能動学習による複合語を考慮した専門用語抽出

小田倉史麿† 小林 滉河† 若林 啓 ††

† †† 筑波大学 〒 305–8550 茨城県つくば市春日 1-2
E-mail: †{s1711496,s1921631}@s.tsukuba.ac.jp, ††kwkaba@slis.tsukuba.ac.jp

あらまし 専門用語抽出とは，コーパス中から専門用語を抽出するタスクである．専門家による人手での抽出では，専門家の作業負担が大きいため，自動で専門用語を抽出する技術の高度化が求められている．既存の研究には，専門家によって与えられた語と並列関係にあたる語を探し出すというアプローチがある．しかし，専門家が念頭におく文脈を十分に考慮できないために，専門家の意図に沿わない用語群が抽出されてしまう懸念がある．本研究では，能動学習を適用することで，人間による教師データ作成のコストを抑えつつ，専門用語の帰納的な類推を行う専門用語抽出の枠組みを提案する．また，複合語を考慮するために，事前に作成した転置インデックスを用いることで，複合語の動的な発見を行う．実験では，既存手法と提案手法に能動学習を適用したうえで，既知の専門用語群をテストデータとし，抽出性能を比較した．

キーワード 自然言語処理，専門用語抽出，能動学習，複合語，転置インデックス

1 はじめに

専門用語を収集および整理することは，今後の研究に資する活動である．しかし，従来では，専門家が人手で研究論文から専門用語を収集することが多く，その作業にかかるコストの大きさゆえに，用語情報を最新の状態に保つことが困難である．このため，コーパス中から自動で専門用語を抽出する技術の高度化が求められている．

中川ら [1] は，単言語コーパスからの用語抽出には 3 つのフェーズがあるとし，第 1 フェーズを「用語の候補の抽出」，第 2 フェーズを「第 1 フェーズで抽出された候補に対する用語としての適切さを表すスコア付けないし順位付け」，第 3 のフェーズを「順位付けられた用語候補集合の中から適切な数の候補を用語として認定すること」と論じた．専門用語の自動抽出は，専門家が行う判断をコンピュータに代替させることに相当する．このため，中川ら [1] による 3 つのフェーズに沿って，次のような課題がある．

まず，コーパス中には大量の語が含まれるため，それら全てについて「スコア付けないし順位付け」を行うには膨大な計算量を要する．このため，コーパス中の語について，一定数まで候補を絞り込む必要がある．次に，絞り込まれた候補に対してどのように「スコア付けないし順位付け」をするかという点においては，専門分野の文脈によって「専門用語としての適切さ」が異なるために，スコア算出の判断基準が曖昧になりやすいという問題がある．また,「専門用語としての認定」を行う段階においても「判断基準の曖昧性」が問題となる．専門用語としての認定は，その分野に精通した専門家でなければ行うことができない．しかし，専門用語を収集および整理するような段階では，専門家にとっても収集すべき専門用語を探っている状態であり，明確な正解が存在しない問題となる．このため，コーパス中から用語を抽出する精度を高めるという視点だけでなく，専門用語の候補を専門家に提示することによって，他の専門用語の想起を支援するという視点も必要となる．

コーパス中から用語を抽出するには，コーパス中の語を何らかの形でコンピュータに認識させる必要がある．そのアプローチのひとつに，語の意味を低次元ベクトルで表現する「分散表現獲得」がある．しかし，Mikolov ら [2] の手法に代表される分散表現獲得手法は，原則単語単位で処理を行うため，複数の単語のまとまりで 1 つの意味を表す語である「複合語」を考慮することができない．専門用語の中には複合語であるものも少なくない．本研究では,「複合語の考慮」と「専門用語の文脈を帰納的に類推し，専門家による専門用語の想起を支援すること」を目的として,「転置インデックスを用いてコーパス中の複合語の出現位置を高速に取得する手法」と「抽出器の学習と専門家への候補語の提示を繰り返す能動学習」を適用した専門用語抽出の枠組みを提案する．

本研究では，コーパスが大量の論文であるため，コーパス中のフレーズ全てに対してスコア計算をすることは現実的ではない．ただし，出現頻度による絞り込みでは得られるフレーズが限定的になる懸念がある．このため提案手法では，既存の言語モデルによる品詞分析のマッチングを用いて抽出候補の絞り込みを行う．また,「候補の獲得頻度」または「共起ベクトルを学習させた SVM（Support Vector Machine）が推定する所属確率」のいずれかを用いて，候補の効果的な並べ替えを試みる．実験では，既存手法に対して提案手法と同様に能動学習を適用したうえで，一定の文脈によって分類された既知の専門用語群をテストデータとして，抽出性能を比較する．

本稿の構成は次の通りである．第 2 章では，本研究に関連する先行研究や概念について紹介する．第 3 章では，複合語の考慮と能動学習を適用した専門用語抽出の枠組みを提案し，比較対象とする既存手法を併せて説明する．第 4 章では，既存手法

と提案手法の抽出性能を比較する実験を行い，複合語の考慮および能動学習が専門用語抽出において有効であるかを評価する．第 5 章では，結論と本研究の今後の展開の議論を行う．

2 関 連 研 究

2.1 同位語抽出

一般に，専門用語抽出と呼ばれる研究では，コーパス中から「専門用語らしい語」を一緒くたに抽出し，既知の正解リストと比較することで抽出性能を検証するアプローチに主眼が置かれる場合が多い．一方で，与えられた語と共通の文脈で専門用語となるような語を探し出していく「同位語抽出」と呼ばれるアプローチが存在する．

大島ら [3] は，並列助詞「や」で並べられる語は同位語である可能性が高いことに着目した．並列助詞「や」の前後に出現するフレーズの出現頻度を考慮することで，単語だけでなく複合語の抽出も可能にする同位語抽出手法を提案した．また，大島らは同手法を英語の場合にも適用させており，日本語の「や」に代わって，接続詞の「or」に注目する手法を提案した [4]．しかし，Web 検索エンジンによって得られる情報のみを用いる手法であるため，同位語が含まれる文章が Web 検索によって十分に得られる場合でなければ抽出は困難である．水越ら [5] は，分散表現による単語間の類似度を用いて同位語候補を獲得し，SVM (Support Vector Machine) によって候補の並べ替えを行う同位語抽出手法を提案した．しかし，この手法は単語に限定した手法であり，複合語を考慮していない．また，Ghahramani ら [6] は，ベイズ推定の枠組みを用いて，与えられた語と並列関係になるような語 (同位語) をランキング形式で返す Bayesian Sets という手法を提案した．

2.2 分散表現獲得

分散表現とは，語の意味を低次元のベクトルで表現する方法である．Mikolov ら [2] の分散表現獲得手法は，Word2vec という名で様々な自然言語処理タスクに幅広く利用されている．分散表現は，$king - man + woman = queen$ というような, 語同士の演算を可能することや，単語間の類似度の計算を可能にすることが特徴である．しかし，Word2vec は基本的に単語にしか対応しておらず，複合語に対応していない．このため，学習した言語モデル中に抽出すべき用語が含まれないという事態が懸念される．

Word2vec の入力に複合語を与える手段として，構成単語のベクトルを合成するといったアイデアが存在する．たとえば，Turian ら [7] は，構成単語ベクトルの和または要素ごとの積によって，複合語ベクトルを合成する方法を示した．しかし，それぞれの構成単語には複合語とは無関係の特徴量も含まれているため，合成されたベクトルが複合語のベクトルとして必ずしも適切であるとはいえない．また，Word2vec は事前に学習した単語のみの分散表現を獲得するため，複合語を Word2vec の出力として獲得することは原則不可能である．そこで Mikolov ら [2] は，前処理によって単語間の結び付きの強さを計算し，結びつきの強い語を複合語とみなして，ひとつの単語のように振舞わせるというアイデアを示した．しかし，前処理によって獲得できる複合語の量に限界があるため，コーパス中に低頻度で出現するような専門用語を獲得することができない懸念がある．分散表現は語に多数の意味属性を同時に持たせることができる表現方法でもある．このため，文脈によって変化する語の多様で曖昧な意味性を考慮するという視点において，専門用語抽出に有効な表現方法である．

2.3 固有表現抽出

固有表現抽出（Named Entity Recognition）とは，文章中から人名・地名・組織名といった固有名詞や，時間表現・金額表現等の語句を抽出する自然言語処理技術である．情報検索や質問応答といった様々な応用タスクの精度を大きく左右させる要因になるため，固有表現抽出に関する研究が盛んに行われている．

専門用語の多くが固有表現に該当するため，固有表現抽出を行うことは，コーパス中の専門用語の候補を発見することに資する．また，複数の語にまたがってタグを付与することができるため，複合語の考慮が可能である．

アノテーションコーパスには，文章中の全ての単語にアノテーションが付いたフルアノテーションコーパスと，文章中の一部の単語にアノテーションが付いた部分的アノテーションコーパスの 2 種類が存在する．専門家分野のコーパスで，大量のフルアノテーションコーパスを得ることは困難である．このため，部分的アノテーションコーパスで学習可能な固有表現抽出手法が有効になる．

部分的アノテーションコーパスを利用した学習手法のひとつに，Jie ら [8] の手法がある．Jie らは，部分的アノテーションコーパスを 2 つのサブデータセットに分割し，固有表現抽出モデルをそれぞれ分割したサブデータセット毎に交互に学習させることで，最終的なモデルを得るという手法を提案した．

2.4 能 動 学 習

少量のアノテーションコストでより良い精度を出す方法として，能動学習 [9] が知られている．能動学習は，ラベル付きデータを得ることに高いコストがかかるタスクや，ラベル付けを行うことに膨大な時間がかかるようなタスクに対して適用される手法である．能動学習では，ラベル付けされたデータを元に機械学習モデルを学習し，そのモデルを用いて大量のラベル無しデータ中から，ラベルが付けられることが学習に効果的であるインタンスを，選択的にアノテータに提示する．アノテータは提示されたラベル無しデータにラベルを付け，再度機械学習モデルを学習させる．これを繰り返すことによって，アノテーションコストを抑えつつもより良い精度を目指すという仕組みである．

能動学習の特徴は，低コストで機械学習モデルが得られることに加えて，アノテータに提示される情報が厳選されていくことにもある．このため，アノテータは単にラベル付に伴う判断を行うだけでなく，提示される情報を元に，新たな知見を得ることが可能である．

3 手　　法

3.1 問題設定

本研究では，専門家が把握するいくつかの専門用語例を手掛かりに，コーパス中から専門用語の候補を自動で抽出し，専門家に提示する枠組みを提案する．専門家が念頭におく文脈に従った専門用語を収集することは，専門用語辞書の作成に不可欠である．しかし，そうした専門用語が大量にある場合，専門家がその全てを列挙することは困難である．ただし，いくつかの具体例を思い浮かべることは比較的容易であるため，はじめに専門家によって少数の事例が与えられるという問題設定とする．

文章の言語は，学術研究の分野でグローバルスタンダードな言語である英語とする．

3.2 複合語の考慮

本研究では，複合語を考慮した専門用語抽出を行う．Mikolov ら [2] による複合語をひとつの単語と見なす方法を用いることで，複合語の分散表現を獲得すること自体は可能であるが，コーパス中のあらゆる複合語の出現位置を獲得することは，その組み合わせの多さゆえに，現実的ではない．コーパス中の複合語の組み合わせの数は，次のように計算できる．n 語からなる文章に対する，複合語の組み合わせ数 $C_{sentence}(n)$ は，$N = 2, ..., n$ における N-gram の数の総和によって，以下の式で計算される．

$$C_{sentence}(n) = \sum_{i=1}^{n} i - n = \frac{1}{2}n^2 - \frac{1}{2}n = \frac{1}{2}n(n-1) \quad (1)$$

コーパス中の文の数を m とし，各文の単語数が $S_j(j = 1, .., m)$ で表されるとき，コーパス中全体の複合語の組み合わせ $C_{corpus}(m)$ は，以下の式で計算される．

$$C_{corpus}(m) = \sum_{j=1}^{m} C_{sentence}(S_j) \quad (2)$$

これをうけて，本研究では，予めコーパス中の複合語に対して処理を行うのではなく，学習の過程において，必要に応じて複合語の出現位置を動的に獲得する手法を提案する．複合語の出現位置を動的に獲得するために，単語単位の出現位置を獲得する転置インデックスを事前に作成し，複合語の構成要素となる各単語の出現位置を効率的に探索する．まず，以下の手順でコーパスから転置インデックスを作成する．

（1） コーパス中の全単語の出現頻度をカウントする．

（2） コーパス中の全単語について，出現頻度が大きい順に単語 ID を 0 から順に振る．

（3） 単語と単語 ID の相互変換辞書を作成する．

（4） コーパス中の各文に文 ID を振る．

（5） コーパス中の各文を単語単位に区切り，単語 ID 列に変換する．

（6） 単語 ID から, コーパス中のどの文に出現するかを逆引きする転置インデックスを作成する (単語 ID を key として与えると, その単語が含まれる文の文 ID の集合が得られる)

（7） 単語 ID と文 ID の組から, 該当単語が該当文中のどの位置に出現するかを逆引きする転置インデックスを作成する (単語 ID と文 ID の組を key として与えると, 該当単語の該当文中の出現位置の集合が得られる)

上記の手続きで作成した転置インデックスを用いることで，指定された単語のコーパス中の出現位置を高速に取得することができる．文 ID の転置インデックスと，出現位置の転置インデックスを別に作成する理由は，両者をまとめた場合の転置インデックスのデータサイズが膨大になるためである．

次に，以下の手順で複合語の出現位置を獲得する．

（1） 与えられた複合語を構成単語に分割する．

（2） 単語 ID が最も大きい語 (コーパス全体での出現頻度が最も小さい語) を基準語とし，複合語中の他の単語について，基準語との相対位置を取得する．

（3） 基準語が出現する文 ID 全てを，転置インデックスを用いて取得する

（4） 文 ID に紐付けられた文章 (単語 ID 列) を取り出す．

（5） 単語 ID 列中の，基準語の出現位置を取得する．

（6） 複合語を構成単語の単語 ID が基準語に次いで大きい順に,「単語 ID 列中の基準語の出現位置」から「基準語との相対位置」だけ離れた語が，複合語の構成単語と一致するかを判定する．

（7） 途中で一致が認められなかった場合は以降の判定処理をスキップし，複合語の構成単語全てにおいて一致が認められた場合，その文を「複合語が出現する文」とする．

上記の手続きでは，出現頻度が少ない単語順に判定処理を行うため，正しい複合語にならない単語列の並びを探索する無駄を抑えることができる．この手続きに基づいて，複合語を判定し，複合語が認められる例を図 1 に，複合語が認められない例を図 2 に示す．

本研究では，この手法に基づいて予め単語単位での転置インデックスを作成しておき，学習の過程において必要に応じて，複合語の出現位置を獲得していく．ただし，複合語の構成単語が全て出現頻度の高い語で構成されるような場合，その中で出現頻度が低い語を基準に探索したとしても，相応の時間を要することが予想される．そういった状況に陥っても，適度な時間で探索を終えるために，第 4 章の実験では，基準語の出現文の

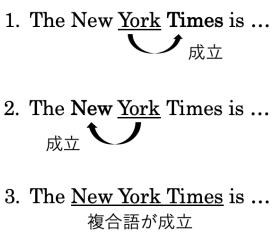

図 1　複合語「New York Times」が認められる例

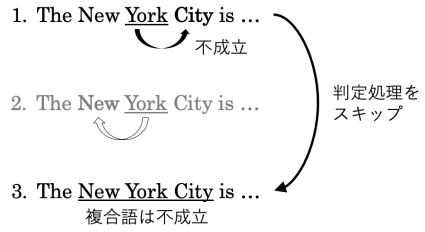

図 2　複合語「New York Times」が認められない例

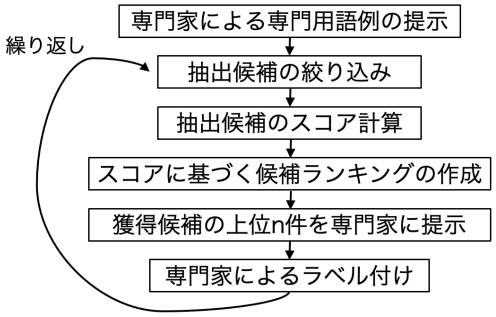

図 3　本研究における能動学習の流れ

文 ID を一定数に制限したり，一定数の出現位置を獲得した時点で探索を打ち切ったりするなどの制限を行う．

3.2.1　能動学習による専門用語抽出

本研究では,「候補の抽出」と「専門家によるラベル付与」を繰り返し行う，能動学習を行う．能動学習は，専門家による作業コスト抑える効果だけでなく，抽出モデルの精度向上に伴って，専門家に提示される情報が次第に洗練されていくことで，専門用語の想起を支援する有効性も見込める．本研究における能動学習の流れを，図 3 に示す．はじめに，専門家によって m 個の専門用語が与えられる．与えられる専門用語は，単語・複合語を問わず，複合語の構成単語の数も問わない．与えられた専門用語をもとに，専門用語候補を獲得する．獲得した全候補に対してスコア付けを行い，そのスコアに基づいて並べ替えた候補の，上位 n 個を専門家に提示する．専門家は提示された n 個の候補に対して，自身が念頭に置く文脈に適合した専門用語である場合は正解ラベルを，そうでない場合は不正解ラベルを付与する．ラベルが付与された n 個のうち，正解ラベルが付与された語をもとに次の専門用語候補を獲得する．スコア付けには不正解ラベルが付与された語の情報も利用される．これを繰り返すことで，専門家の意図を段階的に反映させるような専門用語抽出を行う．次節，次々節で述べる既存手法および提案手法には，本節で述べた形式の能動学習を適用させる．

3.3 既 存 手 法

本節では，既存手法を本研究における問題設定に適用した場合について説明する．

3.3.1　Word2vec による候補獲得と SVM による並べ替え

既存手法の 1 つ目として，水越ら [5] による同位語抽出手法を挙げる．この手法では，複数の単語を入力として，Word2vec における入力単語の分散表現の平均ベクトルを計算し，そのベクトルと類似度が高い語上位一定数を同位語候補とする．そして，同位語候補と入力単語の差のベクトルを SVM に学習させる．

Word2vec と SVM の併用による専門用語抽出の流れは次の通りである．まず，複合語あるいは単語が n_1 個与えられる．与えられた単語あるいは複合語について，分散表現を取得する．単語の場合は Word2vec の学習済みモデル内の分散表現をそのまま利用し，複合語の場合は，構成単語全ての Word2vec の学習済みモデル内の分散表現の平均を，複合語の分散表現として採用する．ある複合語の分散表現 $V_{compound}$ は，その複合語を構成する単語の数を n_2，複合語の i 番目の単語を W_i と表現すると，以下のように計算できる．

$$V_{compound} = \frac{\sum_{i=1}^{n_2} V_{W_i}}{n_2} \tag{3}$$

Word2vec モデルに入力するベクトル V_{in} は，与えられた n_1 個の，単語あるいは複合語のベクトル $V_i (i = 1, ..., n_1)$ を用いて，以下のように計算できる．

$$V_{in} = \frac{\sum_{i=1}^{n_1} V_i}{n_1} \tag{4}$$

V_{in} を Word2vec モデルに入力することで，モデル中に存在する，V_{in} とのコサイン類似度が大きい単語上位 n_3 件を専門用語候補として抽出する．

次に，SVM の利用について述べる．学習の最初期では，正例を n_1 個の V_i とし，負例をコーパス中からランダムに選んだ n_1 個のベクトルとして，SVM の学習を行う．次に，V_{in} と抽出された専門用語候補，$V_{out_i} (i = 1, ..., n_3)$ の差ベクトルを求め，これを SVM 用のベクトル V_{svm_i} とする．この計算式は以下のようになる．

$$V_{svm_i} = V_{in} - V_{out_i} \tag{5}$$

この，V_{svm_i} のラベルを SVM に予測させ，得られた正解ラベルに対する所属確率でソートする．ソートされた上位 n_4 件を専門家に提示し，正解ラベルおよび不正解ラベルを付与してもらう．これを SVM に学習させる．この一連の処理を繰り返すことで，専門用語を獲得する．この手法では，抽出可能な専門用語は単語に限定される．これは，入力に複合語を与えることはできるものの，複合語の分散表現が Word2vec のモデル中にないために，出力に複合語を得ることが出来ないためである．

3.3.2　固有表現抽出による候補獲得

既存手法の 2 つ目として，固有表現抽出における Jie ら [8] の手法を挙げる．この手法では，文中の一部の語にのみ固有表現タグがついた部分的アノテーションコーパスを利用して，固有表現抽出モデルを分割した 2 つのサブデータセット毎に学習させ，制約付きビタビアルゴリズムによってアノテーションを行うことを交互に繰り返す．本研究では，この手法を採用して次のように専門用語抽出を行う．

（1） 与えられた単語あるいは複合語について，その語が出現する文 ID を転置インデックスを用いて獲得する．

（2） 取得した各文について，与えられた語のみに固有表現タグを付与した部分的アノテーションコーパスを機械的に作成する．

（3） 部分的アノテーションコーパスを用いて，Jie らの手法で Bi-LSTM-CRF モデルを学習させる．

（4） Bi-LSTM-CRF モデルに固有表現タグを予測させる．

（5） ラベル予測の自信を表すスコアが大きい文章順に，固有表現タグが付与された語を n 件に到達するまで獲得する．

（6） 獲得した n 件の固有表現を専門家に提示し，正解・不正解のラベルを付与してもらう．

（7） 正解ラベルが付与された語を次の入力として，1〜6 を繰り返す．(部分的アノテーションコーパスは，追加更新していく)

この手法では，機械的に部分的アノテーションコーパスを得ることができるため，専門家に直接アノテーション作業を行ってもらう必要がない．また，入力・出力ともに複合語を考慮することができる．しかし，能動学習の度に Jie らの手法によるモデルの学習を繰り返し行う必要があり，実行時間が膨大になる可能性がある．また，能動学習によって正例が増えていった場合，新たな部分的アノテーションコーパスを得ることができるが，一度の学習で正解がひとつも発見できなかった場合，部分的アノテーションコーパスを追加更新することができないという問題がある．

3.4 提 案 手 法

3.4.1 品詞マッチングによる絞り込み

本研究では，既存の学習済みモデルを用いた品詞マッチングによる絞り込み手法を提案する．次のようにして品詞マッチングを行う．

（1） 与えられた単語あるいは複合語について，その語が出現する文 ID を転置インデックスを用いて獲得する．

（2） 構成単語の類似単語を用いて追加の候補文を獲得する．

（3） 既存の学習済みモデルによって，与えられた語の品詞タグを予測する．

（4） 与えられた語が出現する文，および追加候補文に対して，既存の学習済みモデルによって品詞タグを予測し，与えられた語の品詞タグとマッチする語を全て取り出し，専門用語候補とする．

（5） 次節で述べる候補の並べ替え手法を適用し，その上位 n 件を専門家に提示する．

この提案手法では，固有表現抽出器学習の不安定さや，膨大な学習時間を抑えることが可能である．しかし，既存の学習済みモデルによる品詞タグの予測性能に依存するため，このモデルが品詞タグを適切に予測できない場合，得られる候補が不適切になることが見込まれる．また，もうひとつの問題として，抽出される専門用語が，与えられた語と同じ語数の用語に限定されることがある．また，同一品詞のパターンというのは候補量が非常に多くなってしまうことが予想されるため，その全てをアノテータにラベル付けさせるのは現実的ではない．そのため，次項で述べる候補の並べ替えを効果的に行い，アノテータに提示する上位 n 件中に正解が含まれやすくすることが重要となる．

3.4.2 候補の獲得頻度による並べ替え

前項における品詞マッチングでは，得られる候補の量が膨大になるため，候補に対するスコアを算出し，それに基づいて候補を並べ替える方法を 2 つ提案する．

1 つ目の並べ替え方法として，候補の獲得頻度による並べ替えを提案する．これは，候補を得る際に，その獲得頻度をカウントし，その獲得頻度が大きい順にランキングを作成するというものである．ある複合語 W について，ランキングに用いる $Score(W)$ は，獲得された頻度 $f(W)$ をそのまま使い，以下のように表される．

$$Score(W) = f(W) \tag{6}$$

この方法のメリットは，スコアの計算が単純であるという点にある．一方で，コーパス中の出現頻度が小さいような語を抽出することが困難になる．本研究の専門用語抽出においては，特定の文脈に従った用語群を得ることを目的にしているため，コーパス中の出現頻度が大きい語が必ずしも重要な語であるとは限らない．このため，獲得頻度よる並べ替え手法は，簡易的な手法のひとつとして位置付ける．

3.4.3 共起ベクトルを学習させた SVM が推定する所属確率による並べ替え

獲得頻度よるランキングは，コーパス中の出現頻度が小さいような語を抽出することが困難である．この問題を解消するために，得られた候補全てに対して，共起ベクトルを計算し，SVM を学習させる手法を提案する．共起ベクトルとは，Word2vec において分散表現を獲得する途中過程で計算されるベクトルである．ある語の共起ベクトルは，コーパス中でその語の前後一定範囲に出現する語の出現頻度をカウントし，その頻度を元に作成される．共起ベクトルの要素数は，コーパス中の語彙数分あり，i 番目の要素は，i 番目に該当する単語との共起頻度である．

この共起ベクトルを，正解・不正解の 2 値分類問題として，SVM に学習させる．SVM に学習させるラベルは，0 を不正解，1 を正解とする．ある複合語 W について，ランキングに用いる $Score(W)$ は，SVM が返す，正解ラベルへの所属確率 $P(C = 1|W)$ となり，以下のように表される．

$$Score(W) = P(C = 1|W) \tag{7}$$

並べ替えられた候補ランキングの，上位 n 件を専門家に提示し，ラベルの付与・抽出候補の獲得・SVM の学習を繰り返していく．

共起ベクトルを学習させた SVM が推定する所属確率による並べ替えでは，候補語における，文章中の出現情報を元にランキング付けできるため，コーパス中の出現頻度が小さいような語であっても抽出できる．しかし，共起頻度の計算には相応の時間がかかる上に，得られた候補全ての特徴量を計算するとなるとその時間は膨大になる．また，得られる正例に比べて得られる負例が多い状況が想定されるため，SVM の性能を向上させるには，ある程度まで学習を繰り返す必要性が見込まれる．

4 評価実験

4.1 実験に用いるコーパス

本実験では，コーパスとしてウィキメディア財団 [10] が公開している記事のダンプデータ「enwiki-20201001-pages-articles-multistream.xml.bz2」を使用した．Wikipedia の全ての記事をコーパスとすると，コーパスサイズが膨大になるため，一部の記事のみに限定したコーパスを作成した．作成したコーパスの，語彙数は 320,646，文の数は 736,975，記事数 5858 である．このコーパスにおいて，3.2 節で述べた手法にしたがって，転置インデックスを作成した．また，Word2vec モデルを，次元数を 200，最小出現頻度数を 2，window size を 5，epoch 数を 5 として学習させた．

4.2 評価に用いる既知の専門用語群

本研究では，既知の専門用語群を正解リストとして，専門家によるラベル付与を擬似的に行う．これは，本研究の能動学習では「専門用語として適切かどうか」のラベルを専門家が付与することを想定しているものの，実験段階として特定分野に精通した専門家に協力を依頼することが困難であるためである．

実験には，次の 2 つの用語群を用いる．

- ノーベル賞受賞者
  - 総数：574
  - コーパス中に存在する数：535
  - 含まれる複合語の数：535
- 内陸国
  - 総数：44
  - コーパス中に存在する数：39
  - 含まれる複合語の数：3

4.3 実験方法

能動学習による複合語を考慮した専門用語抽出の有効性を検証するために，既存手法と提案手法の抽出性能を比較する実験を行う．既存手法では，第 3 章で述べた「Word2vec の類似度計算による絞り込みおよび SVM によるソートを行う能動学習」と,「Jie らの手法による固有表現抽出を行う能動学習」の 2 つを行う．提案手法では，第 3 章で述べた「品詞マッチングによる絞り込みおよび獲得頻度で並べ替えを行う能動学習」と,「品詞マッチングによる絞り込みおよび共起ベクトルを学習させた SVM で並べ替えを行う能動学習」の 2 つを行う．実験結果は，横軸を「専門家がラベルを付与した候補語の件数の累積」，縦軸を「新たに発見された正解の累積」とするグラフで示す．グラフの凡例は,「Word2vec の類似度計算による絞り込みおよび SVM によるソートを行う能動学習」を「word2vec+svm」,「Jie らの手法による固有表現抽出を行う能動学習」を「jie's method」,「品詞マッチングによる絞り込みおよび獲得頻度で並べ替えを行う能動学習」を,「proposed method frequency sort」,「品詞マッチングによる絞り込みおよび共起ベクトルを学習させた SVM で並べ替えを行う能動学習」を「proposed method svm sort」と表記する．

実験では，以下の項目を変化させる．

- 正解ドメイン（「ノーベル賞受賞者」または「内陸国」）
- 専門家がはじめに与える用語（初期正例）
- 能動学習の繰り返し数
- 一度の学習で専門家がラベルを付与する候補数

本実験では，実行時間を抑えるために探索の制限を行う．与えられた候補語が含まれる文を取得する数 20 件にするなど，一定の制限を行うが，実験全てで統一する．以下に，各実験のパラメータを示す．

- 実験 1
  - 正解ドメイン：ノーベル賞受賞者
  - 初期正例：Max Planck, Ivan Pavlov, Albert Einstein
  - 能動学習の繰り返し数：20
  - 一度の学習で専門家がラベルを付与する候補数：25
- 実験 2
  - 正解ドメイン：ノーベル賞受賞者
  - 初期正例：Max Planck, Ivan Pavlov, Albert Einstein
  - 能動学習の繰り返し数：100
  - 一度の学習で専門家がラベルを付与する候補数：5
- 実験 3
  - 正解ドメイン：ノーベル賞受賞者
  - 初期正例：Max Planck, Ivan Pavlov, Albert Einstein
  - 能動学習の繰り返し数：10
  - 一度の学習で専門家がラベルを付与する候補数：50
- 実験 4
  - 正解ドメイン：内陸国
  - 初期正例：Afghanistan, Bhutanm, North Macedonia
  - 能動学習の繰り返し数：20
  - 一度の学習で専門家がラベルを付与する候補数：25
- 実験 5
  - 正解ドメイン：内陸国
  - 初期正例：Afghanistan, Bhutanm, North Macedonia
  - 能動学習の繰り返し数：100
  - 一度の学習で専門家がラベルを付与する候補数：5
- 実験 6
  - 正解ドメイン：内陸国
  - 初期正例：Afghanistan, Bhutanm, North Macedonia
  - 能動学習の繰り返し数：10
  - 一度の学習で専門家がラベルを付与する候補数：50

4.4 実験結果

実験 1〜6 の結果を，図 4，図 5，図 6，図 7，図 8，図 9 に示す．また，実験 1〜6 について，手法ごとの実行時間を，表 1 に示す．実験 1〜3 の提案手法おいて発見された正解には，「Werner Heisenberg」「Walther Nernst」「Niels Bohr」などが，不正解には「David Oistrakh」「John Gofman」「Trinity College」などが見られた．実験 4〜6 の提案手法において発見された正解には,「Nepal」「Mongolia」「Laos」などが，不正解には「Iran」「Ljubljana」「Academy」などが見られた．

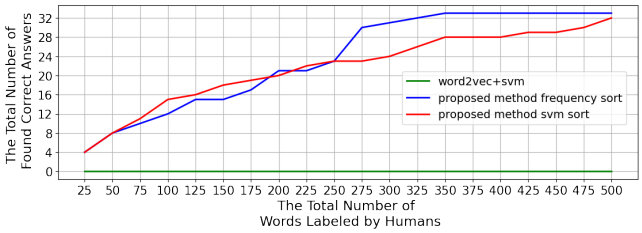

図 4　実験 1 の結果

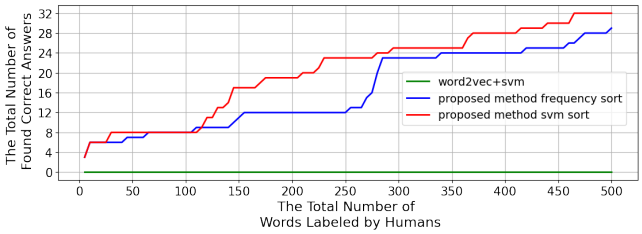

図 5　実験 2 の結果

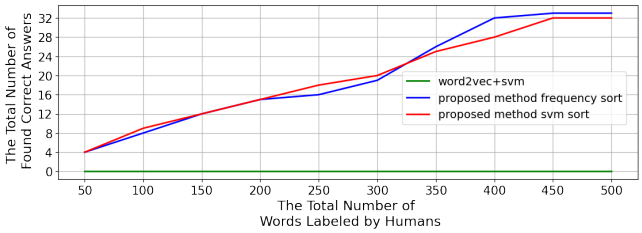

図 6　実験 3 の結果

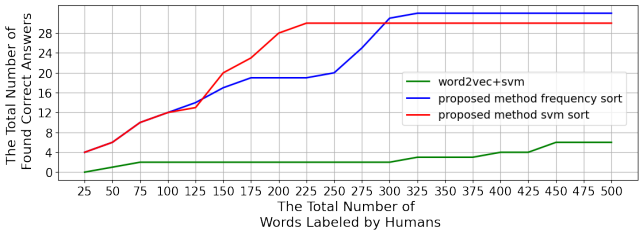

図 7　実験 4 の結果

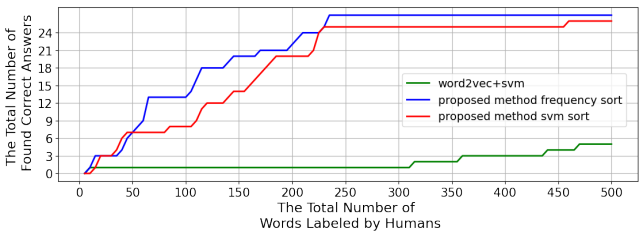

図 8　実験 5 の結果

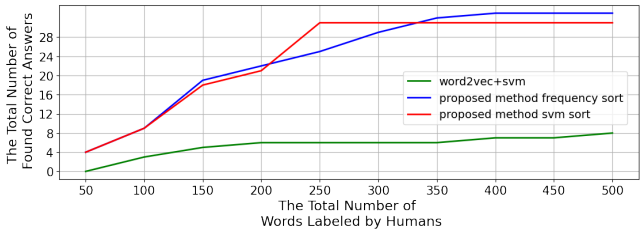

図 9　実験 6 の結果

表 1　実験 1〜6 における手法ごとの実行時間 (秒)

	word2vec +svm	jie's method	proposed method frequency sort	proposed method svm sort
実験 1	4	264	15	121
実験 2	19	1204	15	234
実験 3	2	125	16	121
実験 4	5	351	18	79
実験 5	20	1687	13	238
実験 6	3	169	15	53

実験 1〜6 において，Jie らの手法 (固有表現抽出) では候補語を全く獲得できなかった．このため，専門家によるラベル付与のプロセスが発生せず，グラフ上にプロットされなかった．Jie らの手法において，学習時に得られた部分的アノテーションコーパスを確認すると，与えられた候補以外の固有表現タグを全く予測できていなかった．

4.5 考　　察

図 4，図 5，図 6 では，「word2vec+svm」が全く正解を獲得できなかったが，実験 1〜3 では正解リストに複合語しか存在しないため，当然の結果である．一方で「proposed method svm sort」および「proposed method frequency sort」では，最終的に 25〜30 程度の正解を獲得できた．これにより，提案手法が複合語の専門用語を一定数抽出できることを確認できた．また，図 7，図 8，図 9 においても，「proposed method svm sort」および「proposed method frequency sort」は，「word2vec+svm」の抽出性能を上回った．実験 4〜6 で用いた「内陸国」の正解リストには，複合語でない正解が多く存在する．このため，複合語でない正解が正解リストに多く含まれる場合においても，提案手法が既存手法の抽出性能を上回る場合があることを確認できた．この原因としては，「word2vec+svm」では複合語のベクトルを適切に学習できなかったことが懸念される．このことから，提案手法は，複合語の専門用語を獲得できるだけでなく，複合語の入力を単語と同じように受け付けるという点でも，一定の効果があると推察する．

実験 1〜6 では，最終的に専門家がラベルを付与する件数を 500 で統一し，能動学習の繰り返し数と，一度の学習で専門家がラベルを付与する件数のみを変化させた．その結果，図 4，図 5，図 6 および図 7，図 8，図 9 における提案手法では，最終的に得られる正解の数は同程度に収束した．本研究の能動学習では，「出来るだけ人間のラベル付与作業が少ない段階で，多くの正解を発見できている」ということが好ましい．すなわ

ち，グラフの傾きが大きくなるタイミングが学習初期に現れる事が好ましい．しかし，途中段階での差は僅かにあるものの，グラフ全体としては同程度に右肩上がりになっている．この原因としては，候補集合への依存性が懸念される．提案手法および既存手法は共通して，得られた正解を次の入力として芋づる式に候補を探し出す仕組みである．このため，与えられた候補から得られる次の候補は，SVM が専門用語の共通点を学習できている度合いに関わらず，同様の語になってしまう．すなわち，SVM がうまく学習できているか以上に,「得られた語から探索できる候補文に含まれる正解語」が影響してしまったことが懸念される．これは,「proposed method svm sort」と「proposed method frequency sort」に顕著な差が見られなかった原因としても挙げられる．また，候補中に含まれる正解語の数が，専門家に提示する候補数に及ばなかった場合などでは，並べ替えの効果に関わらず，一定数の不正解語が必ず提示されてしまうことになる．しかし，段階的に新たな正解が次々と発見されていくことから，与えられた正例を元にさらなる候補を探し出すことを繰り返すという枠組み自体は有効であるといえる．このため，与えられた語と共通する文脈をより効果的に学習さる方法を候補の獲得方法と併せて再検討することで，ラベル付与作業が少ない段階で多くの正解を発見できるようにすることが今後の課題である．

表 1 において，実験 1〜6 の全てで,「jie's method」が最も実行時間が長く,「proposed method svm sort」が 2 番目に実行時間が長い結果となった．この原因としては，Jie らの手法では，部分的アノテーションコーパスの作成と Bi-LSTM-CRF モデルの学習を能動学習ごとに繰り返すため，実行時間が長くなったことが懸念される．また,「proposed method frequency sort」よりも「proposed method svm sort」の方が，最低 3 倍以上実行時間がかかった．このことから，候補文の獲得よりも共起ベクトルの作成および SVM によるソートにかかる時間の方が長いことがうかがえる．また,「jie's method」および「proposed method svm sort」は，実験 2 と実験 5 において比較的長い実行時間を要した．本実験において実行時間が長期化することは，現実には専門家の待ち時間が増加することを意味する．このため，抽出の実行時間を短縮することが今後の課題である．

5 おわりに

本研究では，複合語の考慮および能動学習を行う専門用語抽出の枠組みを提案した．コーパス中の単語の出現位置を転置インデックスによって高速に取得することで，既存の手法では不可能であった，複合語の出現位置の動的な取得を可能にした．また，能動学習の枠組みを適用することで，専門家によって与えられた少数の語を起点に，専門家と対話を繰り返しながら，新たな専門用語を段階的に抽出することを可能にした．

実験結果では，提案手法が複合語を抽出可能であること，および既存手法に比べ，高性能に抽出できる状況があることを示した．ただし，候補の並べ替えについては，共起ベクトルを用いる方法が，獲得頻度を用いる場合と比べて有効であることは示せなかった．このため，挙げられた用語群に共通した文脈を類推するという部分には課題が残った．

また，実行時間という課題がある．提案手法は複合語の出現位置取得や共起ベクトルの計算に時間を要する．実験では既知の専門用語群を用いて機械的にラベル付けを行ったが，実際には専門家が繰り返しラベル付け作業を行うことを考えると，専門家の待ち時間が長くなってしまう懸念がある．

そして，候補集合への依存性という課題もある．案手法および既存手法は共通して，得られた正解を次の入力として芋づる式に候補を探し出す仕組みである．これにより，正解を抽出できるかは，獲得できた候補への依存が強く，候補の並べ替えが効果的に機能しない懸念がある．

このため，候補取得や特徴量計算のアルゴリズムをより高速にすることや，候補集合への依存性を少なくした能動学習を可能にすることが今後の目標である．

謝　辞

本研究の一部は，JSPS 科研費（課題番号 19K20333）および JST CREST (#JPMJCR16E3) AIP チャレンジの助成によって行われた．

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] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, Vol. 26, pp. 3111–3119, 2013.

[3] 大島裕明, 小山聡, 田中克己. Web 検索エンジンのインデックスを用いた同位語とそのコンテキストの発見. 情報処理学会論文誌データベース（TOD）, Vol. 47, No. 19, pp. 98–112, 2006.

[4] Hiroaki Ohshima, Satoshi Oyama, and Katsumi Tanaka. Searching coordinate terms with their context from the web. In Karl Aberer, Zhiyong Peng, Elke A. Rundensteiner, Yanchun Zhang, and Xuhui Li, editors, *Web Information Systems – WISE 2006*, pp. 40–47, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[5] 水越俊希, 杉本徹. 単語の分散表現を用いた同位語の抽出. 言語処理学会 第 23 回年次大会 発表論文集, pp. 683–686, 2017.

[6] Zoubin Ghahramani and Katherine A Heller. Bayesian sets. *Advances in neural information processing systems*, Vol. 18, pp. 435–442, 2005.

[7] Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 384–394, Uppsala, Sweden, 2010. Association for Computational Linguistics.

[8] Zhanming Jie, Pengjun Xie, Wei Lu, Ruixue Ding, and Linlin Li. Better modeling of incomplete annotations for named entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 729–734, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[9] Burr Settles. Active learning literature survey. *Machine Learning*, Vol. 15, No. 2, pp. 201–221, 1994.

[10] Wikimedia Foundation Inc. enwiki dump progress on 20201001. https://dumps.wikimedia.org/enwiki/20201001/, 2020.