人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第12回)
SIG-AM-12-01

語彙の概念化とWikipediaを用いた英字略語の意味推定手法

Meaning Estimation Method of Alphabetical Abbreviation using Conceptualization of a Word and Wikipedia

後藤　和人[1*]　土屋　誠司[2]　渡部　広一[2]
Kazuto Goto[1]　Seiji Tsuchiya[2]　Hirokazu Watabe[2]

[1] 同志社大学大学院　理工学研究科
[1] Graduate School of Science and Engineering, Doshisha University
[2] 同志社大学　理工学部
[2] Faculty of Science and Engineering, Doshisha University

Abstract: In this paper, we propose the method to estimate the meaning of an alphabetical abbreviation. To resolve the polysemy of alphabetical abbreviations, this method uses the concept base, Wikipedia, and, the evaluation approach for semasiological association between sentences. The concept base allows Conceptualization of a word. We apply Calculation of Degree of Association or Earth Mover's Distance (EMD) to evaluate semasiological association between sentences. This paper used 129 articles to evaluate the proposal method. 129 articles included 129 meanings of alphabetical abbreviations and 58 literation of alphabetical abbreviations. The experiments showed that the accuracy of the proposal method was close to 80%.

1　はじめに

元来から日本は，外来語を受け入れやすい環境にあるといわれており，数多くの外国の言葉を片仮名として表記し，そのまま使用している．近年になり，よりグローバル化が激しくなると共に，外来語が益々増加する中，外来語の発音を片仮名表記にしないケースが見受けられる．特に，英語の場合，外国語の表記をそのまま利用するシーンも増えてきている．また，英単語などの頭文字をつなげて表記する，いわゆる略語もよく利用されるようになっている．例えば,「IC」といった英字略語がそれにあたる．

しかし，英字略語は英単語の頭文字から構成される表現であるため，まったく別のことを表現しているにも関わらず，同じ表記になることが多い．先の英字略語「IC」を例にとると,「IC」には「集積回路」という意味や，高速道路などの「インターチェンジ」の意味がある．さらには，ある業界では，これらとはまた別の意味で使用されることもある．

このように，英字略語は便利な反面，いわゆる一般的な単語よりも非常に多くの意味を有する多義性の問題を持っている．そのため，英字略語が利用されている情報は，すべての人が容易に，また，正確に把握できているとは言い難い．

特に，知能を有し，情報収集をはじめとする手段により上記問題を解決できる人と比べ，機械にとって，英字略語の理解はより困難な問題といえる．近年，機械は我々の生活・社会と密接に関与し，必要不可欠な存在となりつつある．機械の目指すべき姿は「人と共存する機械（ロボット)」だといえる．二足歩行ができる身体能力に優れたロボットなどが数多く開発されたことにより，その一部が実現されつつある．今後，ロボットが真に「人と共存」するためには，優れた身体能力を持つロボットに「知能」を持たせ，その知能に基づいて会話することが重要になる．

ロボットに知能を持たせるためには，人が日常的に行っている常識的判断を実施できるようにすることが必要になる．そのためには，ロボットに単語を理解させることが必要であり，その手段として，語彙を概念化する研究が行われている．例えば，[1] では，ある概念に対して，当該概念の意味特徴を表す属性と属性の重みを登録したデータベースを構築している．そして，構築したデータベースをもとに，概念同士の意味的な近さを計算できる関連度評価アルゴリズムを提案している．また，[2] では，シソーラスを利用する方法と特異値分解を用いる数学的方法を併用して，単語間の属性空間を構築する方式が提案されている．さらに，[3] は，コーパスから抽出した情報をもとに，格フレームを

*同志社大学大学院理工学研究科情報工学専攻
〒 610-0394　京都府京田辺市多々羅都谷 1-3
E-mail: eup1102@mail4.doshisha.ac.jp

自動的に構築する手法を提案している．これらの手法は，それぞれアプローチは異なるものの，ロボット（計算機）に単語や文章を理解させることを目的している．語彙を概念化することで，単語の表記に頼らず，単語や単語同士の関係を表現することを実現している．結果，単語の意味を考慮した情報検索や応答が可能となり，常識を有するロボットの実現につながることが期待されている．

さらに，ロボットの知能を発展させるために，ユーザの入力（会話など）に対する応答ではなく，ロボットが起点となり，会話文を提示する研究がなされている [4]．この手法では，新聞記事のヘッドラインを手掛かりに，取得した情報の意味を理解した上で，情報を会話文形式に変換し，会話文を自動生成することを実現している．このとき，新聞記事に英字略語が含まれており，かつ，英字略語の意味を正しく理解できなかった場合，ロボットはユーザに誤った会話文を提示する可能性がある．このように，知能を有するロボットの研究開発において，多義性を有する英字略語の意味を正しく理解できなければ，自然な知的対話の実現は困難であるといえる．

英字略語の意味を明示するために，文章（新聞記事など）の中で最初に英字略語が使用される箇所において，括弧書きでその意味を日本語で併記する処置をとっていることが多い．しかし，よく知られている英字略語にはそのような処置をとらないなど，完全に対処されているわけではない．また，文章中において，最初の箇所にのみ上記のような処置がとられており，それ以降はその意味が併記されないことが多い．そのため，途中から文章を読んだ場合，最初にその英字略語が出現した箇所を探すことが必要になる．その結果，解読にはひと手間が必要となり，理解の妨げとなる．さらに，このように英字略語の意味を併記するという処置がとられていないこともある．

これらを踏まえて本論文では，英字略語の意味を推定する手法について提案する．本研究と同様の主旨の研究には，[5] の研究が報告されている．この手法では，先に述べたように新聞記事において英字略語の意味を括弧書きで併記する表現に着目し，英字略語の意味の自動推定を実現している．しかし，前述のように，すべての英字略語に対してこのような表現が適用されているわけではない．そのため，うまく自動推定できない場合がある．また，[6] では，片仮名表記の外来語を英語に復元した後に辞書を用いて日本語訳を獲得する手法が提案されている．しかし，本手法では，対象が片仮名語に限定されており，かつ，多義性を有する語彙には対応できない問題がある．

そこで，本論文では，多義性を有する英字略語に対して，意味の推定を実現する方法を提案する．提案手法では，我々がすでに提案している，あらゆる語彙の意味的な近さを判断できるメカニズムを組み合わせることで，英字略語の意味を推定する．具体的には，ある概念から様々な概念を連想する語彙の概念化処理が可能な概念ベース [7, 8, 9]，及び，世界で最も収録語数が多いとされている Wikipedia[10] を使用する．さらに，概念化した語彙の意味的な近さを判断するために，関連度計算 [11]，または，Earth Mover's Distance を応用した文章間関連度計算手法 [12] を用いる．これらを用いることで，英字略語の多義性を解消し，英字略語の本来の意味を推定する．この技術により，英字略語の意味を理解しやすくすることができ，自然な知的対話の実現に加え，情報検索や自動要約，情報推薦や自動翻訳など多くのアプリケーションの性能向上に寄与することも期待できる．

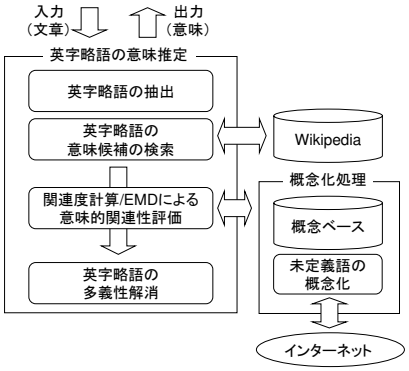

図 1: 英字略語の意味推定手法の概略図.

2 提案手法の概要

図 1 に提案手法である英字略語の意味推定手法の概略図を示す．英字略語が含まれる文章を入力として，その文章から英字略語を抽出する．その英字略語を Wikipedia で検索し，意味が 1 つであれば，その意味を出力する．意味が複数ある場合には，それらの意味と入力文章との意味的な近さを判断し，最も近いと判断した意味を決定する．なお，ここで述べる「意味」とは，英字略語の意味を表現する語，つまり，英字略語の基となっている英単語の日本語での表現を「意味」と定義している．例えば，前述した英字略語「IC」の意味を推定する場合,「集積回路」や「インターチェンジ」という語を「意味」として出力することになる．

3 章では，本論文において使用した要素技術として，語彙を概念化する手法と，概念化した語彙の意味的な近さを判断する手法に関して詳細に説明する．

表 1: 概念ベースの例.

概念	属性
医者	(医師, 0.34), (患者, 0.11), (病院, 0.08), ⋯
病院	(医院, 0.25), (手術, 0.18), (施設, 0.04), ⋯
患者	(病人, 0.52), (看病, 0.21), (治療, 0.12), ⋯
⋯	⋯

3 使用要素技術

3.1 語彙の概念化処理

3.1.1 概念ベース

概念ベース [7, 8, 9] とは，複数の電子化国語辞書などの見出し語を概念，その語義文に使用されている自立語を概念の意味特徴を表す属性と定義して構築された大規模なデータベースである．本論文で使用した概念ベースは自動的に概念および属性を構築した後，人間の常識に沿った属性の追加や削除を人手で行ったものであり，概念数は約 9 万語である．

概念ベースでは，ある概念 A は m 個の属性 a_i とその属性の重要性を表す重み w_i の対によって構成されており，以下のように表現することができる．ここで，属性 a_i を概念 A の一次属性と呼ぶ．

$$\text{概念 } A = \{(a_1, w_1), (a_2, w_2), \cdots, (a_m, w_m)\}$$

概念ベースの大きな特徴として，属性である単語は概念として必ず定義されている点がある．これにより，概念 A の一次属性である属性 a_i を概念とみなし，更に属性を導くことができる．概念 a_i から導かれた属性 a_{ij} を，元の概念 A の二次属性と呼ぶ．概念ベースの具体例を表 1 に示す．例えば，表 1 のように，概念「医者」の一次属性である「患者」は，概念「患者」としても定義されている．また，この概念「患者」の一次属性である「病人，看病，治療，…」は，元の概念「医者」の二次属性ということになる．

このように，概念ベースにより概念の意味特徴を定義し，連鎖できる構造を利用することで語彙の意味の近さを評価することができる．詳細は 3.2 節で説明する．

3.1.2 未定義語の概念化

前節で述べたとおり，概念ベースは複数の電子化国語辞書などを用いて構築されており，大規模かつ品質が高いというメリットがある．しかし，その反面，すべての語彙を網羅できていないという欠点もある．そのため，概念ベースに登録されていない未定義語は概念化されておらず，意味の近さを評価することはできない．

そこで，未定義語については，Web 上の言語情報を利用し，自動的に概念化すること [13, 14] で対処する．具体的には，Web 検索エンジン [15] により未定義語をキーワードとして情報検索し，検索結果の上位 100 件の検索結果ページの内容を取得する．その内容から概念ベースに登録されている自立語のみを抽出し，それらを未定義語の一次属性とする．また，一次属性に対する重みは，情報検索の分野で広く用いられている tf/idf[16] の考え方を応用することで算出する．

語の網羅性である tf 値は，検索結果ページ A 中に出現する自立語 $Word_{Ai}$ の出現頻度 $tfreq(Word_{Ai}, A)$ を，検索結果ページ A 中のすべての自立語の語数 $tnum(A)$ で割ることでで算出される．算出式は以下のようになる．

$$tf(Word_{Ai}, A) = \frac{tfreq(Word_{Ai}, A)}{tnum(A)}$$

次に，語の特定性である idf 値については，$SWeb\text{-}idf$ 値 [13, 14] を用いる．idf 値の算出には，対象となる全文書空間の情報が必要になる．しかし，Web を利用する場合，Web 上のすべての情報が必要ということになり，正確な idf 値を算出することは現実的には不可能である．

そこで，無作為に選択した固有名詞 1000 個をそれぞれキーワードとして，Web 検索エンジン [15] で検索する．続いて，検索結果の上位 10 件ずつ検索結果ページの内容を取得する．そして，それらの内容に含まれるすべての自立語の集合を疑似的に Web の全情報空間とみなし $SWeb\text{-}idf$ 値を算出する．$SWeb\text{-}idf$ 値の算出式は以下のように定義される．ここで N は，固有名詞 1000 個を検索キーワードとした際の各検索結果上位 10 件の合計ページ数（$N = 10000$），$df(Word_{Ai})$ は，$Word_{Ai}$ が出現する検索結果ページ数である．

$$SWeb\text{-}idf(Word_{Ai}) = \log \frac{N}{df(Word_{Ai})}$$

この 10000 ページから，複数の国語辞書や新聞などから概念を抽出したデータベースである概念ベースの収録語数である約 9 万語以上の単語数が得られたことから，獲得した 10000 ページを Web の全情報空間とみなしている．なお，固有名詞の選び方を変えても $SWeb\text{-}idf$ 値に大きな変化は見られないという報告がなされている [13].

以上に示した式より，自立語 $Word_{Ai}$ へ付与する重み w は次の式で定義される．つまり，ある自立語の重みは，網羅性を表す tf 値と特定性を表す $SWeb\text{-}idf$ 値を掛け合わせることで与えられる．

$$w = tf(Word_{Ai}, A) \times SWeb\text{-}idf(Word_{Ai})$$

Web 上の言語情報を利用するため，品質という面では概念ベースより劣るというデメリットはあるが，す

べての語彙を概念化できるという大きなメリットがある．また，[13, 14] などの研究成果から，本論文での処理においても十分な性能を確保していると考えられる．

3.2 意味的な関連性評価手法

3.2.1 関連度計算

関連度計算とは，概念と概念の関連の強さを定量的に評価するものである．概念と概念の間にある関連性を定量的に評価する手法として，ベクトル空間モデルが広く用いられている．しかし，本論文では，概念を定義する属性集合とその重みを含めた一致度に基づいた関連度計算方式を利用している．これは，関連度計算方式が有限ベクトル空間によるベクトル空間モデルよりも良好な結果が得られるという報告がなされているためである [1]．本論文では，重み比率付き関連度計算方式を使用する [11].

任意の概念 A，B について，それぞれ一次属性を a_i，b_j とし，対応する重みを u_i，v_j とする．また，概念 A，B の属性数を L 個，M 個（$L \leq M$）とする．なお，各概念の一次属性の重みは，その総和が 1.0 となるよう正規化している．

$$\begin{aligned} A &= (a_i, u_i) \mid i = 1 \sim L \\ B &= (b_j, v_j) \mid j = 1 \sim M \end{aligned}$$

このとき，概念 A，B の重み比率付き一致度 $MatchWR(A, B)$ は以下のように定義される．

$$\begin{aligned} MatchWR(A,B) &= \sum_{a_i = b_j} \min(u_i, v_j) \\ \min(\alpha, \beta) &= \begin{cases} \alpha & (\beta > \alpha) \\ \beta & (\alpha \geq \beta) \end{cases} \end{aligned}$$

この定義は，概念 A，B に対し，$a_i = b_j$ となる属性（概念 A，B に共通する属性）があった場合，共通する属性の重みの共通部分，つまり，小さい重み分のみ一致するという考えに基づいてる．

次に，属性の少ない方の概念を A とし（$L \leq M$），概念 A の属性を基準とする．

$$A = \{(a_1, u_1), \cdots, (a_i, u_i), \cdots, (a_L, u_L)\}$$

そして，概念 B の属性を，概念 A の各属性との重み比率付き一致度 $MatchWR(a_i, b_{xi})$ の和が最大になるように並び替える．

$$B_x = \{(b_{x1}, v_{x1}), \cdots, (b_{xi}, v_{xi}), \cdots, (b_{xL}, v_{xL})\}$$

これによって，概念 A の一次属性と概念 B の一次属性の対応する組を決める．対応にあふれた概念 B の属性は無視する．ただし，一次属性同士が一致する（概念表記が同じ）ものがある場合 $(a_i = b_j)$ は，別扱いにする．これは概念ベースには約 9 万語の概念が存在し，属性が一致することは稀であるという考えに基づく．従って，属性の一致の扱いを別にすることにより，属性が一致した場合を大きく評価する．具体的には，対応する属性の重み u_i，v_j の大きさを重みの小さい方にそろえる．このとき，重みの大きい方はその値から小さい方の重みを引き，もう一度，他の属性と対応をとることにする．例えば，$a_i = b_j$ で $u_i = v_j + \alpha$ とすれば，対応を決定する箇所は (a_i, v_j) と (b_j, v_j) であり，(a_i, α) はもう一度他の属性と対応させる．このように対応を決め，対応がとれた属性の組み合わせの数を T 個とする．

重み比率付き関連度とは，重み比率付き一致度を比較する概念の各属性間で算出し，その和の最大値を求めることで計算する．概念 A，B の関連度 $DoA(A, B)$ は以下の式で定義される．

$$\begin{aligned} DoA(A,B) = \sum_{i=1}^{T} \{ & MatchWR(a_i, b_{xi}) \cdot (u_i + v_{xi}) \cdot \\ & (\min(u_i, v_{xi}) / \max(u_i, v_{xi}))/2\} \end{aligned}$$

以下，重み比率付き一致度を一致度，重み比率付き関連度を関連度と略し，この関連度 [11] を用いる．関連度は概念間の関連の強さを 0〜1 の間の連続値で表す．

3.2.2 Earth Mover's Distance

前節において，概念間の関連の強さを評価する方法として関連度計算について記載した．関連度計算は関連性が高い順に属性の対応をとることで計算を行う，つまり，1 対 1 で対応をとる手法である．そのため，両概念の中で少ない方の属性の数しか対応がとれない．例えば，概念 A が持つ属性が 3 語，概念 B が持つ属性が 100 語であった場合，概念 B の属性 97 語は計算の対象外となる．本論文では，両概念の属性数に差がある状況にも対応するため，関連度計算に加えて，M 対 N で対応をとることができる Earth Mover's Distance（EMD）を用いて意味的な関連性を評価する手法を利用する．

EMD を用いた意味的な関連性評価手法は，ヒッチコック型輸送問題 [17]（需要地の需要を満たすように供給地から輸送を行う際の最小輸送コストを解く問題）で計算される距離尺度である EMD を概念間の関連性評価に適用したものである．EMD は，2 つの概念間の関連性を定量的に表現することが可能であり，[12] の研究によりその有用性が報告されている．

EMD とは，2 つの離散分布があるとき，一方からもう一方の分布への変換を行う際の最小コストを指す．離

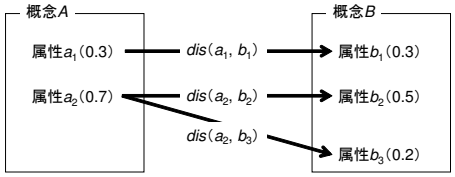

図 2: EMD を用いた意味的な関連性評価手法.

散分布はそれを構成する要素と重みの対の集合で表現される．コスト算出の際には，変換前の離散分布の要素が持つ重みを供給量，変換先の離散分布の要素が持つ重みを需要量と考え，要素間の距離を供給量，需要量にしたがって重みを運送すると考える．できるだけ短い距離で，かつ，需要量に対して効率的に重みを運送する経路が EMD となる．

これを概念間の関連性評価に適応させる際には，概念の一次属性を要素として捉え，一次属性の集合を離散分布と考える．ある概念の離散分布を違う概念の離散分布へ変換すると考えると，その際のコストが最小となる概念が元の概念に最も近い概念となり概念間の関連性評価へ適用することが可能となる．

EMD を用いた意味的な関連性評価手法について，図 2 に示すような簡略図を用いて説明する．ある概念 A と B があったとき，概念 A を概念 B に変換する際のコストを考える．それぞれの概念をそれらの一次属性 a_i，b_j の離散分布と考える．EMD では変換コストの算出を行う際に離散分布を構成する要素同士の距離を用いる．EMD を用いた意味的な関連性評価手法では，この距離を一次属性同士の関連性であると考え，一致度によってこれを求める．

属性 a_i，b_j の距離 $dis(a_i, b_j)$ は次の式で表される．一致度は関連性が高いと値が大きくなる．また，一致度の最大値は 1 であるため，1 から一致度を引いた値を距離としている．

$$dis(a_i, b_j) = 1 - MatchWR(a_i, b_j)$$

ここで，図 2 の例における a_1 と b_1 の間の変換コスト $cost(a_1, b_1)$ は次の式で算出される．これは a_1 と b_1 の距離に重みを掛けたものである．a_1 と b_1 が持つ重みは同じく 0.3 であるため供給量と需要量が合致し，a_1 からの重みの運送はこの時点で終了する．

$$cost(a_1, b_1) = dis(a_1, b_1) \cdot 0.3$$

同様にコストの計算を行っていき，最終的にすべての運送経路のコストを足し合わせたものが EMD となる．図 2 の例では概念 A，B の EMD は次のように表される．

$$EMD(A, B) = cost(a_1, b_1) + cost(a_2, b_2) + cost(a_2, b_3)$$

以上のような式で算出された EMD の値の最小値を最適化計算で求め，概念間の関連性（意味的な近さ）を算出している．

4 英字略語の意味推定手法

4.1 英字略語の抽出

本論文で提案する英語略語の意味推定手法の処理の対象として扱う英字略語とは，英単語の頭文字から構成される表記とする．例えば，商品の型番や「W 杯」のように記号や数字，日本語などアルファベット以外の文字が混じる表記の場合，それらは英字略語ではないものとする．また，1 文字で構成される英字略語の場合，英単語の頭文字ではなく，例えば，S 字カーブの「S」のように，アルファベットの形状などに起因する意味で使用されることがある．本研究は，語彙の意味に着目し，多義性を有する英字略語の意味を推定することを目的としている．そのため，本論文では，2 文字以上のアルファベットのみで構成されている語を英字略語として扱うこととする．

入力として受け付ける情報は，英字略語が含まれている文章とし，その文章から 2 文字以上のアルファベットの羅列を英字略語として抽出する．

4.2 Wikipedia による意味候補の検索

4.1 節で抽出した英字略語の意味を Wikipedia[10] で検索する．検索の結果，意味が 1 つであった場合には，その意味を出力することになる．意味を複数有する場合には，次節で述べる意味的な近さに基づく多義性の解消を行うため，それぞれの意味を概念化する．概念化には，3.1.1 節で述べた概念ベースと 3.1.2 節で述べた Web を用いた未定義語の概念化手法を用いる．

一例として，英字略語「IC」を Wikipedia で検索した際の結果を表 2 に示す．英字略語「IC」の場合，11 種類の意味を有していた．

なお，Wikipedia に掲載されている意味の中には，商品の型番やある種のコードなど英字略語ではないもの（表 2 の例では 9 番と 10 番）が多数含まれている．また，英字略語の意味を概念化する際に，不要な情報が付加されていることが多い．そこで，Wikipedia に掲載されている意味候補の中から，表 3 に示した規則を上から順に適用した上で意味候補を取得する．ここで，表 3 におけるストップワードは，表 4 に示した通りである．

表 2: 英字略語「IC」を Wikipedia で検索した際の結果.

1. 集積回路 (Integrated Circuit) - 電子機器に用いられる部品。関連：IC カード
2. インタークーラー (Inter Cooler)
3. インターチェンジ (Inter Change) - 道路交通同士が接続するための合分流構造。
4. イメージカラー (Image Color)
5. イオンクロマトグラフィー（Ion Chromatography）の略
6. インフォームド・コンセント (Informed Consent)
7. インターシティ (InterCity) - （特にヨーロッパの）都市間特別急行列車
8. インデックスカタログ - 星団や星雲、銀河を収載した 2 つの星表のこと (Index Catalogue)
9. NHK 富山放送局のラジオ第 2 放送・教育テレビのコールサイン（JOIC/JOIC-DTV）
10. イリノイ・セントラル鉄道の報告記号 (Illinois Central railroad)
11. 間質性膀胱炎 (Interstitial cystitis) の略

表 3: Wikipedia から英字略語の意味候補を取得する規則.

• 意味候補にストップワード（表 4 参照）が含まれる場合，意味候補から除外
• 意味候補内の括弧開き（「），右矢印（→），ハイフン（-）より後ろの語を削除
• 意味候補内の「など」を削除
• 意味候補内の「のこと」,「の略」,「の略符」,「の通称」,「の愛称」, 「の名称」,「の英文略称名」より前の語を取得
• 意味候補内の「の一つ、」より後ろの語を取得

4.3 英字略語の多義性解消

4.2 節で検索した英字略語が複数の意味を有した場合，その多義性を解消する必要がある．具体的には，4.2 節で概念化された意味候補と入力された英字略語を含む文章との意味的な近さを評価することで実現する．この際，概念化された意味候補と英字略語の意味的な近さを評価するため，英字略語も概念化する必要がある．

入力された文章に含まれる自立語をすべて抽出し，それを英字略語の一次属性と見立てる．これにより，英字略語を疑似的に概念化することができる．なお，英字略語の一次属性とした自立語の中には概念ベースに登録されている語と未定義の語が存在する．未定義語については，3.1.2 節で説明した手法によりそれらを概念化する．この処理により，英字略語を概念とし，一次，二次へと属性を展開することができ，意味的な近さを評価することが可能になる．

なお，英字略語を疑似的に概念化する際，概念ベースに登録されていない未定義語である一次属性に対する重みは，3.1.2 節の未定義語の概念化の際の属性への重み付けの考え方と同様にして付与する．

具体的には，語の網羅性である *tf* 値は，入力された文章 A 中に出現する自立語 $Word_A$ の出現頻度 $tfreq(Word_A, A)$ を入力された文章 A 中のすべての自立語の語数 $tnum_A$ で割ったもので算出される．算出式は以下のようになる．

$$tf(Word_A, A) = \frac{tfreq(Word_A, A)}{tnum_A}$$

語の特定性である *idf* 値については，*SWeb-idf* [13, 14] の代わりに *SA-idf* 値を用いる．これは，疑似的な全文章空間の情報としては，*tf* 値を算出する際に使用した文章と同じカテゴリ，ジャンルである文章集合を利用する必要があるためである．*SA-idf* 値の算出式は以下のように定義される．ここで，N はこの文章集合の全文章数，$df(Word_A)$ はその文章集合の中で $Word_A$ が出現する文章数である．

$$SA\text{-}idf(Word_A) = \log \frac{N}{df(Word_A)}$$

以上に示した式より，自立語 $Word_A$ へ付与する重み w は次の式で定義される．

$$w = tf(Word_A, A) \cdot SA\text{-}idf(Word_A)$$

このように，概念化された英字略語と意味候補との意味的な近さを 3.2 節で説明した語彙の意味的な近さを評価する手法により評価する．その結果，最も意味的に近いと判断された意味候補を英字略語の意味とする．

表 4: 設定したストップワードのリスト.

型番，型式，形式，シリーズ，略号，単位，ドメイン，拡張子，記号，符号，係数，コマンド， 国名コード，行政区画コード，県名コード，郵便コード，空港コード，港コード，IATA コード， 航空会社コード，形式コード，通貨コード，言語コード，作品，楽曲，登場，アルバム， コールサイン，一覧，上記，その他，以下

5 評価実験

5.1 実験条件

評価実験のデータとしては，新聞記事から英字略語を無作為に抽出し，当該英字略語が含まれている記事を入力文章とした．使用した記事数は 129 記事であり，その中で，表記が異なる英字略語の数は 58 個であった．つまり，129 種類の英字略語の意味と 58 種類の英字略語の表記が含まれる記事を評価実験データとして使用した．また，当該 58 個の英字略語の表記を Wikipedia で検索した結果，859 個の意味を取得できた（1 つの英字略語の表記につき，平均で 14.8 個，最少で 2 個，最多で 39 個の意味が存在）．提案手法により推定した英字略語の意味が，当該英字略語を含む新聞記事における意味と一致した場合を正答として評価した．

なお，4.3 節で述べたとおり，入力した文章を用いて英字略語を疑似的に概念化する際には，3.1.2 節での考え方に基づき属性に重み付けを行う．今回の実験の場合，入力の対象は新聞記事である．そのため，概念ベースに登録されていない未定義語である一次属性に対する重み付けに必要な *SA-idf* 値の算出には，1 か月分の新聞記事集合を疑似的な全文章空間の情報として用いた．

5.2 評価結果

英字略語の意味推定結果として，関連度計算を利用した場合に 76%，EMD を利用した場合に 79%の正答率を得ることができた．意味推定結果の一例を図 3 に示す．

図 3 において例 1 では，英字略語 AFC に対して，Wikipedia から取得した 13 個の意味候補の中から正しい意味を推定できており，英字略語の意味を十分に理解できていることが分かる．

次に例 2 では，英字略語 FS に対して，25 個の意味候補の中から EMD では正しい意味を推定できた一方，関連度計算では意味の推定に失敗している．これは，関連性の高い属性を 1 対 1 で対応をとって計算を行うために，他に関連性の高い属性を除外してしまう関連度計算に比べ，全ての属性を計算に利用する EMD が有効に機能したためだと考えられる．

最後に例 3 では，英字略語 FB に対して，15 個の意味候補の中から正しい意味を推定することができなかった．これは，入力文章に金銭に関連する語が多く含まれていたため，正しい意味であるフェイスブックより金銭に関連する意味候補である政府短期証券のほうが意味的に近いと判定されたためだと考えられる．

Wikipedia は現存する辞書の中で収録語数が最も多いとされているが，提案手法には，処理の拠り所である辞書に登録されていない単語には対応できないという問題がある．これに関しては，辞書を使用する以上，避けることができない問題である．ただし，評価実験で使用したデータとは異なる 100 件の新聞記事を無作為に調査した結果，Wikipedia に登録されていない英字略語の出現頻度は約 4.5%であった．このことから，新聞記事に登場するような比較的一般的な英字略語を対象とする場合，Wikipedia に未登録の語があることに起因する正答率の低下は 5%程度であると考えられる．

6 まとめ

本論文では，英単語の頭文字から構成される表現である英字略語に焦点を当て，その意味を推定する手法について提案した．提案手法では，語彙の概念化手法と語彙の意味的な近さを評価できる関連性評価手法を使用する．さらに，Wikipedia を辞書として用いることで，英字略語の多義性を解消し，英字略語の本来の意味を推定することを実現した．

提案手法に対して，129 件の新聞記事（英字略語の意味：129 種類，英字略語の表記：58 種類）を入力文章として評価を行った．評価結果より，関連度計算を使用した場合は 76%，EMD を使用した場合は 79%と，最高で 80%近い正答率を得ることができた．

本論文で提案した技術により，英字略語の意味を理解しやすくすることができ，自然な知的対話の実現や情報検索など多くのアプリケーションの性能向上に寄与することができると考えられる．

謝辞

本研究の一部は，科学研究費補助金（若手研究（B）24700215）の補助を受けて行った．

例	入力文章	英字略語	英字略語の意味	意味候補の数	評価手法	推定結果	判定
1	・・・ 国際サッカー連盟理事と **AFC**会長だった2008～11年の間に、 度重なる倫理規定違反が認められたのが理由。 ・・・	AFC	アジア サッカー連盟 （Asian Football Confederation）	13	関連度計算	アジア サッカー連盟	○
					EMD	アジア サッカー連盟	○
2	・・・ 一方、1Q業績は、 前年同期が四半期**FS**を作成していないため 前年同期比較はないが、 経常利益は、6億7000万円。 ・・・	FS	財務諸表 （Financial Statements）	25	関連度計算	Future Systems プロジェクト	×
					EMD	財務諸表	○
3	ソーシャル・ネットワーキング・ サービス世界最大手、米**FB**の株式上場の際、 情報開示が不公正だったとして、 米マサチューセッツ州の証券監督当局は17日、 上場手続きを取り仕切った幹事社の 米証券大手モルガン・スタンレーに 罰金500万ドル（約4億2000万円）の 支払いを命じた。 ・・・	FB	フェイスブック （FaceBook）	15	関連度計算	政府短期証券	×
					EMD	政府短期証券	×

図 3: 英字略語の意味推定結果の一例.

参考文献

[1] 渡部広一, 河岡司: 常識的判断のための概念間の関連度評価モデル, 自然言語処理, Vol. 8, No. 2, pp. 39–54 (2001)

[2] 笠原要, 稲子希望, 加藤恒昭: 単語の属性空間の実現方法, 人工知能学会論文誌, Vol. 17, No. 5, pp. 539–547 (2002)

[3] 河原大輔, 黒橋禎夫: 格フレーム辞書の漸次的自動構築, 自然言語処理, Vol. 9, No. 5, pp. 93–110 (2005)

[4] Eriko, Y., Misako, I., Seiji, T., Hirokazu, W.: Computer-generated Conversation Using Newspaper Headline, *Computer Technology and Application*, Vol. 4, No. 8, pp. 387–394 (2013)

[5] 岡崎直観, 石塚満: 日本語新聞記事からの略語抽出, 第 *21* 回人工知能学会全国大会論文集, 2G4-4, pp. 1–3 (2007)

[6] 吉田辰巳, 遠間雄二, 増山繁, 酒井浩之: 可読性の向上を目的とした片仮名表記外来語の換言知識獲得, 電子情報通信学会論文誌, Vol. J88-D-II, No. 7, pp. 1237–1245 (2005)

[7] 小島一秀, 渡部広一, 河岡司: 連想システムのための概念ベース構成法-属性信頼度の考え方に基づく属性重みの決定, 自然言語処理, Vol. 9, No. 5, pp. 93–110 (2002)

[8] 広瀬幹規, 渡部広一, 河岡司: 概念間ルールと属性としての出現頻度を考慮した概念ベースの自動精錬手法, 信学技報, NLC2001-93, pp. 109–116 (2002)

[9] 奥村紀之, 土屋誠司, 渡部広一, 河岡司: 概念間の関連度計算のための大規模概念ベースの構築, 自然言語処理, Vol. 14, No. 5, pp. 41–64 (2007)

[10] Wikipedia: Wikipedia, *https://jp.wikipedia.org*

[11] 渡部広一, 奥村紀之, 河岡司: 概念の意味属性と共起情報を用いた関連度計算方式, 自然言語処理, Vol. 13, No. 1, pp. 53–74 (2006)

[12] 藤江悠五, 渡部広一, 河岡司: 概念ベースと Earth Mover's Distance を用いた文書検索, 自然言語処理, Vol. 16, No. 3, pp. 25–49 (2009)

[13] 辻泰希, 渡部広一, 河岡司: www を用いた概念ベースにない新概念およびその属性獲得手法, 第 *18* 回人工知能学会全国大会論文集, 2D1-02, pp. 1–4 (2004)

[14] 後藤和人, 土屋誠司, 渡部広一, 河岡司: Web を用いた未知語検索キーワードのシソーラスノードへの割り付け手法, 自然言語処理, Vol. 15, No. 3, pp. 91–113 (2008)

[15] Google: Google, *http://www.google.co.jp*

[16] 徳永健伸: 情報検索と言語処理, 東京大学出版会, (1999)

[17] A.J.Hoffman.: On simple linear programing problems, *In Proc of the Seventh Symposium in Pure Mathmatics of the AMS*, pp. 317–327 (1963)