析器というツールによって高い精度で形態素解析が行われている。形態素解析器は、形態素解析器が採用している機械学習モデルと学習コーパスによるパラメータ推定から出力されるパラメータ、そして形態素情報が付与された辞書を使用して、文章の形態素解析を行っている。代表的な形態素解析器として、パラメータの学習に隠れマルコフモデル (HMM) を採用している ChaSen[1]、条件付き確率場 (CRF) を採用している MeCab[2] 等がある。

しかし、これらの形態素解析器はパラメータの学習コーパスに含まれない分野のテキストを解析する際や形態素解析器内の辞書に含まれていない単語が含まれる新規ドメインのテキストを解析すると解析精度が落ちてしまう。先行研究では、一般的なテキストに対しては 96〜99 %の精度 [2] が有るのに対して、新規ドメインでは 80 % [3] 近くまで落ちている結果も出ている。これは、形態素解析器内の辞書に登録されていない未知語（辞書に登録されていない単語）が出現した場合や、形態素解析に多義性が存在する場合に誤った解析を行うためである。

この問題に対して、先行研究では新規ドメインの未知語を形態素解析器内に追加するという方法がある。しかし、各分野において専門用語や新語などの未知語が日々増加していくため、人手で獲得して、辞書に登録するのはコストが高くなる。そのため、未知語をコーパスから自動的に獲得する必要がある。

しかし、未知語抽出の先行研究の数は膨大であり、単名詞と複合語を両方行っている研究、同じ条件で実験を行っているものが少ないため、どの手法が形態素解析器の辞書に追加すべき未知語を出力できるのかわかりにくい。

そのため、本研究は、既存の形態素解析器では解析できない、新規ドメインの

未知語に対して、新規ドメインのコーパスを使用して、先行研究の手法を比較・分析を行う。

本論文の構成は以下の通りである。まず、２節において、まず、抽出すべき未知語について述べる。３節では、先行研究をあげる。4節では、先行研究の中から使用する手法の説明を行う。5節では、4節で述べた手法の実験の設定を述べ、評価実験を行い、その実験結果を述べ、結果の比較・分析する。6節では、結果の分析から考察する。最後７節でまとめを述べる。

2　未知語の形態素解析

本節では、抽出対象である未知語について述べる。未知語とは、辞書に辞書に存在しない単語のことをいう。人名、地名、商品名などの固有名詞、擬音、顔文字などが未知語である。本研究で抽出したい未知語は、おもに固有名詞で、形態素解析を行う上で、未知語は２つの種類に分かれている。１つ目は、”けいれん”、”グルミン”、”遺伝子”など１つの形態素で構成されている「単名詞」、２つ目は、”顔面__けんれん”、”軟体__動物”などの２つ以上の形態素で構成されてる「複合語」である。それぞれの形態素解析器の辞書に存在しない場合どのような形で出力されるのか述べる。

2.1　複合語

以下に、複合語の形態素解析結果を示す。

複合語の形態素解析例

例：コレステロール血症が**動脈硬化**のリスクになる

形態素解析例

コレステロール 名詞, 一般,*,*,*,*, コレステロール, コレステロール, コレステロール,,

血 名詞, 一般,*,*,*,*, 血, チ, チ,,

症 名詞, 接尾, 一般,*,*,*, 症, ショウ, ショー,,

か 助詞, 副助詞／並立助詞／終助詞,*,*,*,*, か, カ, カ,,

名詞, 固有名詞, 組織,*,*,*,*

動脈 名詞, **一般**,*,*,*,*, **動脈**, **ドウミャク**, **ドーミャク**,,

硬化 名詞, **サ変接続**,*,*,*,*, **硬化**, **コウカ**, **コーカ**,,

の 助詞, 連体化,*,*,*,*, の, ノ, ノ,,

リスク 名詞, 一般,*,*,*,*, リスク, リスク, リスク,,

に 助詞, 格助詞, 一般,*,*,*, に, ニ, ニ,,

なる 動詞, 自立,*,*, 五段・ラ行, 基本形, なる, ナル, ナル, なる/成る,

例は、形態素解析器 MeCab を利用した形態素解析結果である。形態素と品詞情報を出力している。複合語の形態素解析結果から、未知語となる複合語は 1 つの語りとして出力されておらず、複数の形態素で出力されていることがわかる。しかし、複合語の場合は分かれた未知語の形態素の品詞は名詞系として分類されることが多く、形態素を品詞を利用して合成する事で抽出可能である [4]。

2.2 複合語

以下に単名詞の形態素解析結果を示す。
単名詞抽出の形態素解析例
例：痴呆症患者では高率に**せんもう**状態を生じる
形態素解析例
痴呆 名詞 , 一般,*,*,*,*, 痴呆, チホウ, チホー,,
症 名 詞, 接尾, 一般,*,*,*, 症, ショウ, ショー,,
患者 名詞 , 一般,*,*,*,*, 患者, カンジャ, カンジャ,,
で 助 詞, 格助詞, 一般,*,*,*, で, デ, デ,,
は 助 詞, 係助詞,*,*,*,*, は, ハ, ワ,,
高率 名詞 , 一般,*,*,*,*, 高率, コウリツ, コーリツ,,
に 助 詞, 格助詞, 一般,*,*,*, に, ニ, ニ,,
せ 動詞, **自立**,*,*, **サ変・スル**, **未然ヌ接続**, **する**, **セ**, **セ**,,
ん 助 動詞,*,*,*, **不変化型**, **基本形**, **ん**, **ン**, **ン**,,
もう 副詞 , **一般**,*,*,*,*, **もう**, **モウ**, **モー**,,
状態 名詞 , 一般,*,*,*,*, 状態, ジョウタイ, ジョータイ,,
を 助 詞, 格助詞, 一般,*,*,*, を, ヲ, ヲ,,
生じる 動詞, 自立,*,*, 一段, 基本形, 生じる, ショウジル, ショージル, しょうじる/生じる,

例は、形態素解析器 MeCab を利用した形態素解析結果である。形態素と品詞情報を出力している。単名詞の形態素解析結果では、「せんもう」という単語が形態素解析器内の辞書に掲載されていない場合、複合語の形態素解析結果とは違

い、名詞ではなく、動詞、助詞、副詞とわけられて出力されてしまうため、複合語の場合と違い、品詞を利用した抽出は難しい。

3　関連研究

新規ドメインに対する形態素解析器の精度向上の先行研究で、形態素解析器内の辞書を追加する手法 [5] では、まず、新規ドメインでの未知語候補抽出が行われ、その後、獲得した未知語の最適化を行うためにスコア付けを行う。従って、本節では、新規ドメインでの未知語抽出、抽出した未知語のスコア付けについて述べる。

3.1　新規ドメインでの未知語抽出

未知語抽出、また、専門用語抽出は従来より盛んに行われている。専門用語には重要な性質として、影浦ら [9] によると、Termhood(用語性)、unithood（単位性）が上げている。用語性とは、専門用語が、領域あるいは対象分野固有の概念と関連する度合い、単位制とは、専門用語において語順、構文構造、意味的関係のある関係等が安定して用いられる度合いの事である。また、影浦らの研究をふまえて、湯本ら [6] は、書き手の持っている概念に注目して、専門用語の構造は用語性、単位性と深く関わっていると述べている。

また、第２節より、未知語、専門用語の他の性質として、これ以上分けることのできない名詞の単名詞と、単名詞などが複数固まって更生される複合語の２つの種類に分けることができる。

このように、未知語、専門用語にはそれぞれ特性があり、単名詞を抽出する場合と、複合名詞を抽出する場合で方法が異なるため、２つに分けて説明を行う。

3.1.1　複合語の抽出

複合語抽出は、単名詞抽出と比べて先行研究 [6, 4, 7, 8, 10, 11, 12] が多い、未知語や専門用語はほとんどが複合語であることが多いためである。先行研究として、スコア付けでは、湯本ら [6]、小山ら [4]、池野ら [7] はコーパスや Web などの膨大なデータの文章を形態素解析を行い、品詞情報を利用して未知語、専門用語となりうる候補を抽出して、スコア付けを行っている。湯本ら [6] の研究では抽

出した候補に隣接する単名詞の頻度に注目して、ある単名詞が複合名詞を形成するために連続する名詞の頻度を利用してスコア付けを行い、高い精度での抽出を行っている。小山ら [4] は、候補を抽出する際に、抽出対象となる専門用語の候補の内部構造と、テキスト内での候補の前後に対する接続関係に制約を設けて、適合率を下げること無く専門用語の抽出を可能にしている。池野ら [7] は、候補を抽出する際に、統計的に候補を抽出し、スコア付けの際には、専門用語の属性を判別するという手段で専門用語を抽出しており、高い再現率を可能にしている。三浦ら [8] は、他の 3 つの手法とはことなり、低頻出単語の抽出を考えており、文字列があたえられている時、文字列を更生する n-gram の部分文字列を抽出して、文字列を更生する部分文字列及び周辺文字列をパープレキシティを用いてスコア付けを行っている。また、一般的なスコア付け方法として、用語性に重点を置いた研究のスコア付け方法としてコーパス内に出現する複合語の頻度と複合語の逆文書行列を利用した TF-IDF[10] や専門コーパスと一般コーパスでの複合語の出現頻度の差を利用した Weirdness[10, 11] がある。単位性に重点を置いた研究では、C-value[10, 12] が有用な手段である。

3.1.2 単名詞の獲得

単名詞の抽出は、単名詞のみを抽出しようとする研究は少なく、単名詞と複合語の両方を抽出する研究 [13, 14, 15, 16] を述べる。新規ドメインで単名詞を抽出する場合、抽出すべき単名詞が辞書に掲載されていないため、複合語の場合と異なり、未知語の単語としての境界に注目した手法が多く、複合語の場合とは異なり、形態素解析器を利用した方法を利用しない場合が多く、n-gram 統計を利用している。また、複合語抽出とは異なり、専門用語の抽出ではく、名詞、名詞句といった文字列の抽出を行っている研究も参考にしている。森ら [13] は、n-gram 統計を用いてコーパスからの単語の抽出とその単語がどの品詞に属するのか推定を行っている。下畑ら [14] は、品詞情報、隣接文字情報の付与した学習データを利用して、名詞の前後に隣接する文字の異なり数をエントロピーを利用することで名詞の抽出を行っている。長尾ら [15] は、森らと同様に n-gram 統計を利用して、語句の抽出を行っており、特に、n-gram 統計を使用するテキストの規模に注目し

て研究を行っている。鍛冶ら [16] は、文脈情報を利用した識別モデルを用いて、未知語抽出を行っている。

3.2 研究で使用する先行手法

本研究では、複合語抽出の比較をするために、用語性重視している手法と単位性を重視している手法を一つずつ使用する。用語性重視の手法では「TF-IDF」を単位性重視の手法では「C-value」を使用して複合語を抽出して比較を行う。

単名詞抽出の手法も同様に、先行研究手法から、隣接する文字の異なり数をエントロピーを利用してスコア付けた手法、隣接する文字を学習することでスコア付けを行う手法の２つを参考にして単名詞を抽出して比較する。

4　未知語抽出手法

4.1　手法全体の流れ

本節では、使用した先行研究の具体的な実験手法について述べる。第２節より、複合語抽出では、形態素解析を利用して未知語の候補を抽出できるため、複合語抽出と単名詞抽出では、未知語のスコア付けを行う際の未知語候補抽出の方法に違いがでる。そのため、単名詞抽出と複合語抽出を別々行う。また、既存研究の性能を比較するため、本研究で用いる手法の多くは既存研究の手法を参考・再現したものである。4.2 節では、複合語抽出について、4.3 節では、なぜ単名詞の必要性、単名詞抽出の手法について述べる。

4.2　複合語抽出手法

4.2.1　複合語抽出の流れ

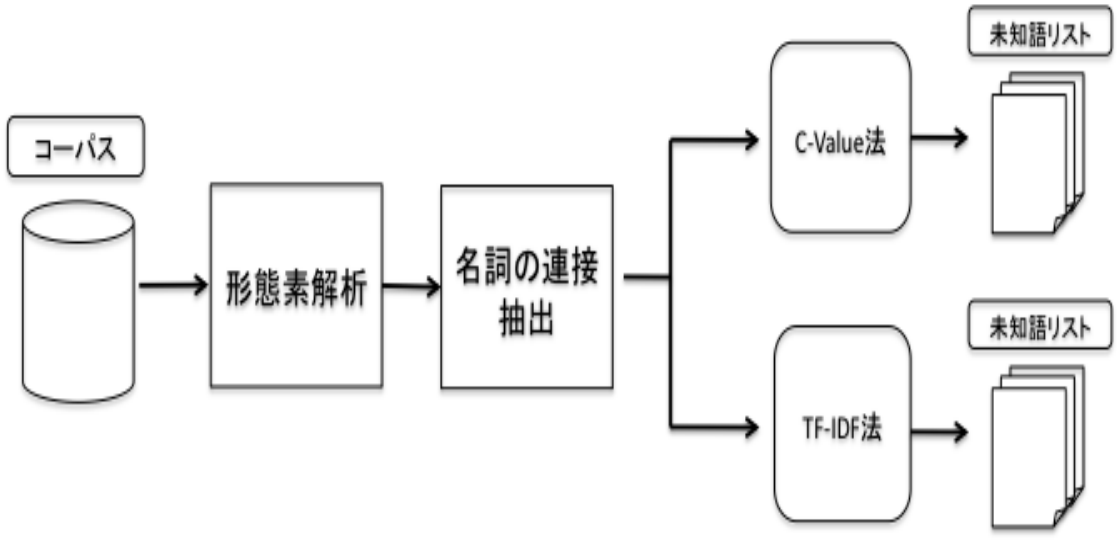

図 1: 複合語抽出の流れ

本研究では、先行研究 [4, 6] を参考にして、複合語の未知語リスト作成を図 1 の様に行う。まず、形態素解析器の利用して、コーパスの文章の形態素解析を行い、形態素解析の品詞情報から、名詞と名詞の連接を未知語候補とする。次に、抽出した未知語候補に対して TF-IDF、C-value でスコア付けを行い、最後にスコア付けした未知語候補から、未知語リストを作成する。

4.2.2 未知語候補の選定

コーパスから、複合語の未知語となりうる未知語候補を選定する。コーパスの文章を形態素解析することで得られる品詞情報をもとに未知語候補を選定する。本研究では、候補となる文字列は、名詞と名詞の連接を候補として選定する。具体的には、以下のような形で未知語候補を選定する。

例：急性の対象喪失反応の段階にあるドナーが家族に接する

形態素解析器の出力
急性 名詞, 一般,*,*,*,*, 急性, キュウセイ, キューセイ,,
の 助詞, 連体化,*,*,*,*, の, ノ, ノ,,
対象 名詞, 一般,*,*,*,*, 対象, タイショウ, タイショー,,
喪失 名詞, サ変接続,*,*,*,*, 喪失, ソウシツ, ソーシツ,,
反応 名詞, サ変接続,*,*,*,*, 反応, ハンノウ, ハンノー,,
の 助詞, 連体化,*,*,*,*, の, ノ, ノ,,
段階 名詞, 一般,*,*,*,*, 段階, ダンカイ, ダンカイ,,
に 助詞, 格助詞, 一般,*,*,*, に, ニ, ニ,,
ある 動詞, 自立,*,*, 五段・ラ行, 基本形, ある, アル, アル, ある/在る/有る,
ドナー 名詞, 一般,*,*,*,*, ドナー, ドナー, ドナー,,
が 助詞, 格助詞, 一般,*,*,*, が, ガ, ガ,,
家族 名詞, 一般,*,*,*,*, 家族, カゾク, カゾク,,
に 助詞, 格助詞, 一般,*,*,*, に, ニ, ニ,,
接する 動詞, 自立,*,*, サ変・－スル, 基本形, 接する, セッスル, セッスル, せっす

る/接する,

未知語候補
急性、対象、対象喪失、対象喪失反応、喪失、喪失反応、反応、段階、ドナー、家族

以上の例のように、名詞と名詞の連接を抜き出す。「対象__喪失__反応」のように名詞が連接した場合には「対象」「対象__喪失」「対象__喪失__反応」「喪失」「喪失__反応」「反応」の様にどの形が未知語であっても抽出できる形を未知語候補として選定する。

4.2.3 TF-IDF

TF-IDF とは、TF（Term Frequency:単語の出現頻度）と IDF(Inverse Document Frequency:逆文書頻度) を掛け合わせたものである。これは、ある文書の集合（コーパス）の中に含まれる 1 つの文書に注目したとき、その文書がどういった単語で特徴づけられるか調べる手法である。情報検索の分野でも主に使用されている。TF-IDF は、ある特定のドキュメントでのみ出現し、かつ、出現頻度の高い単語を抽出するための手法である。

以下の式で表される

$$TFIDF(w) = f(w) \log \frac{N}{df(w)} \tag{1}$$

式のそれぞれの値は、$f(w)$：単語 w を含む文書の数、N：文書総数、$df(w)$：単語 w を含む文章数、である。

4.2.4 C-value

C-Value とは、出現頻度、文字の長さ、Nested Term と呼ばれる語の包括関係を利用した、用語らしさのスコア付けを行う用語抽出の手法の 1 つである。単独で出現し、かつ、構成する形態素の数が多い単語を抽出するための手法。以下の式で C-value と、Nested Term について述べる。

$$C-Value(w) = \begin{cases} log|w| \cdot f(w) & (w \neq NestedTerm) \\ log|w|f(w) - \frac{1}{T_w}\sum_{b \in T_w} f(b) & (w = NestedTerm) \end{cases} \quad (2)$$

式のそれぞれの値は、w：対象とする単語、$f(w)$：ある文章中の単語の頻度、$|w|$：w を構成する形態素数、T_w：w を内部に含むより大きな単語の集合、$f(b)$：w を内部に含むある単語、である。

Nested Term とは、対象の単語が他の単語の構成要素の一部であるとき、対象の単語のことである。例えば、獲得した未知語候補一覧に “contact lens”、“hard contact lens”,”contact lens fluid ” があるとすると、“hard contact lens”、“contact lens fluid” という文字列の構成要素である”contact lens” が Nested Term となる。

この式より、頻度 $f(w)$ が高く、構成要素の数 $|w|$ が多い候補であっても、Nested Term であれば値が小さくなり、単独で出現している単語のスコアが高くなることがわかる。

4.3 単名詞抽出手法

4.3.1 単名詞抽出の流れ

本研究では、先行研究 [13, 14] を参考にして、単名詞抽出を図 2 のような流れで行う。まず、コーパス中から文字 N-gram を獲得する。次に獲得した候補をある条件で選定する。最後に候補を先行研究を参考にした 2 つの手法でスコア付けする。1 つは、未知語候補に隣接文字の異なり数をエントロピーを利用した手法、もう、1 つは外部辞書を利用して、未知語候補の前後に来る隣接文字情報を利用した手法である。

4.3.2 未知語候補の抽出

単名詞抽出の未知語候補抽出にあらゆる文字列を抽出するために文字 N-gram を利用している。

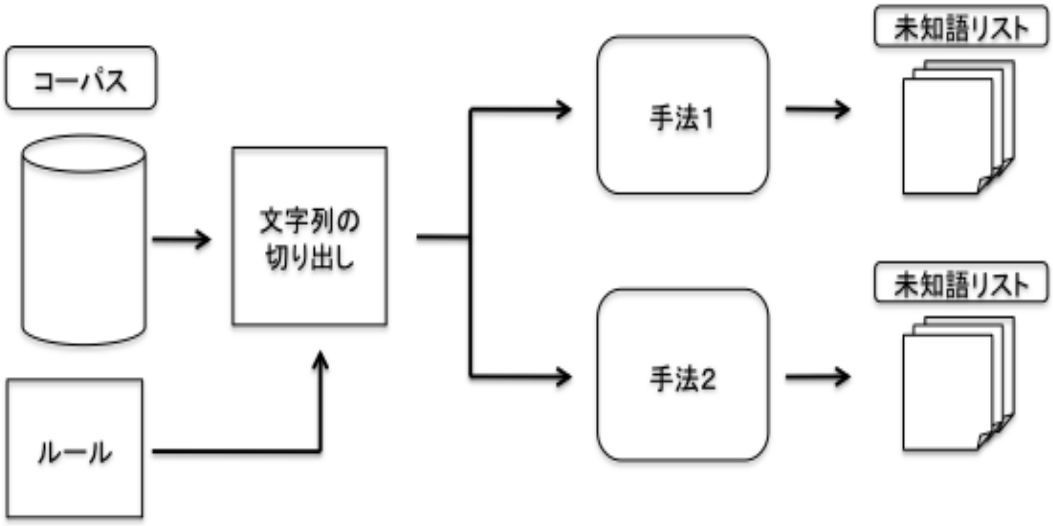

図 2: 単名詞抽出の流れ

N-gram とは、ある文字列から切り出した一定個数の文字の並びの集合である。1 文字続きのものは unigram、2 文字続きのものは bigram、3 文字続きのものは trigram、と特に呼ばれ、4 文字以上のものは、単に 4-gram、5-gram と表現されることが多い。

実験では、1 文字の単語で形態素解析器の辞書に登録されていない専門用語は少ないと考えられる事から、bigram から 10-gram の全てを未知語候補として利用する。N-gram の例として、表 1 に例文「痴呆症患者は高確率でせんもう状態を生じる」を N-gram に分けたときの例を上げる。

この方法では、単名詞のパターンを全てのを抽出でき、形態素解析器では抽出する事ができない可能性のある文字列も抽出できる。

4.3.3 未知語候補の選定

単語となりえない形式の文字列を除いた。本研究では

- 数字のみで構成されている。

表 1: N-gram 例

N	N-gram
2	痴呆、呆症、症患、患者、…、 生じ、じる
3	痴呆症、呆症患、症患者、患者は、…、 を生じ、生じる
4	痴呆症患、呆症患者、症患者は、患者は高、…、 態を生じ、を生じる
5	痴呆症患者、呆症患者は、症患者は高、患者は高確 、…、状態を生じ、態を生じる
6	痴呆症患者は、呆症患者は高、症患者は高確、 患者は高確率、 …、う状態を生じ、状態を生じる
7	痴呆症患者は高、呆症患者は高確、症患者は高確率、 患者は高確率に、…、もう状態を生じ、う状態を生じる
8	痴呆症患者は高確、呆症患者は高確率、症患者は高確率に、 患者は高確率にせ、…、んもう状態を生じ、もう状態を生じる
9	痴呆症患者は高確率、呆症患者は高確率に、 症患者は高確率にせ、 …、せんもう状態を生じ、んもう状態を生じる
10	痴呆症患者は高確率に、呆症患者は高確率にせ、 症患者は高確率にせん 、…、にせんもう状態を生じ、せんもう状態を生じる

- 片方の括弧のみ入っている。
- 文字列の始めに伸ばし棒、点、記号が配置されている。
- 文字列の構成が 4 回以上種類が変化しているもの (種類：漢字、カタカナ、英語、平仮名) ※例：5 回”専門用語の前の単語は” → 専門用語 — の — 前 — の — 単語 — は

以上を単語となり得ない文字列として候補から外す。

4.3.4 エントロピーを利用したスコア付け

図 3 に、エントロピーを利用したスコア付け手法を示す。①のように、ある単語の後ろに決まった文字が出現するときはエントロピーが低くなり、①は 1 つの単語として成り立っていないと考える。次に②のように、単語の後ろに決まった文字が出現しないときエントロピーが高くなり、②は 1 つの単語として成り立っ

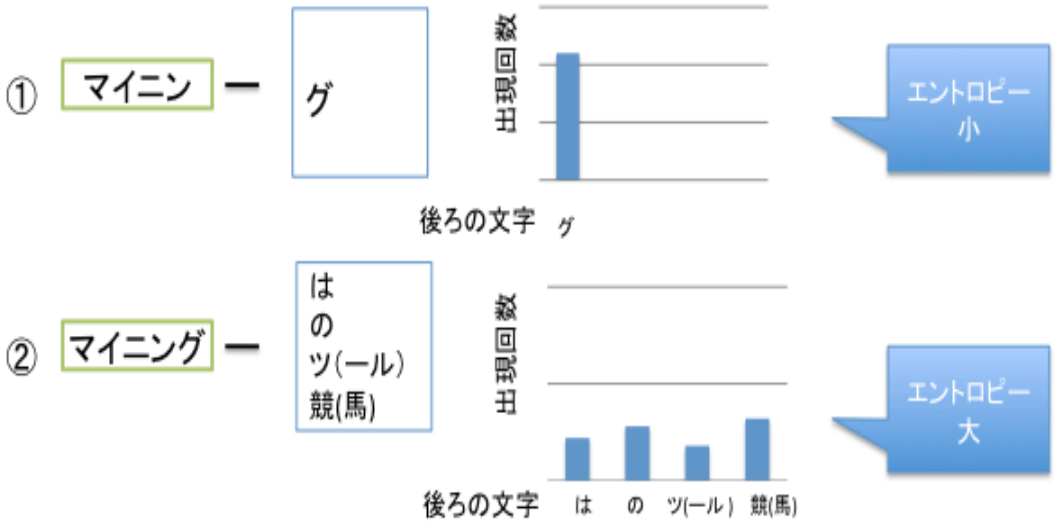

図 3: エントロピー手法の考え方

ていると考える。このように、抽出した未知語候補が 1 つの単語として成立していると考えた時、その未知語候補の前後のエントロピーが高くなるという考えからエントロピーを用いてスコア付けを行う。

実際にエントロピー手法として利用した式を以下に示す

$$Score(w) = f(w) \times \sum_{x} -P(X = x)logP(X = x) \times \sum_{y} -P(Y = y)log(Y = y) \tag{3}$$

式のそれぞれの値は、$f(w)$：候補となる未知語候補の頻度、x：未知語候補の前に出現するある文字、X：未知語候補の前に出現する文字列の集合、y：未知語候補の後ろに出現するある文字、Y：未知語候補の後ろに出現する文字列の集合である。また、この式は

未知語候補の出現頻度×未知語候補の前に出現する文字の異なり数のエントロピー × 未知語候補の後ろに出現する文字の異なり数のエントロピー

を表している。この式で算出される値が高い候補は、頻度が多く、また、前後に出現強いる文字の異なりの種類が大きいため、1 つの文字列である可能性が高い事を示している。

4.3.5 隣接文字の性質を利用したスコア付け

未知語の前後の文字列に出現する文字の分布を利用するものであり、考え方として、4.3.4 節と同様である。4.3.4 節のスコア付けとの違いは、前後に来る文字列を外部の品詞が付与された Web から重みを学習して、利用する点である。

考え方の例として

- **胸焼け**を 解消するにはガムをかむといい
- 虚栄心の強い人 は **てんかん**に かかりやすい
- 虫歯予防 に **キシリトール**の ガムは効果的？

など、普通名詞であれば、「の」「が」「を」「は」等が直前直後に現れやすい文字であり、また、普通名詞は文頭である事が多く、直後に「。」等の句点が現れる事が少ない、という性質がある。

本研究では、この特性を使用して、名詞の前後に出現しやすい文字の重みを Web コーパスを利用して獲得し、スコア付けに利用した。詳しく以下の式で示す。

$$Score(w) = f(w) \times \sum_{N} f(w, N = n) \frac{L(n)}{L} \times \sum_{M} f(w, M = m) \frac{L(m)}{L} \qquad (4)$$

式のそれぞれの値は、w：候補となる未知語候補の頻度、N：未知語候補の前に出現する文字の異なり数、$f(w, N = n)$：未知語候補 w の前に文字'n' が出現した回数、L：Web コーパス中に出現する名詞の出現回数、$L(n)$：n が Web コーパス中の名詞の全単語の前に出現する回数、M：未知語候補の後ろに出現する文字の異なり数、$f(w, M = m)$：未知語候補 w の後ろに文字'm' が出現した回数、 $L(m)$：m が Web コーパス中の名詞の全単語の前に出現する回数、となっている。

この式で算出される値は、頻度が多く、前後に出現している文字が名詞としての性質をもつ可能性が大きい文字を多く備えている候補が高いスコアを示す。

5　実験

5.1　実験設定

本研究で使用したコーパスや形態素解析器等の説明を行う。

本研究では、コーパスは科学研究省の生命・医療分野のコーパスを使用した。コーパスは 4634 文書、約 3700000 文で構成されている。また、使用したコーパスは、文字区切りや、品詞付与がいっさいないため、実験の正解となるデータの作成を行う。以下、評価尺度で詳しく述べる。形態素解析器は『MeCab』を使用した。実験評価、正解データ作成、生命科学の学問領域で使われる専門用語などのオンライン辞書のライフサイエンス辞書 (25510 単語) を使用した。また、3.6 節で名詞の前後に現れる文字の傾向を計算するために、別のコーパスとしてを TSUBAKI という Web コーパスを利用した。

5.2　評価尺度

本研究では、複合語抽出と単名詞抽出では未知語候補の作成方法の違いがあり、同じ条件で比較するため複合語抽出と単名詞抽出を分けて評価する。

5.2.1　評価尺度：複合語抽出

複合語抽出では Precision(精度) を使用して評価する。本研究では以下の式で精度を求める。

$$Precision\ (\text{精度}) = \frac{\text{専門用語として正しい数}}{\text{抽出した未知語のスコア上位 100 単語}} \tag{5}$$

それぞれの手法についてスコア上位 100 単語を対象として精度を求める。未知語のスコア上位 100 単語で計算するため、形態素解析器内の辞書に登録されている単語は除いて計算する。また、精度の評価は人手で行い、かつ、一人の評価である。

5.2.2 評価尺度：単名詞抽出

単名詞抽出では、Precision(精度) と Recall(再現率) を使用して評価する。精度の式は、複合語抽出で使用した式 (5) である。再現率は以下の式で求める。

$$Recall\ (\text{再現率}) = \frac{\text{抽出できた正解単語の数}}{\text{作成した正解単語の総数}} \tag{6}$$

再現率について、形態素解析器内の辞書を一部削除し、未知語とする事で再現率を計算するための候補を獲得する。

未知語として使用した単語は、形態素解析器内の辞書から、コーパスに含まれており、かつ、ライフサイエンス辞書に掲載されている単語である。全 2502 単語を正解データとして使用する。

また、精度、再現率を計算する際に既存の形態素解析器内の辞書に含まれる単語を、形態素解析器内の辞書に掲載されているという理由から除く。

複合語の再現率は使用したコーパスに正解となる単語の情報が付与されておらず、コーパスから正解単語を作成する事が困難であり、また、単名詞抽出の正解単語を使用しないのは、単名詞抽出の際に正解単語とした単語は本来、未知語ではない。単名詞で再現率を評価尺度に加えた理由は、単名詞は正解単語としてのは、単名詞の未知語は複合語の未知語と比べて少なく、精度のみでは評価できないと考慮したためである。

5.3 実験結果

以下の各手法によりスコア付けをして、各手法の精度と再現率率を求める。

- 3.2.2 節、TF-IDF
- 3.2.3 節、C-value
- 3.3.5 節、既存の形態素解析器を利用せず、前後情報とエントロピーを利用してしたスコア付け手法（以下、「エントロピー法」）

- 3.3.6 節、既存の形態素解析器を利用せず、前後情報と未知語の性質を利用したスコア付け手法 (以下、「隣接文字法」)

- エントロピー法と隣接文字法のスコアの積をスコアとする手法 (以下、「積手法」)

$$\text{積手法} = \frac{\text{エントロピー法のスコア} \times \text{隣接文字法のスコア}}{\text{出現頻度}} \tag{7}$$

比較する手法の中で積手法を加えたのは、未知語を獲得する上で複数の手法を組み合わせ、未知語を判断する材料を増やす事で精度の向上を測るためである。積手法は式 (7) で表され、エントロピー法と隣接文字法のスコアをかけた数値から、それぞれの手法で使用されている出現頻度の重複を防ぐため、出現頻度で割るという手法である。

5.3.1 複合語抽出の精度

表 2: 複合語抽出の上位 100 単語の精度 (%)

手法	完全一致	部分一致
TF-IDF 法	80	93
C-value 法	72	100

表 2 は複合語抽出の精度を示した表である。表では完全一致と部分一致に分けて精度の結果を出力した。完全一致とは、抽出した未知語が専門用語であるときを正解としたときの精度であり、部分一致は完全一致に加えて、抽出した未知語が専門用語だと判断した単語を文字列の中に含む場合も正解として扱ったときの精度である。部分一致の例として、「再生__医療」を正解と判断したときの「再生__医療__プロジェクト」などである。表 2 では TF-IDF が最も優れた結果を残しており、C-value 法は完全一致こそ精度が低いが、部分一致では TF-IDF 法の精度より優れた結果を示した。また、単名詞抽出では積手法がエントロピー手法、前後文字手法より、良い結果が出ている事がわかる。

5.3.2　複合語抽出の精度分析

表 2 のより、各手法の精度についての分析を述べる。

表 3: スコアの上位の 10 単語

TF-IDF						C-value					
1	HIV		6	ダイオキシン類	▲	1	分担研究		6	研究事業	▲
2	高齢者		7	HIV 感染		2	研究報告	▲	7	医療機関	
3	症		8	化学物質		3	厚生科学研究		8	厚生科学	▲
4	感染者		9	DNA		4	科学研究		9	研究目的	
5	障害者		10	糖尿病		5	分担研究報告	▲	10	研究方法	

▲は専門用語と判断した単語と包括関係にある単語、記入がない単語は専門用語として正しいと判断した単語

表 3 より、複合語手法が精度の結果について考察をする。表 3 は複合語手法の上位 10 単語を示した表である。

複合語抽出のスコア付けの結果は、用語性を重視した TF-IDF と単位性を重視した C-value で大きく異なる。単位性を重視した C-value では、完全一致のと部分一致の差が TF-IDF よりも大きいことがわかる。

例えば、C-value では「研究〜」「〜研究」という形は、表によく現れるが、「研究」自体はスコアは低い。そのため、「研究」を専門用語として判断した今回の精度において、C-value では部分一致のスコアが高く出力された。これは単位性を重視した C-value の特徴である、Nested　Term のスコアが完全一致の候補を押さえたと考えられる。しかし、部分一致の高い C-value で獲得できた複合語は、表 5.3.2 より「研究報告」「研究目的」「研究方法」など一般的であり、専門用語とはわかりにくい単語が出現した。

一方、TF-IDF では、C-value の特徴である Nested Term ではスコアが低くなってしまう単名詞が上位に食い込んでいる。また、C-value と比べて誤りと判断できる単語がスコア上位に出現した。TF-IDF のスコア上位 100 単語で誤りとした単語は以下の単語である。

- the、and、Fig、0 歳、1 例、 5 月、et、 2 例、 4 月、図 1

いずれもドキュメント全体で出現頻度が高い単語であるが、特定の文章にのみ出現するという単語ではない。「the」「and」「Fig」などはどれも半分以上のドキュメントに出現しているにも関わらずスコア上位に位置している。これは、特定の文章にのみ出現する単語を抽出するためのフィルターである TF-IDF の IDF(逆文書頻度) が上手く機能していないと分析できる。

複合語抽出の分析より、各手法での欠点分析できた。TF-IDF では、ドキュメント全体で出現頻度が高い単語であってもスコア上位に現れてしまい、一般的な単語のフィルターである IDF が上手く機能していないという点、C-value で抽出できる単語は一般的な単語の複合語がスコア上位に現れて、専門用語とはわかりにく単語がスコア上位に来てしまっているという点である。

5.3.3 単名詞抽出の精度

表 4: 単名詞抽出のスコア上位 100 単語の精度 (%)

手法	完全一致	部分一致
エントロピー法	28	28
隣接文字法	44	46
積手法	55	57

表 4 は単名詞抽出の精度を示した。完全一致、部分一致については複合語抽出の精度の場合と同様の設定である。表 4 より、隣接文字法がエントロピー法よりも精度が精度が高い事がわかる。また、抽出した上位 100 単語ではほとんどが完全一致しており、部分一致しているものが少ない。そして、積手法がエントロピー法と隣接文字法より、高い精度を出力した。

5.3.4 単名詞抽出の精度分析

単名詞抽出のスコア付けの結果は、表5より、積手法がエントロピー法、隣接文字法の結果を上回った。これは、２つの手法を掛け合わせる事で精度が向上する事が判明した。また、エントロピー法が隣接文字法と比べて、精度が低くなってしまった原因は、どちらも同じ情報を使用してのスコア付けであるのでエントロピーが単語の境界として、上手に機能しなかったからと考えられる。

表 5: スコアの上位の 10 単語

順位	エントロピー法		隣接文字法		積手法	
1	して	×	して	×	研究	
2	研究	×	研究		して	×
3	である	×	調査		患者	
4	には	×	患者		調査	
5	細胞		よう		細胞	
6	調査		ある	×	である	×
7	への	×	細胞		よう	
8	している	×	いて	×	治療	
9	的な	×	である	×	ること	×
10	であ	×	では	×	遺伝子	

×は専門用語ではないと判断した単語、記入がない単語は専門用語して正しいと判断した単語

表5は各手法上位10単語を示したものである。表5をもとに単名詞抽出の精度を分析する。表４より、積手法、エントロピー法、隣接文字法では名詞以外の文字列も出現している。積手法が他の手法よりも高い精度が出たのは他の２つの手法よりも名詞が多く出現しているからである。また、スコア上位100単語まで見た時、エントロピー法と隣接文字法との違いはエントロピー法では「への」「的な」などが、隣接文字法では出現していなかった。また、この分析ではエントロピー法と隣接文字法の大きな違いが出なかった。違いを分析するために、両手法

の違いであるエントロピーと隣接文字の重みのみで未知語候補のスコア付けで実験する。

表 6: 単名詞抽出手法の出現頻度を除いたスコア上位 20 単語

手法	スコア上位単語（出現頻度抜き）
エントロピー法	などの、等の、を含めた、からの、などによる を中心とした、以外の、等による、などを、 のみが、のみの、時の、としての、を中心とする とその、に関連した、への、等を、型の 。そのものが
隣接文字法	患者、研究、調査、医療、細胞、可能性、方法 遺伝子、高齢者、女性、以上、本研究、 女性、結果、して、ために、生活、疾患、検討、

表 6 は、単名詞抽出の各手法の式 (3)(4) から、出現頻度を除いてスコア付けした際の、スコア上位 20 単語である。上位は「などの」「を含めた」「以外の」などの前後に名詞が出現するようなパターンが多い。名詞のパターンは無限にあると考えられるので、「などの」「を含めた」「以外の」のような文字列のスコアが高くなってしまったと考えられる。一方で獲得したい単名詞の前後に来るでと考えていたパターンとして「の」「が」等の助詞もしくは、単名詞と関係のある名詞が来るのみで、「などの」「を含めた」「以外の」と比べると、エントロピーが低くなってしまうと分析できる。また、隣接文字法の場合では、隣接する文字のスコアがライフサイエンス辞書の専門用語と同じ文字が出現した場合にスコアが高くなるため、「の」「が」「を」などが隣接する単語のスコアが高くなり、対応する単語が前後に出現しない「などの」「を含めた」「以外の」のスコアは低くなり、エントロピー法よりも名詞が抽出でき精度が高くなったと考えられる。

単名詞抽出の精度分析より、大きく３つのことが分析できた。１つ目は積手法では、エントロピー法と隣接文字法を掛け合わせる事で両手法よりも名詞のスコアを上げる事ができること、２つ目と３つ目は、未知語候補のスコア付けの際に出現頻度を除く事で、エントロピー法と隣接文字法の違いがはっきりし、エント

ロピー法は名詞の区切りとしては適切ではないということ、隣接文字法は、名詞の区切りとして正しく機能し、出現頻度を除いた場合のスコア上位の方が名詞が多いということである。

5.3.5 再現率

表 7: 単名詞抽出のスコア上位単語の再現率 (%)

手法	5000 単語	10000 単語	50000 単語	100000 単語
エントロピー法	15.4	22.4	42.8	52.7
隣接文字法	19.2	27.0	46.5	55.4
積手法	22.0	29.3	49.6	58.2

表 7 では、手法ごとのスコアが上位 5,000-100,000 単語までの再現率を示している。単名詞抽出では、エントロピー手法、隣接文字手法、積手法のスコアは上位 5000 単語から、300-550 の正解データが含まれていることがわかる。さらに、スコアが上位 100,000 単語まで、見ると 50%以上の再現率が獲得できる。

5.3.6 再現率分析

表 7 より、得られた結果を使用して再現率の分析を行う。

まず、全体の再現率より、積手法が、エントロピー法、隣接文字法よりも再現率が高く、精度と同様にエントロピー法と隣接文字法を掛け合わせる事で再現率も向上させる事ができると判明した。

次に、エントロピー法と隣接文字法では隣接文字の方が再現率が高い事がわかる。これは、精度の分析で述べた、エントロピー法では「などの」のような前後に名詞が来る文字列のスコアを高くしてしまうため、目的である単名詞が隣接文字の手法と比べて抽出しにくいと分析できる。しかし、名詞が前後に出現しやすい文字列以外では、名詞の場合のエントロピーが比較的ほかの文字列より高くな

る。そのため、5,000 単語では 4%近く再現率の差が開いていたが、10,000 単語では 3%を切るまでになっている。

次に、どのような単語が各手法で抽出できたのか分析する。

表 8: 正解データの出現頻度ごとの抽出単語数 (個))

手法	1000 以上 (856 単語)	100 以上 (785 単語)	10 以上 (563 単語)	10 未満 (298 単語)
エントロピー法	848	471	0	0
隣接文字法	854	533	0	0
積手法	853	602	0	0

表 8 は各手法のスコア上位 100,000 単語を対象としたとき、再現率に使用した正解単語のコーパス中出現回数を桁ごとに分けたものである。それぞれの手法の上位 10 万単語で調べた。この表より、積手法では出現回数が多いものほど獲得しやすい事がわかる。特に、正解単語が 1000 回を超えている単語についてはほぼ全て抽出できているという結果が出た。また、今回の手法において、出現回数が 100 回未満の単語を獲得できないがわかる。この結果より、出現頻度が低い正解単語はスコア上位に現れないことがわかった。

5.4 ライフサイエンス辞書にのみ掲載されている単語の再現率

本研究の実験設定において、再現率を計算する際にライフサイエンス辞書に登録されていて、かつ、形態素解析器の辞書にも登録されている単語を正解データとして再現率を計算した。この実験設定は形態素解析器の辞書に掲載されていない場合、誤りが発生するとかんがえられる実験設定のするために設定したものである。

しかし、実際には形態素解析器の辞書に登録されている単語である。そこで、形態素解析器内の辞書に登録されていないライフサイエンス辞書の専門用語を未知語として扱った場合、どのような単語が抽出する事ができるの分析する。

ライフサイエンス辞書に掲載されていて、コーパス中に存在するし、そして形態素解析器の辞書に登録されていない単語は全 22669 単語である。この単語に対して、複合語抽出、単名詞抽出の再現率を示す。

使用する手法は本実験の時に用いた、TF-IDF、C-value、エントロピー法、隣接文字法、積手法の 5 つの手法で比較する。また、再現率は以下に示す。

5.4.1 実験尺度

本節では、実験尺度として複合語抽出、単名詞抽出両方とも Recall(再現率) を使用して評価する。再現率は以下の式で表す。

$$Recall\ (\text{再現率}) = \frac{\text{抽出できた単語数}}{\text{専門用語辞書にのみ掲載されている単語数}} \tag{8}$$

5.4.2 実験結果：複合語抽出

表 9: 複合語抽出のライフサイエンス辞書にのみ出現している専門用語の再現率 (%)

手法	再現率 (スコア上位 100000 単語)
TF-IDF	13.1
C-value	5.4

表 9 は、複合語抽出手法のスコア上位 100,000 単語の再現率である。この表より、複合語抽出では、C-value より TF-IDF の方が再現率が高い事がわかる。また、表 7 と比べて、再現率が著しく低い事がわかる。

5.4.3 実験分析：複合語抽出

表 10: 複合語抽出の正解データの出現頻度ごとの抽出単語数 (個)

手法	1000 以上 (650 単語)	100 以上 (2805 単語)	10 以上 (7203 単語)	10 未満 (12011 単語)
TF-IDF	531	1987	451	1
C-value	100	478	652	1

表 10 は、複合語抽出手法のスコア上位 100,000 単語を対象とした時、再現率に使用した正解単語のコーパス中の出現回数を桁ごとに分けたものである。

表より、まず、ライフサイエンス辞書に登録されている単語の出現回数はコーパスにおいて全体的に少なく、85%以上が 100 回も出現していない単語であるがわかり、出現回数の低い単語はほぼ抽出できない事わかる。

また、複合語抽出の２つの手法を比べた際の抽出できる単語が出現頻度で大きく分かれている事がわかる。TF-IDF では、単純に出現頻度が高いものほどスコア上位に来やすいという結果が出ている。対して、C-value では、TF-IDF と比べて出現頻度が 10 未満の単語を除き満遍なく抽出できるという結果が出た。特に 10 単語以上 100 単語未満の単語では、TF-IDF よりも抽出できる単語が多かった。

今回の実験より、複合語抽出では TF-IDF と C-valeu で獲得できる単語に大きく差が出る事がわかった。

なぜ、このような結果になったのか分析する。

まず、C-value では 1000 回以上出現する高出現頻度の単語が抽出できないのか分析するこの問題については、抽出できていない高い頻出単語が様々な問題で抽出できていないことがわかった。まず、抽出できなかった多くの正解単語が単独で出現する事が少ないという C-value のスコアを算出する際に問題となる点、次に、“新生物” が” 最新生物” や” 生新生物” など形態素解析器で単語を分かち書きにしたときに正解単語と違う切れ方をしてしまう点、最後に” 高コレステロール血症” の” 高” が接続詞であるなどの名詞の連接で出力されていない場合である。

次に.TF-IDF で抽出でき無かった 1000 回以上出現する高出現頻度単語を分析する。C-vaelu のときと同様に形態素解析器で分析したときに起きる問題が多いが、その他の問題として、全ドキュメントに満遍なく出現するが特定のドキュメントで数が多いという単語”IL” などである。

研究分析より、専門用語を抽出する際の手法では、多くの研究で出現頻度をスコアの計算の際に利用しているが、実際にコーパス中で出現している専門用語は出現頻度が少ないことが分析できた。次に単名詞抽出では、形態素解析器の精度や未知語候補を抽出する際の形態素解析に問題があったり、今回の未知語候補とする名詞の連接以外のパターンも考慮しないと抽出できない単語があり、未知語

候補を作成する際のルールを強化しないと抽出できないことが分析できた。

5.4.4 実験結果：単名詞抽出

表 11: 単名詞抽出のライフサイエンス辞書にのみ出現している専門用語の再現率 (%)

手法	再現率 (スコア上位 100000 単語)
エントロピー法	4.3
隣接文字法	5.8
積手法	6.6

表 11 は、単名詞抽出手法のスコア上位 100,000 単語の再現率である。この表より積手法がエントロピー法、隣接文字法の再現率を上回り、組み合わせる事でスコアがしたことがわかる。また、複合語抽出と同様に表 7 と比べて、再現率が低い事がわかる。

5.4.5 実験分析：単名詞抽出

表 12: 単名詞抽出の正解データの出現頻度ごとの抽出単語数 (個))

手法	1000 以上 (650 単語)	100 以上 (2805 単語)	10 以上 (7203 単語)	10 未満 (12011 単語)
エントロピー法	321	658	0	0
隣接文字法	337	981	1	0
積手法	336	1152	5	0

表 12 は、単名詞手法のスコア上位 100,000 単語を対象とした時、再現率に使用した正解単語のコーパス中の出現回数を桁ごとに分けたものである。

表より、３つの手法で大きく差が出る事なく、積手法がエントロピー法、隣接文字法よりも良い結果を出力できたという結果が出た。ライフサイエンス辞書と

形態素解析器の辞書に掲載さていたときの表 8 と比べると、高出現頻度の再現率も落ちている。特に、形態素解析器の辞書に載っているときでは、1000 回以上出現している単語はほぼ 99%取れていたのにもかかわらず、今回の実験では、約半分のしか抽出できなかった。

最後になぜ、表 12 のような結果が出たのか分析する。

単名詞抽出ではライフサイエンス辞書のみに載っている単語の場合、再現率が落ちてしまったのか分析する。このような結果が出たのは、ライフサイエンス辞書にのみ出現する専門用語には英語が含まれて、約 300 単語が英語である。単名詞抽出の全手法で英語が抽出できなかった。エントロピー方では、英語の候補の前後の文字列が英語か空白、句読点などの記号のみとなり、日本語と比べてエントロピーが低くなってしまい。スコア上位に現れず、スコア上位が日本語の単語のみとなってしまっている。一方で隣接文字法では隣接文字の重み付けの際に利用したコーパスには英語が少ないため、英語の前後には「の」「が」「は」などは出現せず、エントロピー法と同様に日本語がスコア上位を占めている事がわかった。しかし、英語以外の単語についてはだいたい抽出できていることが分析でわかった。

研究分析より、単名詞抽出では、今回の手法では言語統一をしない限り、どれだけ出現していても英語は抽出 することができないことが分析できた。

6　未知語抽出手法の改善法の提案と実験

今回の実験結果より、複合語抽出、特に C-value で抽出した未知語が一般的に使用されているような単語ばかりが上位に出現してしまった。

原因として考えられるのが、TF-IDF、C-value のスコアを計算する際にドキュメント全体で計算しているため、ドキュメント全体で出現する単語のスコアが高くなってしまう。また、C-value では TF-IDF の IDF のような全体的に出現する単語のスコアを低くするフィルターが無いためこのような結果になっていしまったと考えられる。

そこで TF-IDF、C-value を本節では、改良した方がよいのではと考え、手法を実装して、実験を行った。

6.1　実験手法

本節で TF-IDF、C-vaelu の専門性を向上させるための手法を以下に示す。

1. ドキュメント毎に計算する (手法 1)

$$TF-IDF(w) = \arg\max_{d} f(w,d) log \frac{N}{df(w)} \tag{9}$$

$$C\text{−}Value(w) = \begin{cases} \arg\max_{d} log|w| \cdot f(w,d) & (w \neq NestedTerm) \\ \arg\max_{d} log|w| \{ f(w,d) - \frac{1}{T_{w,d}} \sum_{(b,d)\in T_{w,d}} f(b,d) \} & (w = NestedTerm) \end{cases} \tag{10}$$

2. C-value のスコアに IDF をかける (手法２)

$$C\text{−}Value(w) = \begin{cases} log|w| \cdot f(w) \cdot log \frac{N}{df(w)} & (w \neq NestedTerm) \\ log|w| \{ f(w) - \frac{1}{T_w} \sum_{b\in T_w} f(b) \} \cdot log \frac{N}{df(w)} & (w = NestedTerm) \end{cases} \tag{11}$$

表 13: TF-IDF のスコア上位 100 単語の精度 (%)

手法	完全一致	部分一致
TF-IDF	80	93
手法 1	85	98

6.2 実験結果:TF-IDF

表 13 は TF-IDF の精度である。手法 1 の方が完全一致、部分一致の精度が高い事がわかった。結果の分析とスコア上位単語が専門性のある単語になっているのか分析する。

6.3 実験分析:TF-IDF

表 14: TF-IDF のスコア上位 10 単語

TF-IDF						手法 1					
1	HIV		6	ダイオキシン類	▲	1	xs		6	TTV	
2	高齢者		7	HIV 感染		2	EET		7	Fc ε RI 発現	
3	症		8	化学物質		3	gp100		8	CalDAG	
4	感染者		9	DNA		4	日本手話		9	10Na	▲
5	障害者		10	糖尿病		5	ジンセノサイド		10	安全対策担当者	▲

▲は専門用語と判断した単語と包括関係にある単語、記入がない単語は専門用語として正しいと判断した単語

表 15 に TF-IDF のスコア上位 10 単語を示した。まず、表よりドキュメント毎に TF-IDF を行った手法 1 では、「EET（試薬の名前)」「ジンセノサイド（栄養分)」「gp100(抗原)」など、一般的には使用されないような専門単語が上位に出現することがわかる。しかし、手法 1 ではでは生命・医療分野とは直接関係のな

い単語を確認できる。例えば、スコア上位 100 単語の「xs（プログラム)」「クロス集計表（表の種類)」などである。これらの単語は特定の文章でのみ出現しておらず、出現回数も多い。しかし、これらの単語は TF-IDF での抽出目的である用語性とは大きくはなれており、出現頻度の多い単語となってしまっている。

次に、手法 1 でスコア上位 100 単語のスコアを確認すると、改善前の TF-IDF と比べて出現するドキュメント数が大きく減っていた。改善前の TF-IDF では、スコア上位 100 単語でも全 4643 ドキュメント中 3000 ドキュメント以上出現する単語もあったが、手法 1 では、20 ドキュメントを超える単語が存在しなかった。これは、出現頻度をドキュメント毎にすることで、スコアに対する出現頻度の割合を減少させる事ができたと考えられる。

6.4 実験結果：C-value

表 15: C-value のスコア上位 100 単語の精度 (%)

手法	完全一致	部分一致
C-value	72	100
手法 1	86	100
手法 2	73	100

表 15 は C-value の精度である。手法 1 では完全一致が上がり、手法 2 もわずかだが精度が向上した。TF-IDF と同様に、スコア上位単語を分析する事で精度結果について分析する。

6.5 実験分析：C-value

次に C-value について分析する。表 16 にスコア上位 10 単語を示した。手法 1 、手法 2 ともとの C-value で多く変化した。表より、手法 1 では、手法 2 や C-value とは違い「研究」「報告」などの単語が入っている単語が少なく、「身体障害」「口

表 16: C-value のスコア上位 10 単語

順位	C-value		手法 1		手法 2	
1	分担研究		身体障害		化学物質	
2	研究報告	▲	医療機関		研究分担研究者	▲
3	厚生科学研究		口腔乾燥		医療機関	
4	科学研究		拠点病院		教授研究要旨	▲
5	分担研究報告	▲	更生相談	▲	母子	
6	研究事業	▲	添付文章	▲	精神保健福祉	▲
7	医療機関		静脈注射		特定疾患	
8	厚生科学	▲	研究開発		母子保健	
9	研究目的		診療記録開示	▲	特定疾患対策研究事業	▲
10	研究方法		HIV 感染		精神保健	

▲は専門用語と判断した単語と包括関係にある単語、記入がない単語は専門用語として正しいと判断した単語

腔乾燥」などの一般的には使用されていない単語がスコア上位に来ている。また、手法 2 では、C-value とは大きく違う結果となったが、手法 1 と比べると「研究所」「事業」「福祉」といった単語が後ろにつく事が多く、完全一致が低い結果となった。また、手法 2 でスコア上位単語になった単語では、TF-IDF と同様に特定のドキュメントにのみ出現しているというわけではなく、スコア上位単語の中には 1000 ドキュメント以上出現している単語も現れている。

複合語抽出の専門性を向上させるために行った実験より、コーパスが大きく、出現頻度が IDF に対して大きい時、IDF は一般的な単語のフィルターとはなりにくいということが考察できた。また、このように IDF が上手く機能しない場合では、ドキュメントごとにスコア計算することで精度の向上させることができると考察できる。

7　考察

7.1　形態素解析器の辞書への追加に関する考察

今回使用した手法で抽出できた未知語についての考察を行う。

1. 複合語
2. 単名詞
3. 部分一致
4. 略語
5. 人名、地名
6. 非ドメインの専門用語

7.1.1　複合語

TF-IDF、C-value の複合語獲得で主に抽出できる未知語で「アパタイト—セメント」「共存—症」など、単名詞抽出でも獲得できる。形態素解析器内の辞書に入れる事でドメインでの精度があがると考えられるが、1 つにまとめる必要の無い複合語も存在する「知的障害者更生相談所」「調査資料」などである。複合語抽出において、複合語の未知語として抽出できるものは、単名詞抽出でも抽出することができる。かつ、形態素解析器の出力では、名詞と名詞の連接であるので、形態素解析器の精度向上において、あまり影響が無いと思われる。

しかし、形態素解析は自然言語処理の基礎技術である。始めに紹介したように、固有表現抽出、構文解析、情報検索など様々な技術の基礎技術であり、ぞれぞれの研究で専門用語が専門用語と判明しているだけで次ぎに続く技術に良い影響を与える。例えば、固有表現抽出では、あらかじめ、形態素解析した単語が専門用語としてのかたまりで分析されていれば、精度向上につながる。このように複合語の未知語は形態素解析器の精度向上にはあまり影響がないが、形態素解析器内の辞書に追加するだけの価値はあると考えられる。

また、C-value で抽出できた単語が専門用語とはわからない一般的な単語が取れてきた事を考えると複合語で更生されている専門用語は一般的な名詞の連続でできているわけではなく、何らかの専門用語を含んだ単語の連接でできているのではないかと分析できる。

7.1.2 単名詞

複合語に比べて圧倒的に数が少ない。また、ほとんどが辞書に形態素解析器の辞書に掲載されており、新規で発見できた単語として「産生」「機序」「アポトーシス」などがある。しかし、「産生」「機序」は形態素解析を行うと誤りを発生する。これは、形態素解析あやまりで動詞として出力され、本研究のように名詞の連接を未知語候補とした場合 TF-IDF、C-value では抽出する事が難しく、形態素解析器を利用しない手法は単名詞を獲得する上で必要となってくることがわかる。

7.1.3 部分一致

完全に正しい未知語とは考えにくいが、部分一致してるものは多い。特に C-value 法では多く、「〜群」「〜研究」などの形式を取る事が多い。形態素解析器の辞書にはそのまま追加することは精度向上にはつながらないと考えられる。

7.1.4 略語

TF-IDF で多数確認された、形態素解析器の辞書に追加するのは、略語では無く正式名称の方がいいが、実際に形態素解析を行う際の文章は略語で書かれている事が多い。しかし、今回獲得して略語は確認できる限り英語表記であり、辞書に登録しなくても獲得できる。

7.1.5 人名、地名

TF-IDF、C-valeu で獲得できる。人名は形態素解析で区切りを間違う事が多い。特に「登—四郎」「崇—文」のような、よく使用されている名前が含まれる名前で

間違いが発生する。形態素解析器の誤りを減らすために、追加したい単語だが、人名なので数がとても多い。

7.1.6　非ドメインの専用用語

本研究では、生命・医療分野であるが、論文の形式をしているドキュメント集合がコーパスであるので、研究方法を示す上で使用したプログラム言語や研究結果を示すために表の名前等も獲得できた。ドメインとは関係ないが、文章を解析する上で必要であると考えられる。

7.2　獲得した未知語から見た全体の考察

形態素解析器の精度を向上する上で必要となってくるものが多い、これは、形態素解析の対象となる文章がある分野のドメインに関係ある文章であっても、ドメインと関係のない未知語が存在するためである。

形態素解析器の精度を向上を目指して、辞書に未知語を追加するという手法での解決を目指したが、今回の研究結果より、文章中の未知語はとても多く、ドメインとは関係のない分野の未知語も必要となっており、形態素解析器内の辞書に未知語を追加する以外の手法が必要となると考えられる。

8　おわりに

本研究では、新規ドメインにおける形態素解析器の精度向上のために、形態素解析器内の辞書に未知語を追加するという手法で精度向上を目指した。形態素解析器の辞書に追加すべき未知語を抽出するための手法が大量にあり、まったく同じ条件で比較を行っている研究が無いため、どの手法が適しているのかわからなかった。そのため、本研究では未知語抽出手法を検討するために、未知語抽出手法の先行研究から４つの手法の比較・分析をした。

抽出すべき未知語には複合語と単名詞の二種類あり、複合語抽出では、TF-IDF、C-value という２つの手法を使用した。また、単名詞抽出では、抽出したい未知語に隣接する文字を利用する手法が多く、その中から、エントロピーを利用した手法と隣接文字の性質を利用した２つの手法で比較実験を行った。

科学研究省の生命・医療分野のコーパスを利用して、比較実験を行い、それぞれの手法の実験結果より、複合語の精度では TF-IDF が C-value より優れた結果を出力した。単名詞抽出の精度では、隣接文字法がエントロピー法より優れた結果を出力した。次に、再現率では、複合語抽出手法を除いて、単名詞抽出手法の比較を行った。結果として、エントロピー法より、隣接文字法の方が高い再現率を出力した。また、精度と再現率の両方において、エントロピー法と隣接文字法のスコアの積をスコアとする積手法がエントロピー法と隣接文字法の結果を上回った。

各手法の分析では、それぞれの手法の長所・短所を発見する事ができた。また、各手法で専門性のある単語が抽出できるているのか確かめるために、外部の専門用語辞書を利用して、実験を行い、複合語抽出の分析できた専門性のある単語を抽出するために、実験で使用した手法を改善し、精度向上と共に専門性のある単語抽出を行った。。

各手法で抽出してきた未知語について分析・考察を行い、形態素解析器内の辞書に追加すべき未知語についての考察を行った。

最後に、本研究では、新規ドメインに対しての先行研究の未知語抽出を行い、評価・分析を行ったが、本来の目的は、形態素解析器内の辞書を拡充する事で新規ドメインに対する精度を向上させることである。今後の課題として、実際に形

態素解析器内の辞書に未知語を追加するために、獲得しスコア付けを行ったのち、品詞推定を行い追加すべき未知語推定を行っていきたい。

謝辞

本研究を進めるにあたり、ご指導を頂いた乾健太郎教授、岡崎直観准教授に感謝致します。また、日常の議論を通じて多くの知識や示唆を頂いた乾・岡崎研究室の皆様に感謝します。

参考文献

[1] 松本 裕治、形態素解析システム「茶筌」情報処理、2000

[2] 工藤拓、山本薫、松本祐治、CRF を用いた日本語形態素解析情報処理学会研究報告自然言語処理、 2004

[3] 森信介、中田陽介、NEUBIG Graham、河原達也、点予測による形態素解析 NL198 2010

[4] 小山照夫、日本語テキストからの複合語用語抽出情報知識学会誌、2009

[5] 辻 真太朗、西本 尚樹、小笠原 克彦、形態素解析における放射線技術学分野の用語適用 - 診療放射線技師試験を対象とした未知語の調査日本放射線技術學會雑誌、 2008

[6] 湯本 紘彰、森 辰則、中川 裕志、出現頻度と連接頻度に基づく専門用語抽出情報処理学会研究報告、2001

[7] 池野 篤司、濱口 佳孝、山本 英子、井佐原 均、Web 文書集合からの専門用語獲得情報処理学会論文誌、2006

[8] 三浦 康秀、増市 博、部分文字列のパープレキシティを利用した低頻度専門用語抽出電子情報通信学会技術研究報告、 2007

[9] Kyo KAGEURA, Bin UMINO, Methods of Automatic Term Recognition-A Review Terminology, 1996

[10] Petr Knoth, Marek Schmidt, Pavel Smrz and Zdenek Zdrahal ,Towards a Framework for Comparing Automatic Term Recognition Methods Znalosti、2009

[11] Ahmad, K., Gillam, L., and Tostevin, L. University of Surrey participation in TREC 8: Weirdness indexing for logical document extrapolation and retrieval (WILDER),1998

[12] Katerina Frantzi , Sophia Ananiadou , Hideki Mima , Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method International Journal on Digital Libraries、 2000

[13] 森　信介、長尾　眞、n グラム統計によるコーパスからの未知語抽出電子情報通信学会技術研究報告、1995

[14] 下畑 さより、杉尾 俊之、隣接文字情報を用いた n-gram 抽出文字列からの名詞句の自動抽出情報処理学会研究報告、1996

[15] 長尾 眞、森 信介 大規模日本語テキストの n グラム統計の作り方と語句の自動抽出情報処理学会研究報告、1993

[16] 鍛治伸裕、喜連川優、文脈にもとづく未知語獲得における識別モデルの適用言語処理学会、2009

[17] S. Shimohata, T.sugio and J.Nagata, Retrieving collocations by co-occurrence and word orRetrieving collocations by co-occurrenceder constraints , Proceedings of ACL/EACL 1997