同志社大学大学院　博士論文

連想メカニズムを活用した
語彙の意味推定法に関する研究

後　藤　和　人

同志社大学大学院　理工学研究科
情報工学専攻

2020 年

目 次

図 目 次

表 目 次

第1章　序論

1.1　研究の背景

近年，情報処理技術の進展に伴い，コンピュータをはじめ様々な機械の高度化・知的化が著しい．これらの発展に共通する未来像の一つは「人間と共存する機械」であると言える．人間と共存する機械の一例としては人工知能ロボットが挙げられる．人工知能ロボットは搭載されたコンピュータがセンサから獲得した情報を処理してアクチュエータを動作させることで，ロボットの身体を駆動する構成になっている．人工知能ロボットは知覚，運動，思考，記憶，学習といった機能から構成されている [1]．上記機能のうち，知覚と運動に関する機能については，身体能力に長けたロボットが数多く開発されてきたことにより，その一部が実現されつつある [2, 3]．しかし，ロボットが人間と共存するためには，当該機能に加え，人間と自然な会話ができる機能が必要である．人間は会話をする際に，意識的または無意識のうちに，様々な常識的な概念（場所，感覚，知覚，感情など）を会話文章から判断し，適切な応答を行っている．自然な会話を実現するためには，思考，記憶，学習に関する機能に基づき，ロボットが人間と同じように，常識的に物事を理解し，応答できることが必要である．

ロボットとの自然な会話を実現するにあたり，重要となる能力の一つが言語を理解する能力といえる．ロボットに言語を理解する機能を与える主要な技術分野が自然言語処理である [4]．自然言語処理の分野において，語（単語）の意味を理解するために，単語の意味を何らかの形で定義した様々な言語資源が構築されている．例えば，単語を意味的に分類したシソーラス [5, 6]，概念（ノード）を関係（リンク）で結ぶ意味ネットワーク [7]，文書中の語群を出現頻度からベクトル化するベクトル空間モデル [8]，if-then 形式の記述で条件文による推論の知識を記述するプロダクション規則 [9] などが存在する．しかし，会話文に上述の言語資源に含まれていない単語（未定義語）が含まれている場合，当該会話文を理解することは困難である．そのため，未定義語が持つ意味を取得できる仕組みが必要となる．

これまでに，未定義語の意味を取得することに関連する方法として，言語資源を活用した多くの研究が行われている．[10] は，言語資源として ISAMAP[11, 12] を利用し，シソーラス上に存在しない単語をシソーラスに分類する方法としてコーパス中の出現回数などの統計情報を用いている．[13] は，言語資源として NTT シソーラス [5] を利用した上で，統計的決定理論の 1 つであるベイズ基準を用いてシソーラスに未登録の単語をシソーラスに分類している．一方，[14] では，検索エンジンの検索ヒット数に対して χ^2 値を用いた関連度の指標を用いることで，シソーラスの自動構築を行う方法が提案されている．[15] は，コーパスにおける単語同士の共起頻度を用いて単語をベクトル表現で表すことで，概念ベースを作成している．そして，概念ベースに登録されていない単語のベクトル表現を，意味空間への射影による方法および分散最小性に基づく方法を用いて推定することで，概念ベースを拡張している．これらの研究は，コーパスなどの言語資源に存在し，かつ，シソーラスや概念ベースに存在しない単語に対して，単

語の共起頻度などを利用することで，当該単語をシソーラスや概念ベースに分類・登録している．そのため，言語資源に登録されていない未定義語を対象とする場合，共起頻度を獲得することができないため，対応できないという問題を抱えている．

また，上述した方法のように未定義語を言語資源に分類・登録するわけではないが，単語（語義）の曖昧性を解消する方法が研究されている．[16] は，Wikipedia における各記事の参照情報であるハイパーリンクを利用することで，一般的な単語と当該単語を含む文章を入力した際に，当該単語の曖昧性を解消し，単語の意味推定を実現している．本方法は品詞情報を利用して意味推定を実施しており，有効に機能する単語が限定（当該論文では曖昧性を有する普通名詞の意味推定を実施）されるという問題がある．他には，文書内の単語を知識ベースのエントリにマッピングすることで曖昧性を解消する技術（Entity Disambiguation）として，[17] は Wikipedia から収集した情報をもとにニューラルネットワークを構築し，入力単語と入力文書のペアとエントリ間の類似性を判断している．しかし，本方法は 2 つのニューラルネットワークのトレーニングおよびメンテナンスが必要であり，対象領域などに応じた適切な運用が要求される．

以上の問題に対処するためには，多くの単語（未定義語）の意味を取得できる方法に加え，当該方法が未定義語の品詞や対象領域に限定されず汎用的に適用できることが必要である．このような未定義語の意味取得方法を実現することで，人間と共存する機械に言語を理解する能力を与え，会話文の意味を理解させることを支援できるようになると考えられる．

1.2　研究の目的

本研究の目的は，人間と共存する機械に与える仕組みの一環として，人間と自然な会話を行うために重要な能力である言語を理解する方法を実現することである．より具体的には，基準となる言語資源に登録されていない単語の意味を提示するシステムを実現することを目指す．本研究では，基準となる言語資源に登録されている単語と登録されていない単語（未定義語）を分けて扱う．これは，単語の意味を推定・取得する際にいくつかの言語資源を活用することになるが，対象となる単語の特徴に応じた言語資源を適用することで，より正確に意味の推定・取得が可能になると考えられるためである．具体的には，本研究では，基準となる言語資源として人間が持つような常識を機械に与えるために構築された概念ベース [18, 19, 20] を使用し，概念ベースに登録されていない単語を未定義語と定義する．概念ベースを使用することで，概念ベースに登録されている単語に対して意味を付与することが可能である．一方，辞書などから構築された概念ベースに登録されていない未定義語については，辞書に掲載されていないような単語に関する情報を入手できる Web 上の情報を活用することで意味推定を行う．なお，基準となる言語資源として概念ベースを使用する理由は，2 章で述べる通り，人間が言語を理解するために活用している連想という能力を実現することを目的として構築されているためである．

本論文では，未定義語の意味推定を 2 段階で行う．1 段階目は人間であれば表記から意味を特定できる未定義語に対して意味推定を行う．1 段階目の意味推定では，未定義語の代表的な存在である固有名詞をターゲットとする．また，システムへの入力は単語（未定義語のみ）とし，単語の関係性を定義するシソーラスを活用することで意味の推定・取得を行う．2 段階目はより難解な対象として，多義性を有する未定義語に対して意味推定を行う．2 段階目の意味推定では，多義性を持つ語として代表的な存在である英字略語をターゲットとする．システムへ

の入力は英字略語を含む文章とし，世界で最も収録語数が多いとされ，かつ，語の曖昧さを回避するための情報を持つ Wikipedia[21] を活用することで意味の推定・取得を行う．上述の能力を機械に持たせることができれば，様々な未定義語の意味を推定・取得することができるようになり，人間と自然な会話を行うために必要となる言語を理解する能力の実現につながると考えている．

未定義語の意味を推定・取得するためには，ある単語から概念を想起し，さらに，その概念に関係のある様々な概念を連想できる能力が重要な役割を果たす．そこで，すでに提案されている連想能力を表現するメカニズムである連想メカニズムを活用して，未定義語の意味を推定・取得する方法を提案し，その有効性を評価する．

1.3 論文の構成

第 2 章では，未定義語の意味推定を行う上で重要となる，ある単語（概念）から様々な概念を連想できる連想メカニズムについて解説する．連想メカニズムは語の概念化方法（単語に属性と重みの集合を与える方法），意味的関連性評価方法，および，シソーラスから構成されている．本章では，語の概念化方法として，電子化国語辞書などから構築した概念ベース，インターネット百科事典である Wikipedia から構築した文書データベース，インターネット上の言語情報を活用するオートフィードバックを用いる方法について述べる．また，意味的関連性評価方法として，概念と概念の関連の強さを定量的に評価できる関連度計算と Earth Mover's distance を説明する．さらに，単語を意味的に分類した分類体系であるシソーラスについても述べる．

第 3 章では，文書データベースを用いて固有名詞の意味推定を行う方法を提案する．提案方法では，文書データベースを用いて固有名詞の概念化を行い，概念化した固有名詞をシソーラスのノードに分類（当該固有名詞と最も関連が高いノードを算出）することで意味推定を実現する．

第 4 章では，文書データベースを用いて英字略語の意味推定を行う方法を提案する．提案方法では，文書データベースを用いて英字略語の概念化を行い，Wikipedia を活用して英字略語の多義性を解消することで意味推定を実現する．

第 5 章では，オートフィードバックを用いて固有名詞の意味推定を行う方法を提案する．提案方法では，オートフィードバックを用いて固有名詞の概念化を行い，概念化した固有名詞をシソーラスのノードに分類することで意味推定を実現する．

第 6 章では，オートフィードバックを用いて英字略語の意味推定を行う方法を提案する．提案方法では，オートフィードバックを用いて英字略語の概念化を行い，Wikipedia を活用して英字略語の多義性を解消することで意味推定を実現する．

第2章　連想メカニズム

2.1　はじめに

情報処理技術における自然言語の意味理解は，整理されたデータ群による意味定義を必要とする方法が一般的であり，整理されたデータ群として様々な言語資源が構築されている．シソーラス [5, 6] は語の語彙的な意味における互いの関係を分類することで木構造による意味定義を提供している．意味的ネットワーク [7] は対象世界に存在する概念と概念を，その間の関係性を示すリンクでつないだ集合体によって意味を定義する．また，ベクトル空間モデル [8] では，語を有限のベクトル空間によって定義している．その際，概念ベクトルに関する情報を国語辞書などから取得し，シソーラス [5, 6] のノードを基とした基底ベクトルを構築したり，特異値分解による潜在的意味インデキシング [22] によって基底次元を削減したりすることで，語を表現している．このように定義した語に関する知識をもとに，ベクトル余弦に基づく語と語の関連性評価を行う [23] ことで，発音・表記などは異なるが同じ意味を持つ同義性や，類似した意味を持つ類義性を識別し，語と語の関係を定量的に扱うことが可能となる．

しかし，人間は語彙的な意味のみに依存するわけではなく，また分類・体系付けられた構造では表現しきれない自然言語の意味や関連を理解することができるため，一般的な会話においては，同義性や類義性にとらわれることなく語と語の関連の強さを評価することが求められる．例えば,「赤ちゃん」という語から「子供」という語を導くことは容易であるが，人間は「おもちゃ」のような「赤ちゃん」という語からは同義性や類義性の観点からは関連を見出すことが困難な語も導くことができる．これは，人間が「おもちゃは赤ちゃんが使用する道具」と認識しているために,「赤ちゃん」から「おもちゃ」を関連する語として導くことを可能にしている．

このように，曖昧さや柔軟さを持って語の意味を理解することを，人間は「連想」という能力で実現していると考える．連想とは，自身が知っている事柄，情報，概念といった多種多様な「知識」から他の知識を関連付けることであると定義する．この知識とは，人間が持つ「常識」と言い換えることができると考える．人間が「赤ちゃん」と「おもちゃ」の間に関連を見出すことができるのは，これらの語が互いを連想させるからである．人間が持つ連想能力を機械（コンピュータ）上で実現するためには，同義性や類義性を考慮した語彙的な意味や分類（体系）付けられた構造にも依存しない曖昧な関係性，つまり，人間が持つ常識に基づく関係性を表現することが必要となる．しかし，上述した同義性や類似性あるいは 1.1 節で述べた共起情報に基づいて語を表現する方法で連想能力を実現することは困難と考える．そこで，本章では，人間が行っているような常識に基づく連想と意味の理解をコンピュータで表現することを目的としている連想メカニズムを説明する．

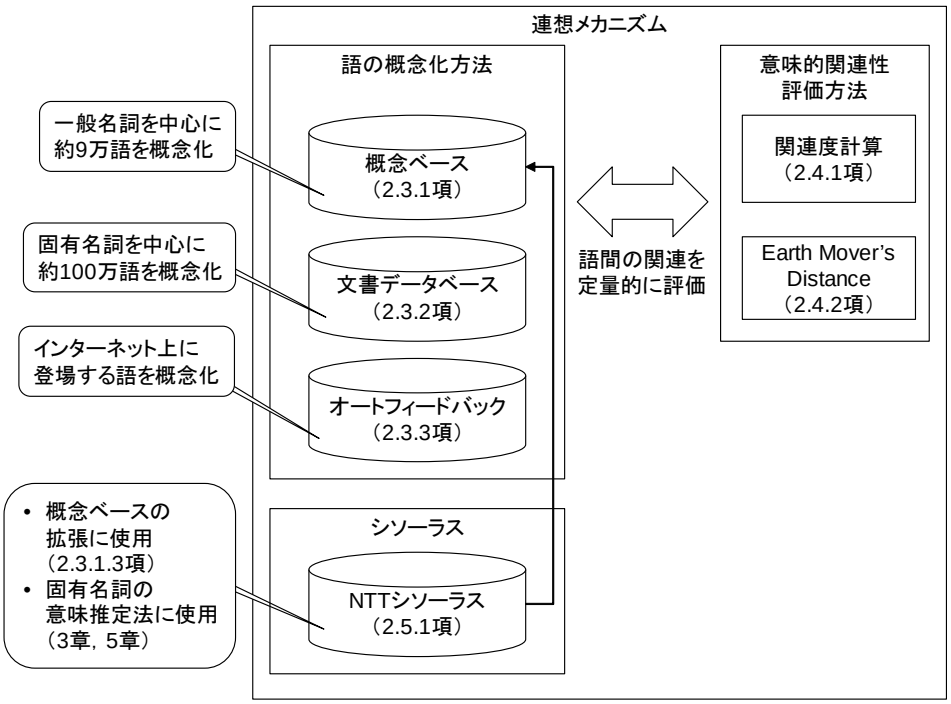


図 2.1: 連想メカニズム

2.2　連想メカニズムの構成

連想メカニズムは，コンピュータに連想機能を持たせるために活用する語の概念化方法と，語と語の関連の強さを定量的に評価することを可能とする意味的関連性評価方法，および，単語を意味的に分類した分類体系であるシソーラスから構成されている．連想メカニズムの構成を図 2.1 に示す．

語の概念化方法として，電子化国語辞書などから構築した概念ベースを用いる方法，インターネット百科事典である Wikipedia から構築した文書データベースを用いる方法，インターネット上の言語情報を活用するオートフィードバックを用いる方法について述べる．各概念化方法の大まかな使い分けは以下の通りである．

まず，電子化国語辞書などに基づいて構築されているために品質が高い概念ベースを用いて，一般的に使用される語（普通名詞など）の概念化を実施する．概念ベースには，一般名詞を中心に約 9 万語が概念として登録されており，その精度は 80%を超えている（2.3.1 項参照）．次に，インターネット百科事典である Wikipedia を基に構築されているために広範な語に対応可能な文書データベースを用いて，概念ベースでは概念化できなかった未定義語（固有名詞など）の概念化を実施する．文書データベースには固有名詞を中心に約 100 万語が概念として登録されており，その精度は 70%を超えている（2.3.2 項参照）．最後に，インターネット情報を利用するために品質はやや低いもののほぼあらゆる語に対応可能なオートフィードバックを用いて，

表 2.1: 概念と属性の例

概念	属性
赤ちゃん	(ベビー, 0.12), (赤子, 0.11), (子供, 0.01)，…
子供	(子沢山, 0.12), (児戯, 0.10), (稚児, 0.09)，…
おもちゃ	(おしゃぶり, 0.20), (玩具, 0.07), (子供, 0.01)，…
…	…

文書データベースでも概念化できなかった未定義語の概念化を実施する．オートフィードバックはインターネット上に登場する語であれば概念化することが可能であり，その精度は 70%程度である（2.3.3 項参照）

意味的関連性評価方法は，ある語から連想される他の語を想起することや，語と語の間の関連を定量的に評価することで意味の近さを測ることを可能とする方法である，当該方法を用いることで，ある語を入力するとその語から人間が連想できるであろう他の語を出力することや，2 つの語を入力することで語間の関連を定量的に表現した値を出力することができる．本章では，意味的関連性評価方法として，関連度計算と Earth Mover's Distance について説明する．

NTT シソーラスは，語を意味的に分類した分類体系であるシソーラスの一つである．本論文では，概念ベースの情報源を拡張し，精度を改善するために NTT シソーラスを使用する．また，本論文で提案する固有名詞の意味推定法においても NTT シソーラスを使用する．

2.3 語の概念化方法

2.3.1 概念ベースを用いた概念化

2.3.1.1 概要

概念ベース [18, 19, 20] は，概念（見出し語）とその特徴を表す複数の語（属性）を対の組として集めて構築されたデータベースである．概念及び属性は，主に国語辞書，新聞記事，シソーラスから取得される．

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．

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

概念定義の一例を表 2.1 に示す．

概念ベースの大きな特徴は，属性である単語は概念として必ず定義されている構造を持つことである．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．上述の構造を表 2.1 に示した例で表現すると図 2.2 のようになる．

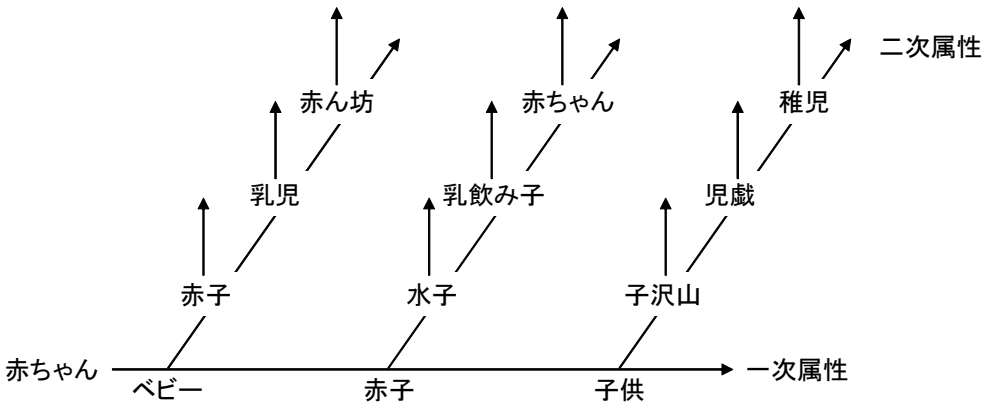


図 2.2: 概念「赤ちゃん」を二次属性まで展開した場合の例

表 2.1 と図 2.2 に示した概念と属性の例が示すように，概念ベースにおける概念の定義は，その概念を示す語の語彙的な意味に留まらない．概念との明確な関係性を定義できない語であっても，人間が何かしらの関連を見いだせる場合は属性として付与する．このような特徴を持つ概念ベースによって，人間らしい連想を実現する可能性を持つ連想メカニズムを構築することができる．

2.3.1.2　概念ベースの精度評価方法

概念ベースの構築方法を説明する前に，構築した概念ベースの精度評価方法を説明する．概念ベースの精度評価には，X-ABC 評価セット [20] を用いる．X-ABC 評価セットは，ある基準概念 X と，概念 X と関連が非常に強い概念 A，概念 A ほどではないが関連があると思われる概念 B，全く関連のない概念 C によって構成された評価セットであり，人手によって作成されている．概念 X，A，B，C を 1 組として，500 組が評価セットとして存在している．表 2.2 に評価セットの例を示す．

概念 X と概念 A の関連度を $DoA(X, A)$，概念 X と概念 B の関連度を $DoA(X, B)$，概念 X と概念 C の関連度を $DoA(X, C)$ とする．そして，評価セットにおける $DoA(X, C)$ の平均値を $AveDoA(X, C)$ とし，各概念間の関連度を比較することで概念ベースを評価する．なお，関連度については 2.4.1 項で説明する．当該評価セットを用いた評価方法として，順序正解率と C 平均順序正解率について以下に述べる．

＜順序正解率＞

評価セットにおける各概念間の関連度に以下の関係がある場合を正解と判定する．

$$DoA(X, A) > DoA(X, B) > DoA(X, C)$$

この評価を全ての組に対して実施した上で，正解となった組の比率を概念ベースの精度とする．

表 2.2: X-ABC 評価セットの例

概念 X	概念 A	概念 B	概念 C
飲食店	食堂	米	青空
仲買	仲介	市場	仕舞う
沿岸	海岸	船	練乳
ご飯	飯	米	青空
意図	志向	内心	帰宅
羽	翼	鳥	返還
延期	順延	日程	関連
演算	計算	処理	芋
…	…	…	…

＜C 平均順序正解率＞

概念 X と関連がない概念 C の関連度 $DoA(X, C)$ は，本来 0.0 となることが理想である．しかし，関連度の算出方法の特性上，概念 X と概念 C に 1 つでも共通した属性が存在すれば，微小な値が算出されることになる．そこで，$DoA(X, C)$ を誤差とみなし，その平均 $AveDoA(X, C)$ を評価セット全体での平均誤差とする．そして，$DoA(X, A)$，$DoA(X, B)$，$DoA(X, C)$ に平均誤差以上の差が存在している場合に，人間の常識に沿った関連度が算出されているとして正解と判定する．具体的には，以下の式を満たす場合に正解とみなす．

$$DoA(X, A) - DoA(X, B) > AveDoA(X, C)$$
$$DoA(X, B) - DoA(X, C) > AveDoA(X, C)$$
$$AveDoA(X, C) = \frac{\sum_{i=1}^{set_{num}} DoA(X_i, C_i)}{set_{num}}$$

この評価を全ての組に対して実施した上で，正解となった組の比率を概念ベースの精度とする．本論文では，C 平均順序正解率における評価セット数 set_{num} は 500 となっている．

2.3.1.3 構築方法と性能評価

本項では，概念ベースの構築方法及び構築した概念ベースの性能評価について述べる．

具体的には，概念と属性の取得，属性の精錬，重みの付与，属性追加といった概念ベースを構築するための各種処理について示し，後述する方法を用いて構築した概念ベースの精度を算出する．

以降，概念ベースを構築する方法として以下を説明する．

- 国語辞書を用いた基本概念ベースの構築（基本概念ベース）
- ルールを用いた属性精錬（ルール精錬概念ベース）

- 新聞記事を用いたルール精錬概念ベースの拡張（新聞概念ベース）
- シソーラスを用いた属性追加（シソーラス概念ベース）

＜国語辞書を用いた基本概念ベースの構築（基本概念ベース）＞
○概要
国語辞書から概念と属性を取得することで構築される基本概念ベースの構築方法 [24, 18] について述べる．
この概念ベースでは，国語辞書 [25, 26, 27, 28, 29, 30] の各見出し語を概念と定義し，各見出し語の語義文中における語を属性候補として抽出する．次に，当該属性候補に対して後述する属性信頼度という指標による選別を行い，適切と判定した属性のみを残す．残した属性に対して，学習データによる重みを付与することで基本概念ベースは構築される．

○構築方法（概念と属性の抽出）
国語辞書には，見出し語とその意味を説明する語義文が存在する．まず，この見出し語を概念とみなし，概念を説明する語義文中の自立語を属性として抽出する．各属性には出現頻度による重みを付与する．
次に,「属性が持つ属性」及び「概念を属性として持つ概念」を新たな属性として追加する．例えば，概念「馬」を定義する際には，国語辞書の見出し語「馬」が持つ語義文から属性として「家畜，たてがみ，… 」といった語を取得することができる．同様に，国語辞書中の見出し語を概念化すると，概念「馬」の属性として得られた「家畜」も概念として定義される可能性がある．概念「家畜」も語義文から属性が抽出されるため，抽出された属性も概念「馬」の属性として使用する．上記処理が「属性が持つ属性」による属性の追加である．また,「概念を属性として持つ概念」による属性の追加では,「馬」という語を属性として持つ概念，例えば，概念「競馬」が属性「馬」を持つ場合，概念「馬」の属性として「競馬」も使用する．
最後に,「概念化されていない属性」及び「全ての概念に出現する属性」を削除する.「概念化されていない属性」は，語義文中には出現したが見出し語としては国語辞書に存在しない語を指す.「全ての概念に出現する属性」は日本語特有の言い回しに由来する語を指し，例えば,「こと」や「もの」といった語である．上記処理による属性の削除を実施した後，所属属性数が閾値より少ない概念を削除する．
以上の処理により，国語辞書からの概念及び属性の抽出が完了する．

○構築方法（属性信頼度の算出）
属性信頼度とは，得られた属性が概念にとってどれぐらい信頼できる情報であるかを表す値であり，人手によるサンプル評価を用いて算出される [18]．属性信頼度によって，各属性をクラス分けし，重み付けを行う．属性信頼度の算出は，概念と属性の関連を見出すために使用する 6 つの手がかりと，ランダムに選択したサンプル概念 100 語の目視評価結果を組み合わせることで行う．
まず，サンプル概念が持つ属性を「適切，どちらでもない，不適切」の 3 段階に分けて評価する．当該評価は，3 名の目視によって実施され，3 名全員が不適切と判断しなかった属性を「適切属性」とする．

次に，6 つの手がかりについて，各手がかりが合致したときに，属性がどれくらい信用できるかを表す属性信頼度を設定する．各手がかりと属性信頼度は以下の通りである．

手がかり（1）：概念と属性の一致
概念と属性の表記が完全に一致している場合，手がかり（1）に合致すると判定する．この場合，概念と属性の関連は疑いようがないため，属性信頼度は 100%となる．

手がかり（2）：語関係データに定義される概念と属性の関係
国語辞書から構築された語の関係を示すデータにおいて，概念と属性が同義，類義，上位下位のいずれかの関係にある場合，手がかり（2）に合致すると判定する．この場合，概念と属性の関係が明確に定義されているため，属性信頼度は 100%とする．

手がかり（3）：出現頻度によって付与された属性の重み
現時点では，基本概念ベースにおける属性の重みは国語辞書中の出現頻度によって与えられている．手がかり（3）では，ある概念において重み（出現頻度）が大きい属性ほど信頼できる属性である可能性が高いと判定する．[18] では，例えば，重みが 0.09 以上の属性のうち，適切な属性の割合は約 80%と示されており，重みが 0.09 以上の属性の属性信頼度は 80%とする．

手がかり（4）：概念と属性の関連度
手がかり（4）では，概念と属性の関連度が高いほど信頼できる属性である可能性が高いと判定する．[18] では，例えば，関連度が 0.1 の属性のうち，適切な属性の割合は約 50%と示されており，関連度が 0.1 の属性の属性信頼度は 50%とする．

手がかり（5）：概念と属性の漢字一致
概念と属性の表記において漢字の一部が一致している場合，手がかり（5）に合致すると判定する．[18] では，概念と属性の表記において漢字の一部が一致している属性のうち，適切な属性の割合は 73%と示されており，この値が当該属性の属性信頼度となる．

手がかり（6）：相互属性
概念 A の属性 a_i を概念として見たとき，a_i の属性に A が存在する場合，属性 a_i を概念 A の相互属性と呼ぶ．手がかり（6）では，概念 a_i の属性 A の重みを用いて，手がかり（3）による属性信頼度を用いることができる．

なお，複数の手がかりに該当する属性の属性信頼度は，独立事象の確率合成とみなして算出する．ある属性 a_i に対して，手がかり（2），（3）から属性信頼度 p_2，p_3 が得られた場合，属性 a_i の属性信頼度 P は以下の式から求められる．

$$P = \frac{p_2 p_3}{p_2 p_3 + (1 - p_2)(1 - p_3)}$$

○構築方法（属性信頼度による重み付与）

表 2.3: 属性信頼度によるクラス分け

クラス	属性信頼度（%）
信頼度 1	100
信頼度 2	80 以上 100 未満
信頼度 3	60 以上 80 未満
信頼度 4	40 以上 60 未満
信頼度 5	20 以上 40 未満
信頼度 6	0 以上 20 未満

ある概念における属性の属性信頼度を算出し，算出した属性信頼度に従って属性のクラス分けを行う．各クラスの条件を表 2.3 に示す．

表 2.3 に示したクラスごとに，重みの付与を実験的に行う．信頼度 2 の属性を基準として 1.0 の重みを付与する．信頼度 1 の属性には概念との関係（同義・類義・上位下位）に応じて 1.0〜16.0 の重みを付与するパターンを用意し，信頼度 3〜6 の属性には 0.0〜1.0 の重みを付与するパターンを用意し，各パターンを組み合わせることで実験を行う．2.3.1.2 項に示した X-ABC 評価セットを重み学習用に作成し，実験の結果，当該セットにおいて C 平均順序正解率が最も高くなったパターンを属性の重みに採用する．

以上の処理によって，国語辞書を用いた基本概念ベースの構築が完了する．

○性能評価

2.3.1.2 項で示した X-ABC 評価セットを用いて基本概念ベースの精度評価を実施した．評価結果を図 2.3 に示す．なお，比較対象としてシソーラス距離 [31] を用いた類似度計算を適用した場合の精度も示している．図 2.3 より，基本概念ベースはシソーラス距離と比較して，順序正解率と C 平均順序正解率の両方で精度を 30%以上改善できていることが分かる．また，[32] では，基本概念ベース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ことが報告されている．

人手で作成された評価セットに対して高い精度を獲得できていることから，概念ベースの構造は人間らしい関連性の判断に適していることが示されている．なお，構築された基本概念ベースは，概念数が約 3.4 万語，1 概念あたりの平均属性数が約 45 個となっている．

＜ルールを用いた属性精錬（ルール精錬概念ベース）＞

○概要

基本概念ベースの属性は国語辞書の語義文に出現する語から得られる語群であるが，属性として採用するか否かの選別は辞書中の出現頻度のみによって行われている．この属性の選別にルールを定めることで，属性の精錬を行う方法 [33, 19] について述べる．

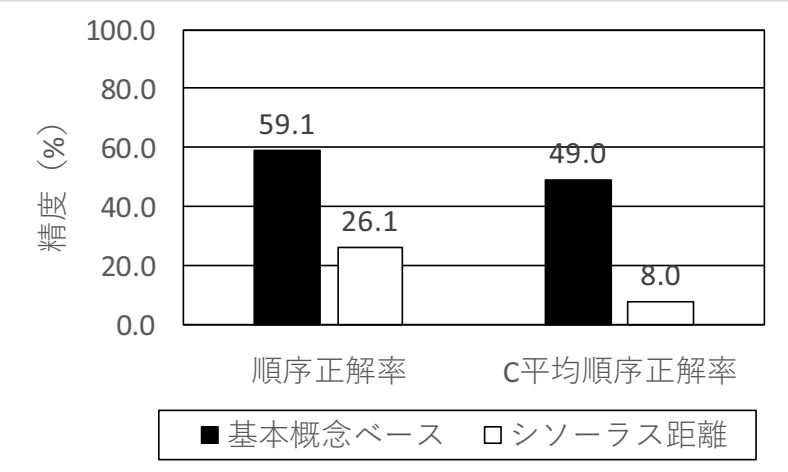


図 2.3: 基本概念ベースの評価結果

○構築方法

基本概念ベースに対して 4 つのルール群による精錬を行うことで精錬用概念ベースを作成する．この精錬用概念ベースにより算出される関連度とルール群を組み合わせることで属性のランク分けを行い，属性の選別を行う．ルールは基本概念ベースの構築において使用した属性信頼度を算出する際に活用した手がかりのうち，語関係データに定義される概念と属性の関係（手がかり（2)），概念と属性の漢字一致（手がかり（5)），相互属性（手がかり（6)）を利用する．さらに,「シソーラスにおいて概念と属性が上位・下位・仲間の関係にある」というルールを加え，これら 4 つのルールのどれにも適合しなかった属性を削除する．この処理により，作成された概念ベースを精錬用概念ベースとして利用することで，基本概念ベースの属性を選別する．

続いて，作成した精錬用概念ベースを利用して，基本概念ベースに定義される概念と属性の関連度を算出する．算出した関連度と前述の 4 つのルールを用いて，基本概念ベースにおける属性のランク分けを行う．具体的には，各ルールに適合する属性と概念の関連度を算出し，当該関連度が閾値以上である場合に属性として適切と判断する．また,「関連度」そのものもルールに加え，ある概念 A の属性 a_i の関連度が閾値以上であった場合に属性として適切と判断する．なお，閾値はルール毎に異なる値をとり，値は実験評価により算出している．

図 2.4 に，属性のランク分けの流れを示す．図 2.4 において，$Judge_1\_High$ と $Judge_1\_Low$ は，4 つのルールと「関連度」のルールのうち，いずれかのルールに適合した属性の選別を行うための閾値である．前述した通り，適合したルールによって閾値は異なる．$Judge_1\_High$ は，各ルールに適合した属性のうち，適切と判断できた属性の数が 8 割以上となるように設定した閾値である．また，$Judge_1\_Low$ は，各ルールに適合した属性のうち，適切と判断できた属性の数が 6 割以上となるように設定した閾値である．そして，N は属性が適合したルールの数を示す．つまり，概念との関連度が $Judge_1\_High$ 以上であり，かつ，2 つ以上のルールに適合した属性はランク A_1 に分類される．また，概念との関連度が $Judge_1\_High$ 以上であり，かつ，1 つのルールに適合した属性はランク A_2 に分類される．概念との関連度が $Judge_1\_Low$ 以上

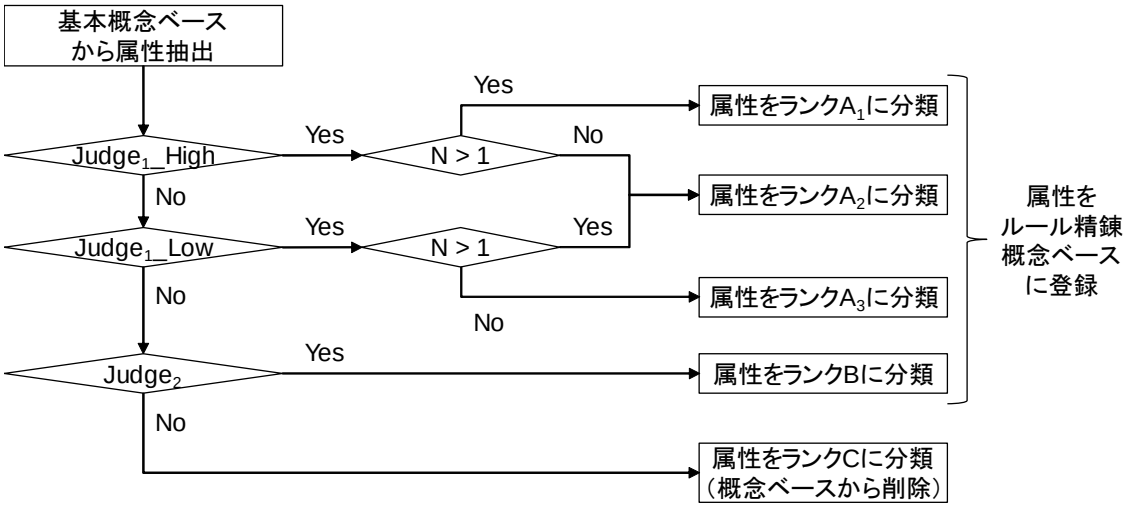


図 2.4: ルールを用いた属性のランク分け

であり，かつ，2 つ以上のルールに適合した属性もランク A_2 に分類される．概念との関連度が $Judge_1\_Low$ 以上であり，かつ，1 つのルールに適合した属性はランク A_3 に分類される．さらに，$Judge_2$ は，関連度を用いた追加のルールであり，「概念 A と属性 a_i の関連度が閾値より高く，属性 a_i 以外の属性との関連度と比べて十分高い」場合に，当該属性をランク B に分類する．当該ルールの閾値は，適切と判断できた属性の数が 3 割以上となる値に設定している．上記全てのルールに適合しなかった属性はランク C に分類され，当該属性は概念ベースから削除される．以上のようなランク分けを行うことで，ルール精錬概念ベースは構築される．

○性能評価

2.3.1.2 項で示した X-ABC 評価セットを用いてルール精錬概念ベースの精度評価を実施した．評価結果を図 2.5 に示す．図 2.5 より，ルール精錬概念ベースは基本概念ベースと比較して，順序正解率と C 平均順序正解率の両方で精度を 5%前後改善できていることが分かる．ルール精錬概念ベースは，概念数は基本概念ベースと同じであるが，1 概念あたりの平均属性数が約 29 個となり，基本概念ベースと比較して約 35%の属性が削除されている．

＜新聞記事を用いたルール精錬概念ベースの拡張（新聞概念ベース）＞

○概要

ルール精錬概念ベースを構築する際に利用した国語辞書には，見出し語が約 20 万語登録されている．しかし，属性の取得範囲が語義文のみであるため，適切に属性が取得できない概念が存在し，結果として概念数は約 3.4 万語に留まっている．そのため，人間が自然と考え付く語句によって構成される X-ABC 評価セットに含まれる語にも，概念化されていない（概念ベースに登録されていない）語が存在する．なお，X-ABC 評価セットに含まれる 2000 語のうち，167 語が基本概念ベース及びルール精錬概念ベースに登録されていない語となっている．これ

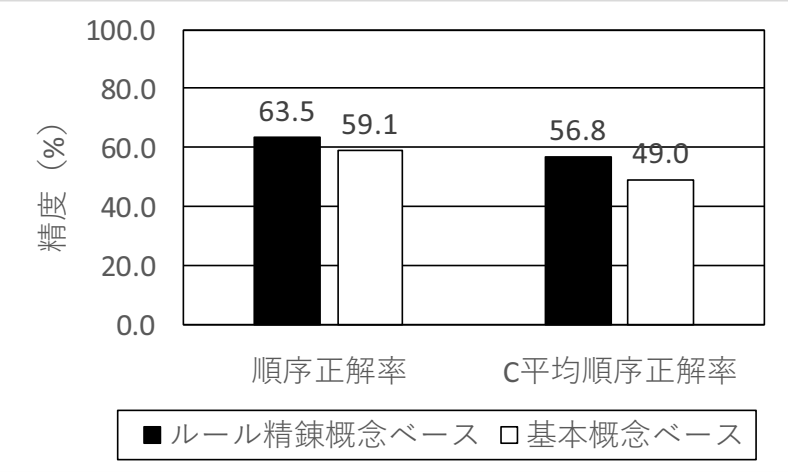


図 2.5: ルール精錬概念ベースの評価結果

は，人間が日常的に扱う語の意味知識を定義できていないということになり，知識が不足していると言える．

そこで，概念ベースを構築する情報源として新聞を加えることでルール精錬概念ベースの拡張 [34] を行う．新聞記事中には日常的に人間が用いる語が頻出しており，記事中に互いに共起する語を概念と属性の関係に見立てることで，連想を実現するために何かしら関連がある語を取得することが期待できる．

○構築方法（概念と属性の抽出）

情報源として新聞記事を用いることで，国語辞書のみからでは属性が付与されず，概念として採用されなかった見出し語に対しても概念化できる可能性がある．そこで，まず，国語辞書に記載されている約 20 万語の見出し語のうち，概念として不適切な表記を除去した約 12 万語を抽出する．ここでは，概念は単独で何かしらの意味を持つ必要があるため，名詞・動詞・形容詞・形容動詞である見出し語を抽出している．抽出された見出し語には，すでに基本概念ベースに登録されている語も含まれているが，新聞記事からの属性取得は基本概念ベースに登録されているか否かを区別せずに行う．つまり，基本概念ベースに登録されている見出し語（概念）に関しては，新聞記事からさらに属性が追加されることになる．なお，当該抽出において，除去された見出し語は,「」による表記の省略や「＜＞」による表記の使用例,「・」による表記の並列などを表す見出し語である．新聞記事には，国語辞書のように「見出し語-語義文」という明確な関係がない．このため，新聞記事から「概念-属性」の関係を抽出するためには，新聞記事内での語と語の共起関係を手がかりとする方法 [19, 20] が有効であると考える．本方法では，ある一定の記事範囲において同時に出現（共起）する語同士には何かしらの関係があるとみなし，それらを属性として取得する．共起を判別する記事範囲は句読点に区切られた範囲とし，当該範囲で互いに共起する語同士をそれぞれ概念と属性の関係とみなす．新聞記事における共起の例を図 2.6 に示す．図 2.6 では，例えば「大学が国立研究所など外部の研究機関に大

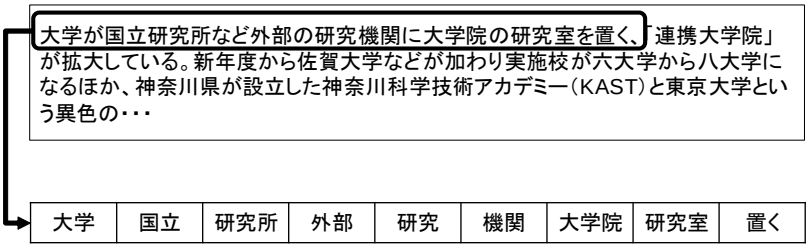


図 2.6: 新聞記事における共起の例

表 2.4: 共起情報を用いた「概念-属性」の関係の例

概念	属性 1	属性 2	属性 3	属性 4	属性 5	属性 6	属性 7	属性 8
大学	国立	研究所	外部	研究	機関	大学院	研究室	置く
国立	大学	研究所	外部	研究	機関	大学院	研究室	置く
研究所	大学	国立	外部	研究	機関	大学院	研究室	置く
外部	大学	国立	研究所	研究	機関	大学院	研究室	置く
研究	大学	国立	研究所	外部	機関	大学院	研究室	置く
機関	大学	国立	研究所	外部	研究	大学院	研究室	置く
大学院	大学	国立	研究所	外部	研究	機関	研究室	置く
研究室	大学	国立	研究所	外部	研究	機関	大学院	置く
置く	大学	国立	研究所	外部	研究	機関	大学院	研究室

学院の研究室を置く」が共起する語を抽出する範囲となる．当該範囲中に存在する名詞・動詞・形容詞・形容動詞を形態素解析により抽出することで,「大学，国立，研究所，… 」が共起している語として取得される．

取得された単語 A，単語 B，単語 C，… の間には，単語 A を概念とした場合，その属性として共起している単語 B，C，… が付与されるという関係にある．図 2.6 の例では，表 2.4 に示した「概念-属性」の関係が取得される．

上述の処理により，各概念に対して取得された平均属性数は約 100 語である．

○構築方法（出現頻度による重みの付与）

新聞記事から獲得した属性には重みが与えられていない．そこで，共起回数から擬似的に重みを付与する．

まず，共起情報を用いて獲得した属性に対して，概念との共起回数を算出する．共起回数は新聞記事全体を句読点で区切った範囲毎に概念と属性が同時に出現する回数を数える．本処理により，ある概念の一次属性の共起回数の合計は，当該概念の新聞記事全体における出現頻度と等しくなる．共起回数を付与した概念ベースのイメージを図 2.7 に示す．

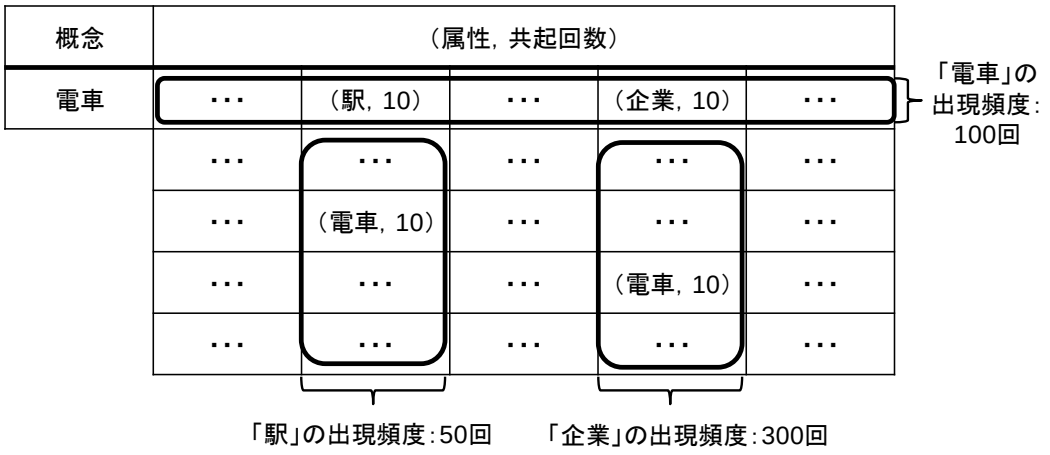


図 2.7: 共起回数を付与した概念ベース（イメージ）

図 2.7 は，概念「電車」の属性である「駅」と「企業」を示している．属性「駅」と「企業」は共に概念「電車」との共起回数が 10 回である．そのため，共起回数をそのまま重みに適用した場合，両属性の価値は等しくなる．ここで，属性「駅」を概念とみなして一次属性を取得した結果,「駅」の出現頻度が 50 回だったとする．同様に属性「企業」の出現頻度が 300 回だったとする．この場合，概念「電車」にとって，記事全体に対して出現頻度が少ない「駅」という語が持つ 10 回の共起と，出現頻度が多い「企業」という語が持つ 10 回の共起が同じ価値を持つことは妥当ではない．そこで，属性に対する擬似重み付与方法として，相互情報量を用いる．概念 A の属性 a_i に対して相互情報量を考慮した擬似重み $W_a np$ は以下の式で定義される．

$$W_a np = \log \frac{q_{Aa}}{\sum_k q_{Ak} + \sum_k q_{ak} - q_{ak}}$$

ここで，q_{Aa} は概念 A と属性 a の共起回数である．$\sum_k q_{Ak}$ は概念 A の一次属性の共起回数の合計，つまり，概念 A の出現頻度である．$\sum_k q_{ak}$ は属性 a を概念とみなした際の一次属性の共起回数の合計，つまり，概念 a の出現頻度である．概念 A の一次属性には属性 a が，概念 a の一次属性には属性 A が存在するため，共起回数が重複して加算されることを防ぐために，上記式の分母において，q_{ak} を減算している．以上の処理によって算出した擬似重みを属性に付与する．

○構築方法（ルールによる属性精錬と概念ベース idf による重みの付与）

算出した擬似重みを用いて関連度を算出することで，ルール精錬概念ベースを構築する際に実施した属性精錬のランク分けを行うことが可能になる．さらに，属性の精錬によって得られる属性のランクを利用して重みを付与する．

まず，基本概念ベースを構築した際に使用したサンプル概念に対してルールによる属性のランク分けを行い，各ランクに分類された属性に対して適切属性の割合を調査する．表 2.5 に適切属性の割合を調査した結果を示す．

次に，新聞記事から得られた属性も同じようにランク分けを行う．各属性の重みとして，表

表 2.5: ランク毎の適切属性の割合

ランク	適切属性の割合
$A1$	0.84
$A2$	0.74
$A3$	0.57
B	0.33
C	0.13

2.5 に示したランク毎に得られた適切属性の割合を付与する．ルール精錬概念ベースを構築した際は，ランク C に分類された属性を不適合属性として削除したが，新聞記事を用いたルール精錬概念ベースの拡張では概念と属性の拡充が目的であるため，削除は行わず重みを小さくすることとする．

続いて，概念ベース *idf* の算出を行う．*idf* とは逆文書頻度のことであり，一般的には「様々な文書が存在する空間において，特定の文書にしか出現しない語は重要である」ことを示す指標である．*idf* は，空間中の総文書数を，語が出現する文書数で割った値の対数によって表され，一般に以下の式で算出される．ここで，N は対象とする文書空間における総文書数，$df(t)$ は語 t が出現する文書数である．

$$idf(t) = \log \frac{N}{df(t)}$$

この考えを概念ベースに対応づけたものが概念ベース *idf* である．概念ベースを仮想的な文書空間と捉え，概念ベース上の *idf* を概念ベース *idf* として算出する．つまり，多数の概念の属性として付与されているような属性は参照頻度が高く，各概念を特徴付ける上で価値が低いと考える．逆に，少数の概念にしか属性として付与されていない属性は各概念において価値が高いものとみなす．

概念ベース *idf* では，空間中の 1 文書を 1 概念が持つ属性群と考える．よって，概念ベースに定義されている概念の総数が全文書数となる．ゆえに，ある概念 X の概念ベース *idf* は，全概念数を概念 X が属性として出現する概念の数で割った値の対数によって表され，以下の式で算出される．

$$CV_N(X) = \log_2 \frac{V_{all}}{df_N(X)}$$

$CV_N(X)$ は概念ベースを N 次属性まで属性を展開した場合における概念 X の概念ベース *idf* である．V_{all} は概念ベースに定義されている全概念数，$df_N(X)$ は N 次属性集合内において概念 X を属性として持つ概念の数である．概念ベースでは概念の持つ属性の範囲を N 次展開によって広げることができるが，新聞記事を用いたルール精錬概念ベースの拡張では 1 概念が持つ属性群の範囲を二次属性まで展開することで概念ベース *idf* を算出している．なお，一般に *idf* を算出する際の抑制関数として常用対数が用いられる．これは情報検索やテキストマイニングなどで対象とする文書数は数億から数十億と大規模な文書を情報源としているためである．

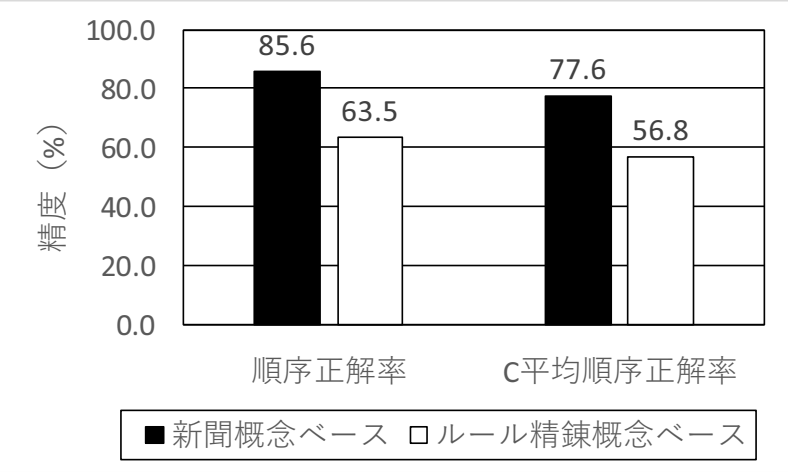


図 2.8: 新聞概念ベースの評価結果

一方，概念ベースは数万語程度の仮想的な文書集合であるため，常用対数を用いると抑制幅が大きくなり値の分布が高密度に圧縮されてしまうことを考慮し，二進対数を使用している．

これまでに述べてきた各属性に対して算出したランク分けによって付与された重みと概念ベース idf を掛け合わせた値を属性の新たな重みとする．以上の処理によって新聞記事を用いたルール精錬概念ベースの拡張が行われる．

○性能評価

2.3.1.2 項で示した X-ABC 評価セットを用いて新聞概念ベースの精度評価を実施した．評価結果を図 2.8 に示す．図 2.8 より，新聞概念ベースはルール精錬概念ベースと比較して，順序正解率と C 平均順序正解率の両方で精度を 20%以上改善できていることが分かる．また，構築された新聞概念ベースは，概念数が約 8.8 万語，1 概念あたりの平均属性数は約 140 語となっている．なお，基本概念ベースとルール精錬概念ベースにおいて定義されていなかった X-ABC 評価セットにおける 167 語についても全て概念化することに成功している．

○サンプル概念を用いない重み付け方法とその性能評価

これまでに述べてきた概念ベースの構築における属性の重み付けには，人手によるサンプル概念の評価結果を使用していた．しかし，新聞記事などの新しい情報源を用いることで概念ベースの規模が大きくなると，数が少ないサンプル概念の評価結果に依存した重み付けでは適切な結果を得ることが難しくなる．そこで，サンプル概念に依存しない重み付け方法として，関連度と概念ベース idf を用いた重み付けを行う．

属性の重み付けにおいて，サンプル概念は，表 2.5 に示したランク毎の適切属性の割合を算出する際に使用されている．ここでは，ランク毎の適切属性の割合に変わり，概念と属性の関連度を用いる．つまり，ある概念 A が持つ属性 a_i の新たな重み w_i は，概念 A と属性 a_i の関連度に属性 a_i の概念ベース idf を掛け合わせた値となり，以下の式から算出される．なお，$DoA(A, a_i)$

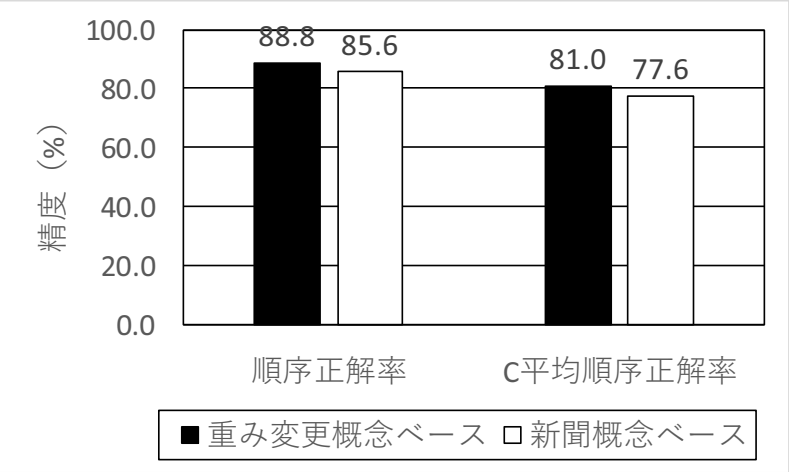


図 2.9: 重み変更概念ベースの評価結果

が概念 A と属性 a_i の関連度，$CV_N(a_i)$ が属性 a_i の概念ベース idf である．

$$w_i = DoA(A, a_i) \times CV_N(a_i)$$

この概念と属性の関連度を用いた重み付け方法を新聞概念ベースに適用し，性能評価を実施した．概念ベース idf における属性の範囲を表す N については，$N = 1 \sim 4$ までを実験に評価した結果，$N = 3$ の場合に最も高い精度が得られたため，$N = 3$ として評価を行っている．評価結果を図 2.9 に示す．なお，当該重み付け方法を適用した概念ベースを重み変更概念ベースと表記している．

図 2.9 より，重み変更概念ベースは新聞概念ベースと比較して，順序正解率と C 平均順序正解率の両方で精度を 3%程度改善できていることが分かる．本重み付け方法は，概念ベース以外の情報源を必要としないため，他の方法による属性追加を実施した場合においても汎用的に用いることが可能である．

＜シソーラスを用いた属性追加（シソーラス概念ベース）＞

○概要

概念ベースの意味知識をより充実させるためには，属性に様々な分野や観点の語を持たせる必要がある．これまでに構築してきた概念ベースは，国語辞書の語義文から概念の直接的な意味を表す属性群を，新聞記事中の語句から概念と何かしら関連があると考えられる多様な属性群を取得してきた．基本概念ベースと比べると，新聞記事を用いて拡張した新聞概念ベースは概念と属性の数が大幅に増加したことに加え，精度も向上している．よって，様々な情報源から概念に対して何かしら関連がある語を属性として取得し，属性数を増加させることで概念ベースの精度向上が期待できる．ここでは，人手によって作成された情報源であり，語の関係を表す情報を持つシソーラスを用いて属性の追加を行う．なお，本論文で使用するシソーラスである NTT シソーラスについては，2.5.1 項で説明する．

表 2.6: NTT シソーラスから追加される 1 概念あたりの属性候補数

関係	最大数	平均数
3 階層までの上位関係	38	4.0
1 階層までの下位関係	639	36.9
仲間関係	1423	150.3

○構築方法（属性候補の選別と追加）

まず，概念を NTT シソーラスのノードまたはリーフと対応付ける．その上で，上位・下位・仲間の関係にあるノードもしくはリーフを属性候補として抽出する．この時，概念の NTT シソーラス上での対応付けは表記一致により行う．つまり，NTT シソーラス上に存在しない概念に関しては属性の追加の対象としない．ただし，NTT シソーラスには特有の表記を持つノードが含まれるため，うまく対応付けられないことがある．そこで，以下に示すルールにより NTT シソーラス上の表記と概念表記を対応付ける．

記号の削除と分割

ノード「乗り物（部分（移動（陸圏））」やノード「飲物・たばこ」のように表記に記号が存在する場合，該当記号を削除した上で概念と対応付ける．この時，括弧記号を持つノードについては，先頭の語句のみを抽出して対応を取る．また，複数の語を併記する記号については，記号の前後で語句を分割し，それぞれの語句について対応付けを行う．

「等」の削除

ノード「嗜好品等」のように，末尾に「等」が付いた語句の総括を表すノードに関しては，「等」を削除して対応を取る．

NTT シソーラス上に対応付けられた概念の属性候補として，各概念から見て 3 階層を上限とした上位関係の語，1 階層を下限とした下位関係の語，ならびに仲間関係である語を取得する．例えば，2.5.1 項で説明する図 2.17 に示した例において，概念「カフェオレ」の属性候補をシソーラスから取得すると，上位関係の語からは「飲物，飲料，浄水」などが得られる．この条件に従うと，1 概念あたりにシソーラスから取得できる属性候補の数は表 2.6 のようになる．

ここで，下位関係と仲間関係から得られる属性候補は数が多く，特に仲間関係に関しては概念の意味知識としては適さない語も多く存在する．例えば，図 2.17 では，概念「カフェオレ」の仲間関係には「ビール」が存在するが，人間は両者に対して深い関連は感じない．そこで，これらの属性候補に対して選別処理を行う．属性の選別は重みを利用して実施する．NTT シソーラスから取得した属性候補群に対して，重み変更概念ベースを構築した際に適用した重み付け方法を用いて重みを付与する．付与した重みの降順に属性候補群を並べ替え，上から順に一定数を追加属性として選別する．

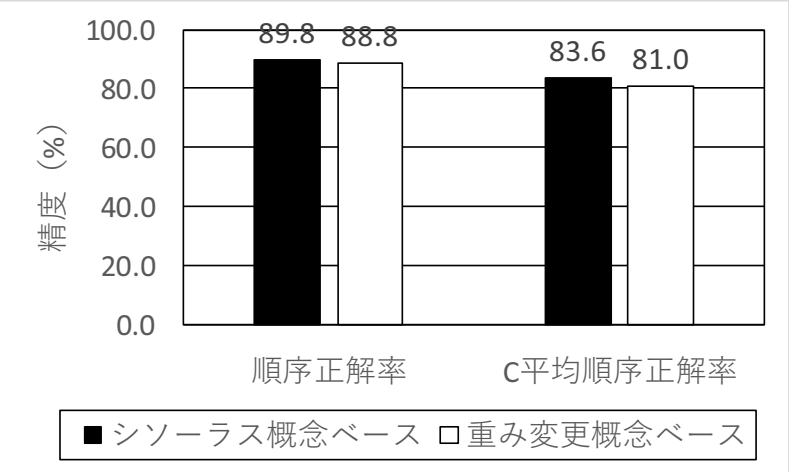


図 2.10: シソーラス概念ベースの評価結果

○性能評価

2.3.1.2 項で示した X-ABC 評価セットを用いてシソーラス概念ベースの精度評価を実施した．評価結果を図 2.10 に示す．なお，下位関係と仲間関係から追加する属性数は実験的に精度を求めた結果，下位関係から重み上位 10 個を，仲間関係から重み上位 15 個を属性として追加している．今回は，重み変更概念ベースにおいて重み上位 30 個 [32] の属性に限定した概念ベースを作成し，当該概念ベースに対して NTT シソーラスから抽出した属性を追加することでシソーラス概念ベースを構築した．

図 2.10 より，シソーラス概念ベースは重み変更概念ベースと比較して，順序正解率と C 平均順序正解率の両方で若干ではあるが精度を改善できていることが分かる．また，構築されたシソーラス概念ベースにおける 1 概念あたりの平均属性数は約 40 語となっている．

＜各概念ベースの X-ABC 評価における関連度の変化＞

これまでに構築してきた各概念ベースについて，X-ABC 評価において算出された関連度を表 2.7 に示す．$DoA(X, A)$ は X-ABC 評価セットにおける基準概念 X と概念 A の関連度，$DoA(X, B)$ は基準概念 X と概念 B の関連度，$DoA(X, C)$ は基準概念 X と概念 C の関連度を表す．なお，表 2.7 には，X-ABC 評価セットにおける関連度の平均値を示している．また，基本概念ベースとルール精錬概念ベースに関しては，X-ABC 評価セット内の基準概念が存在しないセットを除いて算出した平均値となっている．

2.3.1.2 項で述べた通り，X-ABC 評価セットは基準概念 X との関連が適切に表現できているか否かを評価する評価セットである．概念 A が最も基準概念 X と関連が強い語であり，概念 B は概念 A ほどではないが人間なら関連を見出すことができる語であり，概念 C は基準概念 X と全く関連の無い語である．つまり，基準概念 X との関連度は，概念 A が最も高く，概念 B が中程度，概念 C が最も低くなること（理想的には 0.0）が期待される．

表 2.7 より，基本概念ベースと比較して全ての概念ベースについて概念 A との関連度は 3 倍

表 2.7: 各概念ベースの関連度

概念ベース	$DoA(X, A)$	$DoA(X, B)$	$DoA(X, C)$
基本概念ベース	0.0848	0.0311	0.0030
ルール精錬概念ベース	0.2552	0.0704	0.0056
新聞概念ベース	0.4609	0.1039	0.0023
重み変更概念ベース	0.3806	0.0914	0.0034
シソーラス概念ベース	0.4490	0.0988	0.0030

以上の値となっており，各拡張方法や精錬方法が適切であり，大きな効果をもたらしていることが分かる．さらに，新聞概念ベースでは，概念 A に加えて概念 B との関連度が大きく上昇しているが，これは概念と属性の数が増加した結果，2.4.1 項で述べる関連度の計算方法の特性上，合致する属性や関連の強い属性が存在する可能性が高まったことに起因すると考えられる．

2.3.2 文書データベースを用いた概念化

2.3.2.1 概要

2.3.1 項で述べた通り，概念ベースは国語辞書や新聞記事などの情報源から構築されており，品質が高いというメリットがある．しかし，全ての語彙を網羅できていないという欠点もある．概念ベースに定義されていない未定義語は概念化することができないため，2.4 項で述べる意味的関連性評価方法を適用することができない．そのため，コンピュータに未定義語の意味を理解させることはできない．

そこで，未定義語については，インターネット百科事典であり，世界で最も収録語数が多いとされる Wikipedia[21] を利用して構築した文書データベースを用いて概念化を実施する．

2.3.2.2 関連ツール

本項では，文書データベースの構築に関連するツールとして MeCab と TermExtract を説明する．

○ MeCab

MeCab[35] は形態素解析エンジンの一つであり，対象言語における文法の知識や辞書を情報源として使用することで，自然言語で記載された文を形態素に分割し，各形態素の品詞を判別することが可能である．「学校に行く．」という文を MeCab で形態素解析した結果を表 2.8 に示す．

○ TermExtract

TermExtract[36, 37] は名詞の連接頻度に基づき専門用語を抽出するための Perl モジュール

表 2.8: MeCab の実行結果一例

入力語	読み方	基本形	品詞	活用形
学校	ガッコウ	学校	名詞-一般	
に	ニ	に	助詞-格助詞-一般	
行く	イク	行く	助詞-自立	五段・カ行促音便
.	.	.	記号-句点	

である．上述した MeCab に代表されるような形態素解析エンジンは，入力語を形態素単位で分割するため，多くの単語を組み合わせて構成されることが多い専門用語を抽出することは困難である．TermExtract は形態素解析エンジンが解析した結果をもとに名詞を連接した複合語を取得できることを利用して専門用語の抽出を可能としている．また，抽出した複合語や単語の重要度を取得できるという特徴も有する．

2.3.2.3　構築方法

文書データベースは以下の手順で構築される．

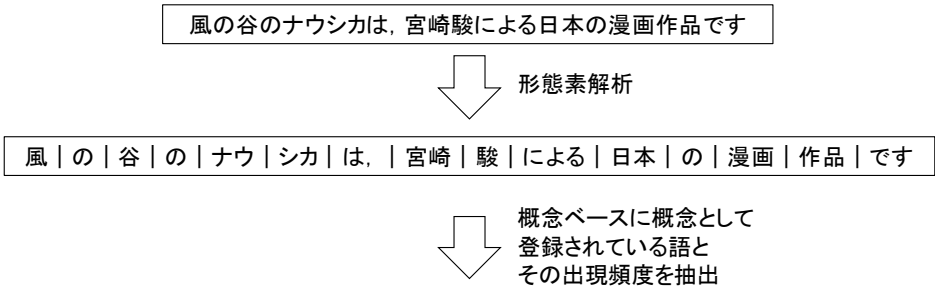
- 記事の取得
- 概念及び未定義語の抽出
- 文書の登録
- 文書の追加
- *idf* の登録
- 未定義語の属性と重みの取得

○記事の取得

Wikipedia（日本語版）の全文を記事毎に分割し，各記事を文書データベースに登録する．なお，文書データベースは NoSQL 型データベースである MongoDB を使用して構築している．本論文で使用した Wikipedia（日本語版）に登録されていた記事数は，約 80 万件である．

○概念及び未定義語の抽出

文書データベースに登録された各記事から概念と未定義語を抽出する．まず，各記事に対して形態素解析を実施することで取得された自立語から，概念ベースに概念として登録されている語とその出現頻度を取得する．概念の抽出例を図 2.11 に示す．



概念	頻度
風	1
谷	1
ナウ	1
日本	1
漫画	1
作品	1

図 2.11: 概念の抽出例

続いて，各記事から未定義語の抽出を実施する．多くの未定義語を抽出するために，以下に述べる 3 つの方法を併用して未定義語を抽出する．

まず，MeCab を利用して各記事の形態素解析を実施し，未知語と判別され，かつ，概念ベースに存在しない語を未定義語として抽出する．次に，TermExtract を利用して抽出された語の中で概念ベースに存在しない語を未定義語として抽出する．最後に，Wikipedia 日本語版の見出し語，共有辞書サービスであるはてなキーワード [38]，及び，日本語入力システムのインプットメソッドである Social IME[39] から辞書を構築し，当該辞書に含まれ，かつ，概念ベースに存在しない語を未定義語として抽出する．未定義語の抽出例を図 2.12 に示す．

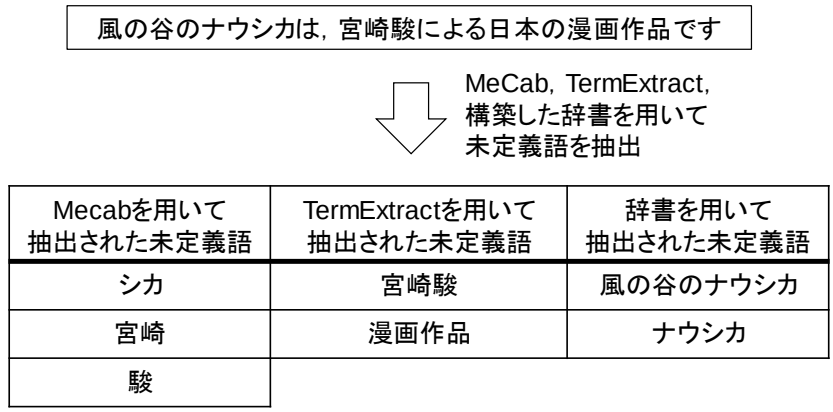


図 2.12: 未定義語の抽出例

表 2.9: レコード（文書）の一例

フィールド名	内容
記事	風の谷のナウシカは，宮崎駿による日本の漫画作品です
概念，頻度	（風，1），（谷，1），（ナウ，1），（日本，1），（漫画，1），（作品，1）
未定義語	シカ，宮崎，駿，宮崎駿，漫画作品，風の谷のナウシカ，ナウシカ

○文書の登録

各記事から抽出された概念とその出現頻度，及び，未定義語をレコードとして文書データベースに登録する．文書データベースに登録されるレコードの一例を表 2.9 に示す．以降，レコードを構成する記事，概念，概念の出現頻度，未定義語をまとめて文書と表記する．

○文書の追加

これまでの処理で構築された文書データベースの情報量を拡張するため，文書の追加を行う．文書の追加は，文書データベースに登録された未定義語を Web 検索エンジン [40] で検索することで実施する．具体的には，当該未定義語を検索キーワードとして情報検索し，上位 100 件の検索結果ページを取得する．そして，当該検索結果ページから概念ベースに登録されている自立語を概念として抽出し，当該概念が出現する頻度と併せて取得する．さらに，概念ベースに存在しない自立語を未定義語として取得する．このようにして取得された検索結果ページ，概念，概念の出現頻度，及び，未定義語も文書として文書データベースに追加登録する．文書を追加登録した結果，文書データベースに登録された文書数は約 630 万件となっている．

○ idf の登録

文書データベースに登録された文書には，概念と当該概念の出現頻度がセットで格納されてい

るが，当該概念の *idf* を追加する．*idf* は，新聞概念ベースを構築する際に述べた通り（2.3.1.3 項），空間中の総文書数を，語が出現する文書数で割った値の対数によって表される．ここでは，総文書数は文書データベースに登録された文書の総数，語が出現する文書数は文書データベースにおいて概念が出現する文書の数となる．なお，概念の出現頻度と *idf* は，後述する未定義語の属性と重みを取得する際に使用する．

○未定義語の属性と重みの取得

まず，未定義語が含まれる文書に格納されている概念が，当該未定義語の一次属性となる．未定義語が複数の文書に含まれている場合は，当該複数文書に登録されている概念が未定義語の一次属性となる．次に，一次属性に対する重みは，当該一次属性の出現頻度と *idf* を掛け合わせることで算出する．当該一次属性が複数の文書に含まれている場合は，出現頻度はその総和を利用する．なお，この重みの算出方法は，情報検索の分野で広く用いられている $tf \cdot idf$ [41, 42] の考え方を応用したものである．

2.3.2.4 改良

上述した処理によって構築した文書データベースには，獲得した未定義語の属性の一部が不適切という問題がある．「一覧」，「ニュース」，「記事」など Web ページに頻出する語が，多くの未定義語の属性として獲得される傾向にあるが，これらの語は雑音に相当し，属性として適切でないことが多い．また，別の問題として，MeCab による形態素解析の失敗により，未定義語が正しく獲得できないことも挙げられる．例えば，記事中に「スマートフォン」という語が含まれる場合，「スマート」と「フォン」に分解され，それぞれが文書を構成する概念として登録されることになるが，大抵これらの語は属性として不適切である．

そこで，文書データベースの改良を実施する．文書データベースの改良は，2.3.2.3 項で述べた 2 番目の手順である「概念及び未定義語の抽出」を変更することと，4 番目の手順である「文書の追加」を実施しないことで行った．以降，改良前の文書データベースを基本文書データベース，改良した文書データベースを改良文書データベースと表記する．

○概念及び未定義語の抽出

基本文書データベースでは，概念及び未定義語の抽出を MeCab，TermExtract，構築した辞書（Wikipedia 日本語版の見出し語，はてなキーワード，Social IME から構築した辞書）を利用して行った．一方，改良文書データベースでは，MeCab のみを利用して概念と未定義語の抽出を実施する．さらに，MeCab のシステム辞書を標準辞書である IPA 辞書から mecab-ipadic-NEologd[43] に変更する．mecab-ipadic-NEologd は，IPA 辞書に新語・固有名詞などを再録して拡張した辞書であり，定期的に更新されるという特徴がある．mecab-ipadic-NEologd を活用することで多くの未定義語を正しく抽出できるようになることが期待されるため，当該辞書を使用した MeCab による形態素解析のみを用いて概念及び未定義語の抽出を実施する．

○文書の追加

基本文書データベースでは，Web 検索エンジンを利用して文書の追加を実施したが，Web 検索エンジンからは玉石混合な情報が得られるために，結果として Web 上に頻出する語が未定義

表 2.10: 評価セットの一例（文書データベース）

年	未定義語			
2012	スカイツリー	iPS 細胞	オスプレイ	…
2013	あまちゃん	黒子のバスケ	クックパッド	…
2014	壁どん	広瀬すず	ソチオリンピック	…
2015	ドローン	マイナンバー	爆買い	…
2016	君の名は	EU 離脱	福原愛	…

語の属性として獲得された可能性がある．そこで，文書の追加は実施せず，改良文書データベースは 2.3.2.3 項で述べた 1 番目の手順である「記事の取得」に記載した通り，Wikipedia（日本語版）の全文のみから構築する．なお，Wikipedia（日本語版）の全文を取り直した結果，改良文書データベースに登録された記事数は約 108 万件となっている．

2.3.2.5　性能評価

構築した文書データベースの性能評価として，精度とヒット率の評価を実施した．

○精度評価

2012 年から 2016 年の各年において話題になった語の中から無作為に選択した 50 語の未定義語を評価セットとして使用した．話題になった未定義語は，Google Trends ランキング [44] または新語・流行語大賞 [45] にノミネートされた語から選定した．評価セットに対して，構築した文書データベースを用いて属性（一次属性）を 30 語獲得した．獲得した属性を 3 人で目視により評価し，2 人以上が適切と判断した属性の割合を精度として算出した．評価に使用した未定義語の一例を表 2.10 に示す．

精度評価結果を図 2.13 に示す．

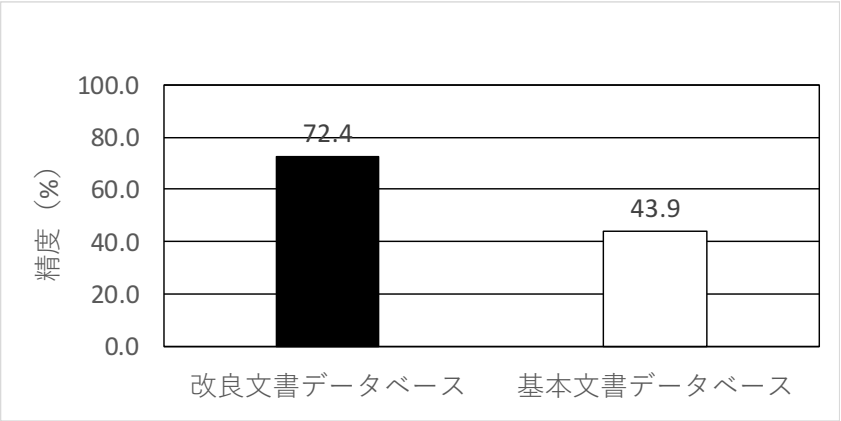


図 2.13: 文書データベースの精度評価結果

表 2.11: 文書データベースを用いて獲得した未定義語「福原愛」の属性の一例

改良文書データベース		基本文書データベース	
属性	評価	属性	評価
卓球	○	卓球	○
優勝	○	優勝	○
選手	○	選手	○
オリンピック	○	ニュース	×
女子	○	一覧	×

図 2.13 より，改良文書データベースは基本文書データベースと比較して精度を 28.5%改善できていることが分かる．これはインターネット百科事典であり質の高い情報で構成されている Wikipedia から得られる情報に限定して文書データベースを構築した結果，雑音に相当する語が除去され，未定義語に関連のある語が属性として追加されたことに起因すると考えられる．表 2.11 に未定義語「福原愛」の属性獲得結果の一部を示す．表 2.11 より，改良文書データベースは Web ページに頻出する語が除去され，未定義語「福原愛」に関連する語が属性として獲得できていることが分かる．

○ヒット率評価

2012 年から 2016 年における Google Trends ランキングまたは新語・流行語大賞にノミネートされていた未定義語 603 語に対して，文書データベースを用いて属性の獲得を行い，属性獲得に成功した割合をヒット率として算出した．ヒット率評価結果を図 2.14 に示す．

図 2.14 より，改良文書データベースは基本文書データベースと比較してヒット率を 16.4%改善できていることがわかる．これは MeCab の辞書を新語や固有名詞が多数登録されている

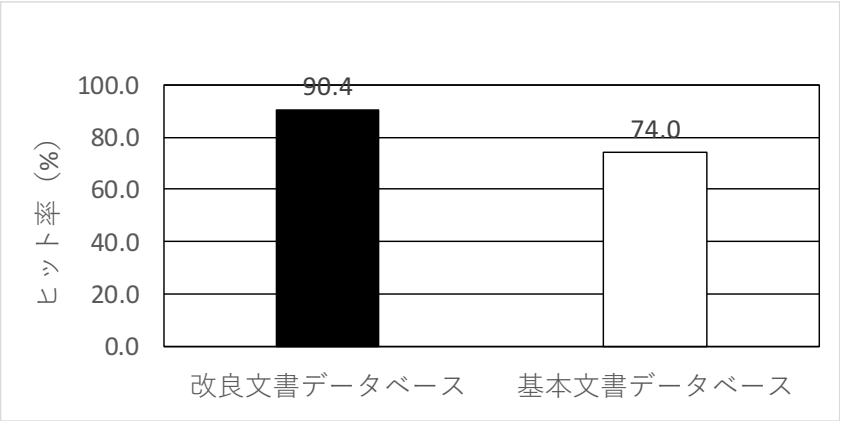


図 2.14: 文書データベースのヒット率評価結果

mecab-ipadic-NEologd に変更した結果，より多くの未定義語を登録できたことに起因すると考えられる．

2.3.3　オートフィードバックを用いた概念化

2.3.3.1　概要

2.3.2 項で述べた通り，文書データベースを使用することで，概念ベースに定義されていない未定義語を概念化することができるようになる．しかし，2.3.2.5 項に示したように，文書データベースは全ての未定義語を概念化できるわけではない．

そこで，インターネット情報である Web 検索エンジンの検索結果を利用することで，ほぼあらゆる未定義語に対応可能なオートフィードバックを用いて概念化を実施する [46]．

2.3.3.2　属性獲得方法

オートフィードバックによる属性獲得は以下の手順で実施される．

- 検索結果ページの取得及び文書の整形
- 形態素解析及びその補正
- 未定義語の属性と重みの取得

○検索結果ページの取得及び文書の整形

Web 検索エンジン [40] を用いて，未定義語を検索キーワードとして情報検索を実施し，検索結果の上位 100 件の検索結果ページを取得する．取得した検索結果ページは HTML 形式であ

るため，スクリプト部分とタグを取り除くことで文書を整形する．

○形態素解析及びその補正

整形した文書に対して茶筅 [47] を用いて形態素解析を行う．ただし，茶筅による形態素解析は，主に以下に挙げる 3 つの要因により解析に失敗することがある．1 番目の要因は，複合語の過多分割である．例えば,「基本情報技術者試験」に対して形態素解析を実施した場合は，単語の切れ目を誤った結果,「基本，情報，技術，者，試験」と分割してしまう．2 番目の要因は，英単語に対応できないことである（英単語は未知語と出力される）．3 番目の要因は，新しいカタカナ語に対応できないことである．例えば,「ストーカー」に対して形態素解析を実施した場合は，形態を誤った結果,「スる（動詞），トー（名詞），カー（名詞)」と分割してしまう．

そこで，上述した形態素解析の失敗を補正するルールを適用する．補正ルールの基本は，必要でない形態素，あるいは，区切り文字が出現するまで形態素を結合することである．ただし，語尾が「〜する」であるサ変動詞については語幹のみを取得している．これは，概念としては，語幹とサ変動詞は同一の語であるとみなすためである．例えば,「走行」と「走行する」は品詞は異なるが，語の意味としては同義と捉えることが可能である．

○未定義語の属性と重みの取得

まず，形態素解析結果から，概念ベースに登録されている概念のみを抽出し，それらを未定義語の一次属性とする．次に，一次属性に対する重みを，$tf \cdot idf$ [41, 42] の考え方を応用して算出する．

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 t の出現頻度 $tfreq(t, A)$ を，検索結果ページ A 中の全ての自立語の語数 $tnum(A)$ で割ることで算出される．計算式は以下のようになる．

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

次に，語の特定性である idf 値は，以下に述べる 3 種類の方法を使用する．

1 番目の idf は Web-Inverse Document Frequency（$Web\text{-}idf$）である．$Web\text{-}idf(t)$ は，Web 上において語 t が持つ大局的な重みを示し，以下の式で算出される．$Web\text{-}idf$ では，idf における対象とする文書空間における総文書数を示す N は，Web 検索エンジンが保有している日本語ページの数となる．ただし，オートフィードバックにおいて使用する Web 検索エンジンが保有している日本語ページの数は公開されていない．そこで，日本語の文書として頻出する「は」で検索を行った場合におけるヒット数を，$Web\text{-}idf$ の算出式における $N_{Web\text{-}idf}$ とした．また，自立語 t が出現する文書数である $df(t)_{Web\text{-}idf}$ は，語 t を Web 検索エンジンで検索を行った時のヒット件数である．

$$Web\text{-}idf(t) = \log \frac{N_{Web\text{-}idf}}{df(t)_{Web\text{-}idf}}$$

あらかじめ概念ベースに登録されている全ての概念と，NTT シソーラス [5] にリーフとして登録されている全ての語に対して，$Web\text{-}idf$ 値を算出しておくことで，未定義語の一次属性の idf 値として使用する．

表 2.12: 評価セットの一例（オートフィードバック）

未定義語		
BSE	サイクロン掃除機	国連平和維持活動
SARS	チェルノブイリ原発事故	水平対抗エンジン
カブトミジンコ	ロード・オブ・ザ・リング	おれおれ詐欺
…	…	…

2 番目の *idf* は Statics Web-Inverse Document Frequency（*SWeb-idf*）である．定義上，*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 を対象とする場合，Web 上の全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 [40] で検索する．続いて，検索結果の上位 10 件の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な Web の全情報空間とみなし *SWeb-idf* 値を算出する．語 t の *SWeb-idf* 値である $SWeb\text{-}idf(t)$ は以下の式で算出される．ここで $N_{SWeb\text{-}idf}$ は固有名詞 1,000 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（$N = 10,000$），$df(t)_{SWeb\text{-}idf}$ は自立語 t が出現する検索結果ページ数である．

$$SWeb\text{-}idf(t) = \log \frac{N_{SWeb\text{-}idf}}{df(t)_{SWeb\text{-}idf}}$$

3 番目の *idf* は Now-Web-Inverse Document Frequency（*Now-Web-idf*）である．*Web-idf* と *SWeb-idf* はいずれも事前に *idf* 値を算出しておく方法であるため，*idf* 値が与えられる語は有限である．そこで，*Now-Web-idf* は，形態素解析結果から未定義語の一次属性の候補となる自立語を獲得する時に，全ての自立語候補に対して，*Web-idf* と同様の方法を用いて *idf* 値を算出する．ただし，*Now-Web-idf* は全ての語に対して *idf* 値を算出できる反面，オートフィードバックによる未定義語の属性を行うたびに，Web 検索エンジンを使用してヒット数を獲得する必要があるため，処理に時間を要するというデメリットを持つ．

以上より，未定義語の一次属性の重みは，*tf* 値に，上述した方法で算出した *idf* 値を掛け合わせることで算出する．

2.3.3.3　性能評価

無作為に選択した 120 語の未定義語を評価セットとして使用した．評価セットに対して，オートフィードバックを用いて属性（一次属性）を 50 個獲得した．獲得した属性を目視により評価し，適切と判断した属性の割合を精度として算出した．評価に使用した未定義語の一例を表 2.12 に，オートフィードバックを用いて属性を取得した評価結果の一例を表 2.13 に示す．

精度評価結果を図 2.15 に示す．

表 2.13: オートフィードバックを用いて獲得した未定義語「BSE」の属性の一例

属性	評価
BSE	○
牛海綿状脳症	○
狂牛病	○
検査	○
リーフレット	×
米 BSE	○
関係	×
牛肉	○
…	…

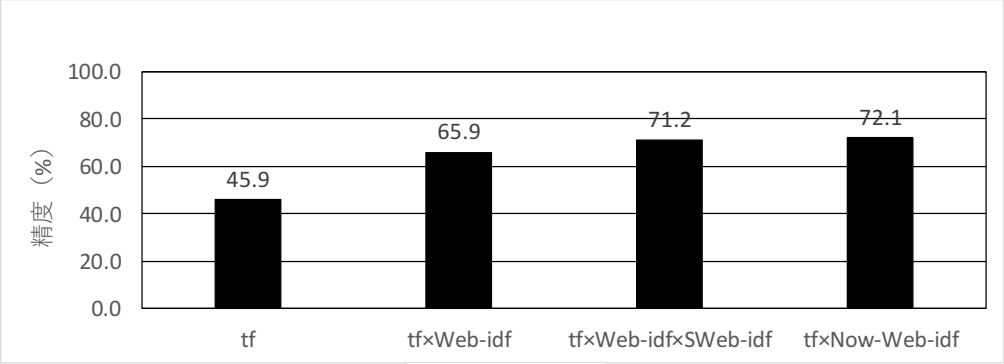


図 2.15: オートフィードバックの精度評価結果

図 2.15 より，属性の重み付けに *tf* 値を単独で使用する方法と比較し，*idf* 値を使用することで，精度を 20%以上改善できていることが分かる．具体的には，*tf* 値に，*Web-idf* 値を掛け合わせることで 20.0%，*Web-idf* 値と *SWeb-idf* 値を掛け合わせることで 25.3%，*Now-Web-idf* 値を掛け合わせることで 26.2%改善している．

精度の改善は，語の特定性を考慮する *idf* を導入したことで，Web ページ上に頻出する語（「ページ」,「キーワード」,「コンテンツ」など）の重みが相対的に低下したことに起因すると言える．3 種類の *idf* については，ヒット数のみを考慮する *Web-idf* を単独で利用する方法と比較し，疑似的ではあるものの Web 上の情報空間に出現する文書数（ページ数）を考慮できる *SWeb-idf* を併用することで，適切な重みを付与することに成功したと考えられる．また，あらかじめ *idf* 値を算出することが必要な *Web-idf* と *SWeb-idf* に対して，*Now-Web-idf* は最新の Web 情報に基づいて *idf* 値を算出することができるため，最も高い精度が得られたと考えられる．ただし，*Web-idf* と *SWeb-idf* を組み合わせた場合と，*Now-Web-idf* を使用した場合の精

度の差分は 1%以下と低い．さらに，*Now-Web-idf* を使用する方法は，他の方法と比べて処理時間が 30 倍近く増大することを確認している．したがって，オートフィードバックにおける重み付けは，*tf* 値に，*Web-idf* 値と *SWeb-idf* 値を掛け合わせる方法を採用する．

2.4　意味的関連性評価方法

2.4.1　関連度計算

関連度計算は概念と概念の間に人間が見出す関連の強さを定量的に表現する計算方法である [48, 49]．関連度計算では，概念同士が持つ各属性（一次属性）を意味が近い属性で対応付ける．その上で対応付けられた属性同士の属性，つまり，元の概念の二次属性同士を比較することで意味の類似度合いを算出する．算出された意味の類似度合いが，概念同士の関連度となる．

まず，属性の対応付けに用いる一致度について説明する．概念 A 及び概念 B の一次属性をそれぞれ a_i，b_j とし，各一次属性に対応する重みを u_i，v_j とする．概念 A が持つ属性数を L 個，概念 B が持つ属性数を M 個（$L < M$）とすると，概念 A 及び概念 B は以下のように定義される．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 及び概念 B の一致度 $DoM(A, B)$ を以下の式で定義する．

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

ここで，$a_i = b_j$ は属性同士が表記的に一致した場合を示すことになる．つまり，一致度は，概念 A と概念 B が共通して持つ属性のうち，小さい方の重みを足し合わせることで算出される．これは，大抵の場合，共通した属性は概念 A と概念 B においてそれぞれ異なる大きさの重みが付与されており，小さい方の重み分は概念 A と概念 B の両方の属性に有効であると考えるためである．つまり，一致度とは，双方の概念にとって有効な属性が持つ重みの和を示す数値である．関連度計算では，この一致度を用いることで，関連度を算出する概念同士の属性を対応付けた上で，関連度の算出を実施する．

次に，関連度について説明する．関連度は共通属性を考慮した関連度と共通属性を除外した関連度を足し合わせることで算出する．

共通属性を考慮した関連度は次のように算出する．関連度を算出する概念が持つ属性のうち，双方の概念に共通して存在する属性を抽出する．概念 A と概念 B の双方に共通して存在する属性が t 個であったとすると，その t 個の属性を優先して抽出し，共通する属性同士が対応付けれる．そして，この対応付けられた属性を利用して，共通属性関連度 $DoA_{com}(A, B)$ は以下の式から算出される．

$$
DoA_{com}(A, B) = \sum_{a_i = b_j} \min(u_i, v_j)
$$

共通属性を考慮した関連度は上に記載した式の通り，共通属性を用いた一致度の算出と同様の方法で算出される．一致度を算出する計算において，共通属性が持つ重みのうち，小さいほうの値が利用される．そこで，用いられなかった大きいほうの値を持つ重みを，以降の関連度を算出する計算に活用するために，概念の再定義を行う．具体的には，$a_i = b_j$ の時に，各属性の重みの関係が $u_i > v_j$ となる場合，属性 a_i の重みを $u_i - v_j$ に更新した上で，属性 b_j を概念 B から除外する．逆に，各属性の重みの関係が $u_i < v_j$ となる場合は，属性 b_j の重みを $v_j - u_i$ に更新した上で，属性 a_i を概念 A から除外する．上述の処理により，概念 A と概念 B に共通属性が存在しなくなる．この再定義された概念を概念 A' と概念 B' とする．

続いて，共通属性が除外された概念 A' 及び概念 B' の関連度を算出する．所持する属性数が少ない方の概念 A' を基準とし，概念 A' が持つ一次属性の並びを固定する．その上で，概念 B' の一次属性を，概念 A' が持つ各一次属性との一致度の和が最大になるように並び替える．この時，概念 A' の属性と重みを (a'_i, u'_j)，概念 B' の属性と重みを (b'_i, v'_j) とし，次のように定義する．なお，T_1 は概念 A から除外された共通属性の数，T_2 は概念 B から除外された共通属性の数である．

$$
\begin{aligned}
A' &= \{(a'_i, u'_i) \mid i = 1 \sim L - T_1\} \\
B' &= \{(b'_j, v'_j) \mid j = 1 \sim M - T_2\}
\end{aligned}
$$

さらに，共通属性を除去した概念間の関連度 $DoA_{def}(A', B')$ を以下の式で定義する．

$$
\begin{aligned}
DoA_{def}(A', B') &= \sum_{k=1}^{L-T_1} \{DoM(a'_k, b'_k) \cdot \frac{\min(u'_k, v'_k)}{\max(u'_k, v'_k)} \cdot \frac{u'_k + v'_k}{2}\} \\
\min(\alpha, \beta) &= \begin{cases} \alpha & (\beta \geq \alpha) \\ \beta & (\alpha > \beta) \end{cases} \\
\max(\alpha, \beta) &= \begin{cases} \alpha & (\alpha \geq \beta) \\ \beta & (\beta > \alpha) \end{cases}
\end{aligned}
$$

そして，共通属性を考慮した関連度 $DoA_{com}(A, B)$ と共通属性を除去した関連度 $DoA_{def}(A', B')$ の合計を，概念 A と概念 B の関連度 $DoA(A, B)$ と定義する．

$$
DoA(A, B) = DoA_{com}(A, B) + DoA_{def}(A', B')
$$

関連度は概念間の関連の強さを 0〜1 の間の連続値で表し，値が高いほど概念間の関連が深いことを示す．表 2.14 に関連度計算の一例を示す．

2.4.2 Earth Mover's Distance

2.4.1 項において，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説明した．関連度計算は関連性が高い順に属性の対応を取ることで計算を行う．つまり，1 対 1 で対応を取り，関連の強さを評価する計算方法である．そのため，両概念に対して，数が少ない方の属性数分しか対応を取ることができない．例えば，概念 A が持つ属性の数が 3 個，概念 B が持つ属性の数が 100 個であった場合，概念 B の属性 97 語は関連度計算の対象外となる．そこで，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

表 2.14: 関連度計算の一例

概念 A	概念 B	関連度
飲食店	食堂	0.167
飲食店	米	0.148
飲食店	青空	0.047
仲買	仲介	0.302
仲買	市場	0.141
仲買	仕舞う	0.086
沿岸	海岸	0.841
沿岸	船	0.154
沿岸	練乳	0.022

対応を取ることができる Earth Mover's Distance（EMD）を用いて概念間の意味的な関連性を評価する方法を利用する．

○概要

EMD[50] は線形計画問題の一つであるヒッチコック型輸送問題 [51] において計算される距離尺度である．輸送問題とは，需要地の需要を満たすように供給地から需要地へ輸送を行う場合の最小輸送コストを解く問題である．輸送問題において EMD は，二つの離散分布において，一定の分布を他方の分布に変換する際の最小コストとして定義される．

EMD を算出する際，ある分布 P は m 個の特徴量 p_i とその特徴量に対する重み w_{p_i} を用いて，以下のように表現される．

$$\text{分布 } P = \{(p_1, w_{p_1}), (p_2, w_{p_2}), \cdots, (p_m, w_{p_m})\}$$

同様に，別の分布 Q も n 個の特徴量 q_i と重み w_{q_i} を用いて，以下のように表現する．

$$\text{分布 } Q = \{(q_1, w_{q_1}), (q_2, w_{q_2}), \cdots, (q_m, w_{q_m})\}$$

ここで，特徴量 p_i と q_j の距離を d_{ij} とし，全特徴量間の距離を $D = [d_{ij}]$ とする．この時，特徴量 p_i から q_j への輸送量を f_{ij} とすると，全輸送量は $F = [f_{ij}]$ となる．そして，以下に示すコスト関数 $WORK$ を最小とする輸送量 F を求めることで，EMD を計算する．

$$WORK(P, Q, F) = \sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij}$$

ただし，コスト関数 $WORK$ を最小化する際，以下の制約条件を満足することが必要である．

【条件 1】　$f_{ij} \geq 0, \quad 1 \leq i \leq m, \quad 1 \leq j \leq n$

【条件 2】　$\sum_{j=1}^{n} f_{ij} \leq w_{p_i}, \quad 1 \leq i \leq m$

【条件 3】 $\sum_{i=1}^{m} f_{ij} \leq w_{q_j}, \quad 1 \leq j \leq n$

【条件 4】 $\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min\left(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{q_j}\right)$

条件 1 は輸送量が正の値であることを示し，特徴量 p_i から q_i に輸送される一方通行であることを表している．つまり，分布 P が需要地の分布を示し，分布 Q が供給地の分布を示すことになる．条件 2 は輸送元である特徴量 p_i の重み w_{p_i} 以上に輸送できないことを表す．条件 3 は輸送先である特徴量 q_j の重み w_{q_j} 以上に受け入れることができないことを表す．条件 4 は総輸送量の上限を示し，その上限は輸送元または輸送先の特徴量における重みの総和の小さい方に制限されることを表している．

以上の制約条件に基づいて求められた全輸送量 F を用いて分布 P，Q 間の EMD を以下の式から計算する．この式より，EMD は 2 つの分布における特徴量の数が異なっている場合であっても計算が可能であることが分かる．なお，最適なコスト関数 $WORK$ を EMD の値としてそのまま使用しない理由は，コスト関数が輸送元または輸送先の特徴量における重みの総和に依存することを考慮し，正規化することでその影響を取り除くためである．

$$EMD(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}}$$

○ EMD の関連性評価への適用

EMD は需要地の特徴量とその重み，供給地の特徴量とその重み，需要地と供給地間の距離を定義できれば，どのような問題にも適用することが可能である．[52] の研究により，EMD を用いて 2 つの概念間の関連性を定量的に表現する方法の有効性が報告されている．以下に EMD を概念間の関連性評価に適用する方法を説明する．

EMD を概念間の関連性評価に適用する場合，概念の一次属性を特徴量とみなし，一次属性の重みを特徴量の重みとみなすことで，一次属性と重みの集合を離散分布と考える．また，後述するが，需要地と供給地間の距離に相当する情報として一次属性間の一致度を利用する．ある概念の離散分布を別の概念の離散分布に変換すると考えると，その際のコストが最小となる概念を元の概念と最も関連が高い概念とみなすことができるため，EMD を概念間の関連性評価に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方法について，図 2.16 に示す簡略図を用いて説明する．ある概念 A と B があった時，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 の離散分布とみなす．EMD では変換コストを算出する際に離散分布を構成する要素同士の距離（需要地と供給地間の距離）を用いる．EMD を用いた意味的な関連性評価方法では，当該距離を一次属性同士の関連性であると考え，一致度によって算出する．

図 2.16 において，属性 a_i と属性 b_j 間の距離 $dis(a_i, b_j)$ は次の式で表される．一致度は属性同士の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属性間の距離としている．

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

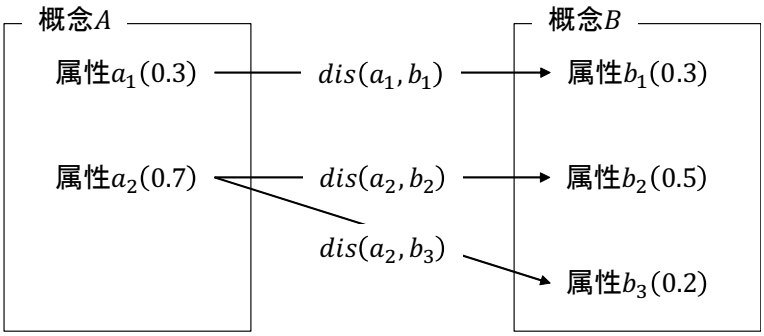


図 2.16: EMD を用いた意味的な関連性評価方法

ここで，図 2.16 の例における属性 a_1 と属性 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．これは属性 a_1 と属性 b_1 の距離に重みをかけたものである．このとき，属性 a_1 と属性 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，属性 a_1 からの重みの輸送はこの時点で終了する．

$$cost(a_i, b_j) = dis(a_i, b_j) \cdot 0.3$$

同様に他の属性の組み合わせについてもコストの計算を行い，最終的に全ての輸送経路のコストを足し合わせた値が EMD となる．図 2.16 の例では概念 A と概念 B 間の EMD は次のように表される．

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

以上の式で算出された EMD の最小値を最適化計算で求めることで，概念間の関連性（意味的な近さ）を算出する．

2.5　シソーラス

シソーラスとは，単語を意味的に分類した分類体系である．シソーラスの多くは木構造を持ち，名詞の集合を分類した名詞シソーラスや用言の集合を分類した用言シソーラスなどがある．また，木構造の葉（以下，リーフ）のみに単語が所属する分類シソーラスや根と中間ノードの両方に単語が所属する上位下位シソーラスがある．

2.5.1　NTT シソーラス

本論文では，木構造を持つ名詞シソーラスであり，上位下位シソーラスの 1 つである NTT シソーラス [5] を用いる．NTT シソーラスは一般名詞の意味的用法を表す 2710 個のノードの上位-下位関係及び全体-部分関係が木構造で示されている．ノードに所属する名詞として約 13 万語のリーフが登録されている．図 2.17 に NTT シソーラスの一部を示す．

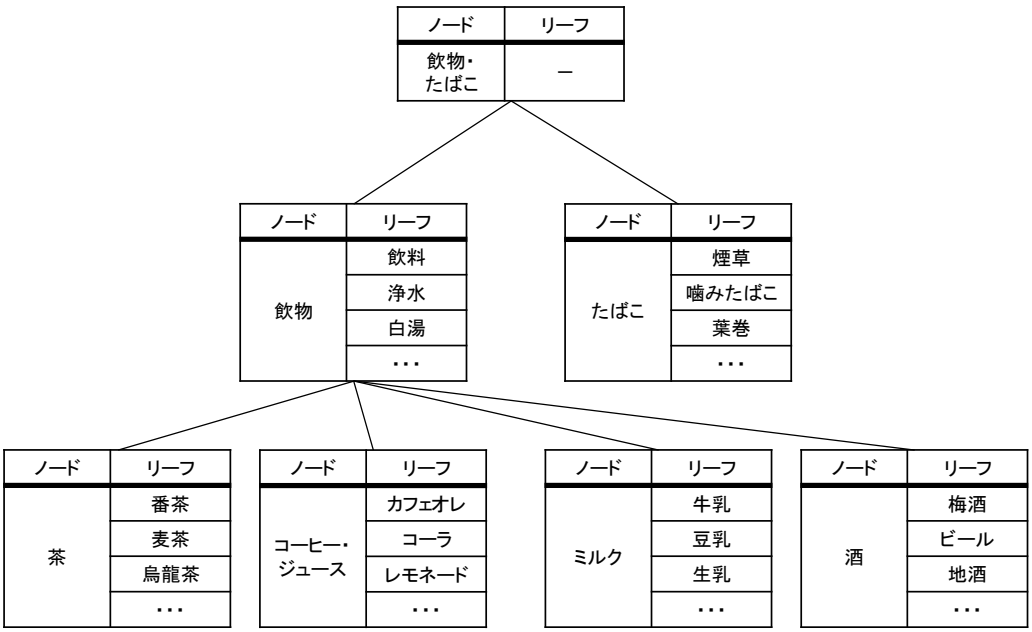


図 2.17: NTT シソーラスの一例

図 2.17 において，上側に位置するノードがより上位のノードであり，抽象的な語になっている．例えば，リーフ「カフェオレ」に着目すると，リーフ「カフェオレ」の 1 階層上位の関係にある語はノード「コーヒー・ジュース」となる．また，1 階層上位に共通のノードを持つ語同士を仲間関係になると定義する．ノード「コーヒー・ジュース」の場合，ノード「茶」，ノード「ミルク」，ノード「酒」が仲間関係になる．

2.5.2 NTT シソーラスのノードの概念化

第 3 章と第 5 章で述べる固有名詞の意味推定法では，未定義語と NTT シソーラスのノードに対して関連度計算を行うことになる．関連度計算を適用するためには，対象となる単語に属性と重みの集合を与える概念化を実施することが必要である．未定義語の概念化は 2.3.2 項や 2.3.3 項で述べた方法を用いて実施することができる．本項では，NTT シソーラスと概念ベースを用いて NTT シソーラスのノードを概念化する方法 [53] について説明する．

まず，ノードに所属する全てのリーフに対して概念ベースを参照し，当該リーフを概念とみなすことで，その一次属性を当該ノードの属性として取得する．上記処理を全てのノードに対して行うことで，NTT シソーラスにおける全てのノードに属性を与えることができる．ノードから取得した属性の一例を図 2.18 に示す．

続いて，取得したノードの属性に対する重みを，$tf \cdot idf$ [41, 42] の考え方を応用して算出する．語の網羅性である tf 値は，概念ベースから取得した属性の重みを使用する．また，語の特定性である idf 値は，以下の計算式から算出する．ここで，N は全てのノードの数，$df(Attribute)$

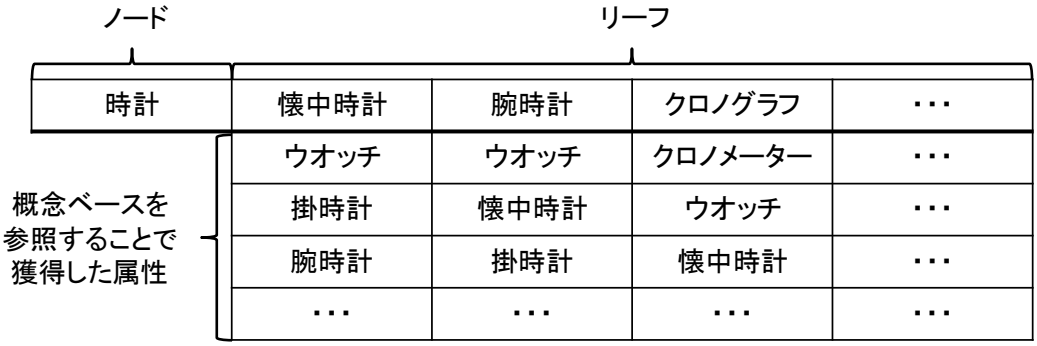


図 2.18: ノードから取得した属性の一例（ノード「時計」）

表 2.15: NTT シソーラスのノードの概念化結果一例（ノード「時計」）

属性	重み
懐中時計	0.133
掛時計	0.097
置時計	0.078
腕時計	0.070
ウオッチ	0.062
…	…

は概念ベースから取得した属性 *Attribute* が出現するノードの数である．

$$idf(Attribute) = \log \frac{N}{df(Attribute)}$$

そして，得られた *tf* 値と *idf* 値を掛け合わせた値がノードの属性（一次属性）の重みとなる．表 2.15 に NTT シソーラスのノードの一つである「時計」を概念化した例を示す．なお，取得した属性の重みは，その総和が 1.0 となるよう正規化している．

2.6　まとめ

本章では，コンピュータに人間が行っているような連想能力や意味理解能力を持たせるために重要となる連想メカニズムについて述べた．本論文において，連想メカニズムは，語の概念化方法，意味的関連性評価方法，および，シソーラスから構成される．

まず，語の概念化方法は，単語に属性と重みの集合を与える方法であり，本論文では 3 つの概念化方法を使用する．

1番目は電子化国語辞書などから構築した概念ベースを用いる方法であり，辞書などに登録されている多くの語を概念化することができる．概念ベースを用いた概念化方法では，電子化国語辞書に加え，ルールを用いて属性を精錬したり，新聞記事やシソーラスを用いた拡張を行ったりすることで，精度を改善できることを示した．

2番目はインターネット百科事典である Wikipedia から構築した文書データベースを用いる方法であり，概念ベースに定義されていない未定義語を概念化することができる．文書データベースを用いた概念化方法では，世界で最も収録語数が多いとされる Wikipedia の各記事に出現する未定義語と概念ベースに登録されている概念を抽出し，当該概念を当該未定義語の一次属性とする．さらに，当該一次属性に対して，$tf \cdot idf$ の考え方を応用して重みを付与することで，未定義語を概念化している．

3番目はインターネット情報である Web 検索エンジンの検索結果を利用するオートフィードバックを用いる方法である．オートフィードバックを用いた概念化方法はインターネット情報を利用するため品質はやや低いもののほぼあらゆる未定義語に対応可能である．そのため，文書データベースでも概念化できなかった未定義語を概念化することができる．当該概念化方法では，未定義語を検索キーワードとして情報検索を実施し，検索結果ページから概念ベースに定義されている概念を抽出し，当該未定義語の一次属性とする．さらに，当該一次属性に対して，$tf \cdot idf$ の考え方を応用して重みを付与することで，未定義語を概念化している．このとき，3種類の *idf* 値を比較した結果，予めヒット件数を活用して *idf* 値を算出する *Web-idf* と，仮想的な Web 上の情報空間を用いて統計的に *idf* 値を算出する *SWeb-idf* を掛け合わせることが有効であることを示した．

次に，意味的関連性評価方法は，概念化された語と語の間の関連を定量的に評価することで意味の近さを測ることを可能とする方法であり，本論文では2つの評価方法を使用する．

1番目は関連度計算であり，概念同士が持つ一次属性を意味が近い属性で対応付けた上で，対応付けられた属性が持つ属性（元の概念の二次属性）を比較することで意味の類似度合いを算出することができる．

2番目は輸送問題における距離尺度である Earth Mover's Distance を用いる方法であり，ある概念を別の概念に変換（輸送）する際に必要となるコストを，概念同士の意味の近さとして算出することができる．関連性が高い順に属性の対応を取るために1対1で対応を取る関連度計算に対して，Earth Mover's Distance を用いる方法は属性の対応を M 対 N で取ることが可能である．

最後に，シソーラスは単語を意味的に分類した分類体系であるが，本論文では木構造を持つ名詞シソーラスであり，上位下位シソーラスの1つである NTT シソーラスを使用する．本章では，NTT シソーラスと概念ベースを用いて NTT シソーラスのノードを概念化する方法について説明した．概念化したノードは，第3章と第5章で述べる固有名詞の意味推定法において，未定義語と関連度計算を行うために使用する．

第3章　文書データベースを用いた固有名詞の意味推定法

3.1　はじめに

本章では，文書データベースを用いて固有名詞の意味を推定する方法について取り上げる．日常的な会話の中では，固有名詞に代表される様々な未定義語が使用される．未定義語に関する知識が無ければ，会話文を理解することは困難である．Web 検索を利用することで未定義語について調べることは可能であるが，Web 上には膨大な情報が存在するため，必要な情報を効率的に得ることは困難である．

そこで，本章では，基準となる言語資源である概念ベースに加えて，文書データベースと NTT シソーラスを言語資源として活用することで，未定義語の意味推定を行う方法を提案する．具体的には，Wikipedia を情報源として構築した文書データベースを用いて未定義語を概念化し，概念化した未定義語が所属するべき NTT シソーラスのノードを提示することで，未定義語の意味推定を行う．なお，本章では，未定義語の代表的な存在である固有名詞を対象として意味推定を行う．

3.2　提案方法

文書データベースを用いた固有名詞の意味推定法は以下の手順で実施される．

まず，入力された未定義語に対して，文書データベースを用いて概念化を行う（2.3.2 項参照）．ここで，概念化された未定義語 U を l 個の属性 u_i と重み w_{u_i} を用いて，以下のように表現する．

$$\text{未定義語 } U = \{(u_1, w_{u_1}), (u_2, w_{u_2}), \cdots, (u_l, w_{u_l})\}$$

次に，NTT シソーラスのノードの概念化を行う（2.5.2 項参照）．同様に，概念化されたノード N_a を m 個の属性 n_{a_i} と重み $w_{n_{a_i}}$ を用いて，以下のように表現する．

$$\text{ノード } N_a = \{(n_{a_1}, w_{n_{a_1}}), (n_{a_2}, w_{n_{a_2}}), \cdots, (n_{a_m}, w_{n_{a_m}})\}$$

最後に，概念化を行った未定義語 U と，概念化を行った NTT シソーラスの各ノード N_a に対して関連度計算を実施し，最も高い関連度を有するノードを当該未定義語が所属するべきノードとして出力する．

なお，本提案方法では，未定義語を最も詳しく説明するノードに分類するという考えから，未定義語を分類するノードを最下位ノード（1926 個）に限定している．さらに，最下位ノードの

表 3.1: 使用する最下位ノードの選別結果一例

ノード	使用有無
時計	○
政治家	○
役人	○
茶	○
企業	○
自称（単数／男）	×
接辞（女／単数）	×
水たまり	×
煙	×
点（形状）	×

中で，未定義語が分類されることはないと判断できるノードを人手で削除している．なお，判断基準としては，3 名の被験者が各最下位ノードに未定義語が分類されるノードか否かを判断し，そのうち 3 名全員が未定義語が分類されることはないと判断したノードを削除している．結果，使用するノード数は 385 個となっている．つまり，上述したノード N_a を示した式において，a は 1〜385 の整数である．表 3.1 に選別したノードの一例を示す．

3.3　評価実験

3.3.1　実験条件

本章では，未定義語に関する表現として固有名詞を扱う．これは，人間が会話を行う際に用いられる一般名詞については，約 9 万語が概念として登録されている概念ベースと，ノードとリーフをあわせて約 13 万語の単語が登録されている NTT シソーラスを利用することで，意味を取得できると考えたためである．評価実験に使用する固有名詞として，20 名の被験者から各 10 個ずつ固有名詞と当該固有名詞が所属するべき NTT シソーラスのノードを提供してもらうことで，200 語のテストセットを作成した．さらに，作成したテストセットの中から，Wikipedia に存在しない語，および，多義性を持つ語を人手で除外することで抽出した 153 語を，本章が提案する方法を評価するための評価セットとして使用した．これは，世界で最も収録語数が多いとされている Wikipedia に存在しない語は一般的な語ではないという考えと，多義性を持つ語の意味推定は第 4 章で扱うため，上述の除外を実施した．作成した評価セットの一例を表 3.2 に示す．

表 3.2: 評価セットの一例（文書データベースを用いた固有名詞の意味推定法）

未定義語	所属ノード（正解ノード）
FinePix	カメラ
スターウォーズ	映画・映像
大谷山荘	宿泊施設
同志社	学校
ストラディバリウス	楽器

3.3.2 評価結果

文書データベースを用いて未定義語の概念化を行い，当該未定義語が所属する NTT シソーラスのノードを出力することで評価を実施した．提案方法が発話を伴う人工知能ロボットとの会話に適用される場合には単一の候補のみを出力することが必要となる．しかし，会話内容が画面に出力されるなど，出力インタフェースに画面を使用する場合には複数の候補を提示することでユーザの理解を支援する運用方法なども考えられるため [54, 55]，本章では順位をつけて複数の結果を出力することにした．なお，順位は未定義語と各ノードの関連度に基づいて付与している．つまり，最も関連度が高いノードが 1 位，2 番目に関連度が高いノードが 2 位となる．本評価では，10 位のノードまで出力を行った．

評価結果を図 3.1 に示す．図 3.1 は出力したノードに正解ノードが含まれる精度を示している．つまり，算出した未定義語と各ノードの関連度に対して，上位 5 個までのノードを出力した場合に正解ノードが含まれる精度は 70.5%であった．図 3.1 より，精度は第 1 位のみを出力した場合は 43.1%であり，第 5 位までを出力した場合は 70%を超え，第 10 位までを出力した場合は 80%を超えており，提案方法はある程度未定義語と関連があるノードを獲得できていることが分かる．

3.4 検索ヒット数を考慮した性能改善

本節では，提案方法の精度を改善するために，Web 検索エンジンから取得した検索ヒット数を利用する方法であるノード動詞と共起ヒット [56] について述べる．

3.4.1 ノード動詞

NTT シソーラスは作成者がある分類基準に従って単語を分類したものである．そのため，NTT シソーラスには「あるノードに所属するリーフは，そのリーフの直後に現れる助詞を伴う動詞が同じである」という関係が存在する．例えば，ノード「茶」に属するリーフ「番茶」や「麦茶」などには,「番茶を飲む」や「麦茶を飲む」など直後に現れる助詞を伴う動詞が共に「を飲む」であることが分かる．ノード動詞とはこの関係を利用してノードに設定した検索キーワードのことである．

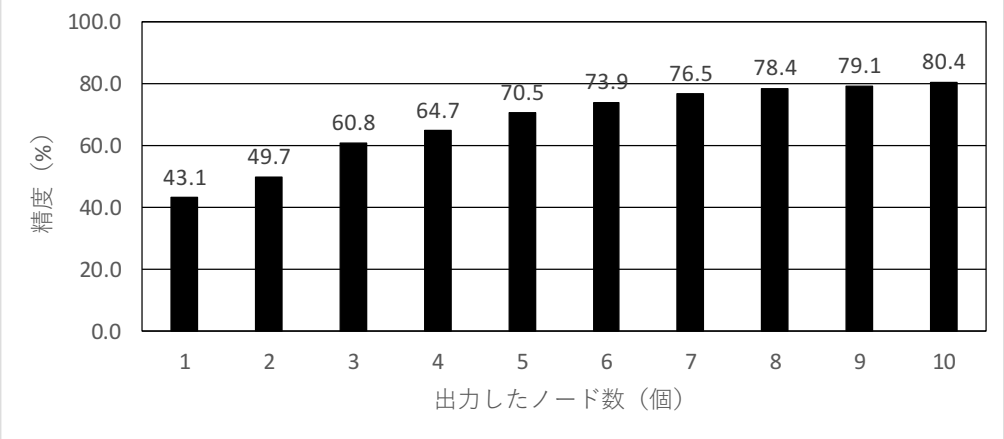


図 3.1: 精度評価結果（文書データベースを用いた固有名詞の意味推定法）

ノード動詞を利用する場合，未定義語に NTT シソーラスのノードごとに対応する助詞を伴う動詞（ノード動詞）を連結したキーワードを Web 検索エンジン [40] に入力し，検索ヒット数を獲得する．そして，獲得した検索ヒット数を未定義語が所属するべきノードの決定に利用する．例えば，未定義語が「FinePix」，ノードが「カメラ」である場合，ノード「カメラ」のノード動詞である「で撮影」を連結した「FinePix で撮影」というキーワードを Web 検索エンジンで検索した時の検索ヒット数を求める．

以下に，ノード動詞の構築方法を示す．

○ステップ 1

NTT シソーラスのノードに属しているリーフを全て抜き出す．

○ステップ 2

各リーフをキーワードとして Web 検索エンジンで検索し，検索結果上位 1,000 件の検索結果ページを取得する．そして，検索結果ページ内において当該リーフの直後に出現する「格助詞＋動詞（サ変動詞含む）」を全て抜き出す．

○ステップ 3

ステップ 2 の処理を全てのノードに対して行う．

○ステップ 4

ステップ 3 の処理で得られた「格助詞＋動詞」に対して，$tf \cdot idf$ を利用して重みを求める．tf は「格助詞＋動詞」の出現頻度から算出する．また，idf は全てのノードの数と「格助詞＋動詞」が出現するノードの数から算出する．そして，算出した tf と idf を掛け合わせて重みを算出し，最も大きな重みを持つ「格助詞＋動詞」をノード動詞に決定する．

表 3.3: ノード動詞の一例

ノード	ノード動詞
茶	を飲む
カメラ	で撮影
映画・映像	を鑑賞
レンズ	を研磨
星	を観察

構築したノード動詞の一例を表 3.3 に示す.

3.4.2 共起ヒット

「単語の意味は，どのような単語と共起するかという観点から特徴づけられる」という Harris の分布仮説 [57] から，関連のある二語はある文書に共に出現するという関係を持つと考えられる．共起ヒットとはこの関係を利用して設定した検索キーワードのことである．

共起ヒットを利用する場合，未定義語と NTT シソーラスのノード名の And 検索を Web 検索エンジンで行い，検索ヒット数を獲得する．そして，獲得した検索ヒット数を未定義語が所属するべきノードの決定に利用する．例えば，未定義語が「FinePix」，ノードが「カメラ」である場合,「FinePix」と「カメラ」の And 検索を行ったときの検索ヒット数を求める．

3.4.3 評価結果

検索ヒット数を考慮して未定義語を分類するべきノードを決定する計算式を以下に示す．検索ヒット数を考慮する場合，未定義語 U が所属するべきノードとしてノード N_a の中で最もノード得点 $NodeValue(U, N_a)$ が高いノードを出力する．ここで，$DoA(U, N_a)$ は未定義語 U とノード N_a の関連度，$VerbHit(U, N_a)$ は未定義語 U にノード N_a のノード動詞を連結したキーワードを Web 検索エンジンで検索した時の検索ヒット数，$CoincidenceHit(U, N_a)$ は未定義語 U とノード N_a の And 検索を Web 検索エンジンで行ったときの検索ヒット数を示す．なお，log の計算は検索ヒット数が 3 件以上のときに実施し，2 件以下の場合は 1 としている．

$$
\begin{aligned}
&NodeValue(U, N_a) = \\
&DoA(U, N_a) \cdot \log(VerbHit(U, N_a)) \cdot \log(CoincidenceHit(U, N_a))
\end{aligned}
$$

本評価では，3.3.2 項と同様，ノード得点が高い順に 10 位のノードまで出力を行った．検索ヒット数を考慮した場合の評価結果を図 3.2 に示す.

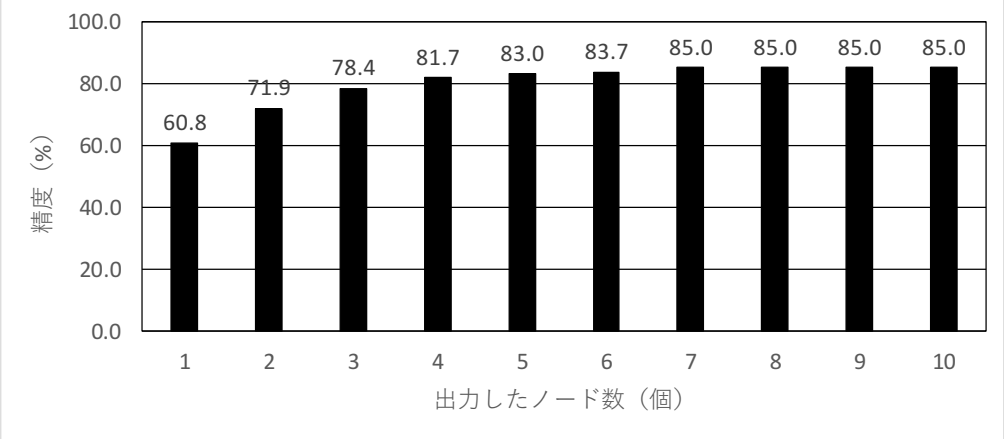


図 3.2: 検索ヒット数を考慮した場合の精度評価結果（文書データベースを用いた固有名詞の意味推定法）

図 3.1 と図 3.2 を比較することで，ノード動詞と共起ヒットを利用することで，5〜20%程度精度が改善していることが分かる．特に，第 1 位のノードに正解ノードを出力できる精度が 15%以上改善し 60%を超えており，検索ヒット数を考慮する方法の有効性を示すことができた．一例として，未定義語「FinePix」と「スターウォーズ」に対して検索ヒット数を考慮した場合の評価結果を表 3.4 に示す．

表 3.4: 検索ヒット数を考慮した場合の評価結果一例（文書データベースを用いた固有名詞の意味推定法）

未定義語（正解ノード）	ノード	関連度	ノード動詞		共起ヒット		ノード得点
			検索キーワード	検索ヒット数	検索キーワード[※]	検索ヒット数	
FinePix（カメラ）	カメラ	0.074	FinePixで撮影	1,110	FinePix & カメラ	1,890,000	7.500
	レンズ	0.081	FinePixを研磨	0	FinePix & レンズ	784,000	1.010
スターウォーズ（映画・映像）	映画・映像	0.050	スターウォーズを鑑賞	8,080	スターウォーズ & 映画OR映像	21,700,000	7.600
	星	0.090	スターウォーズを観察	0	スターウォーズ & 星	8,320,000	1.434

[※]ノード名に含まれる「・」は「OR」に変換

表 3.4 より，未定義語「FinePix」と「スターウォーズ」のいずれの場合であっても，正解ノード（カメラ，映画・映像）ではないノード（レンズ，星）のほうが高い関連度を得ているが，ノード動詞と共起ヒットを利用することで正解ノードのほうが高いノード得点を獲得できており，正解ノードの出力に成功していることが分かる．

3.5 まとめ

本章では，固有名詞の意味推定法として，Wikipedia を情報源として構築した文書データベースを用いて未定義語を概念化し，概念化した未定義語が所属するべきシソーラスのノードを提示する方法を提案した．未定義語として固有名詞を取り上げ，評価実験を行った結果，第 1 位のノードのみを出力した場合の精度が 43.1%であり，第 10 位のノードまで出力した場合の精度が 80%を超えることを示した．この結果より，提案方法が未定義語とある程度関連があるノードを出力できているといえる．さらに，提案方法に検索ヒット数を考慮する方法を組み込むことで，第 1 位のノードに正解ノードを出力できる精度が 60%を超えることを示し，その有効性を確認した．

本章で提案した方法により，未定義語（固有名詞）の意味を理解しやすくできるようになるため，機械が言語を理解する能力を実現するための一助となることが期待できる．

第4章　文書データベースを用いた英字略語の意味推定法

4.1　はじめに

本章では，文書データベースを用いて英字略語の意味を推定する方法について取り上げる．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，今まで以上にグローバル化が進展すると共に，外来語がますます増加する中，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ことも増えてきている．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，まったく別の意味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」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」という意味がある．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，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つ．そのため，英字略語が利用されている情報は，全ての人が容易に，また，正確に把握できるとは言い難い．

例えば，新聞記事などでは記事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理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がとられていないなど，完全に対処されているわけではない．また，記事中の最初の箇所にのみ上記のような処置が取られており，それ以降はその意味が併記されていないことが多い．そのため，記事の途中から文書を読んだり，関連する記事が複数のページにわたって掲載されている時に先頭のページではない部分から記事を読んだりした場合には，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．結果として，解読には一手間が求められ，記事の理解を妨げることになる．さらに，一般的な文書の場合では，このような英字略語の意味を併記する処置がされている方が珍しいと言える．他には，インターネットなどを利用して英字略語の意味を調べることも可能であるが，上述の通り，英字略語は多義性を持つことが多く，所望の結果が得られるまでにそれなりに時間を要することがある．

そこで，本章では，基準となる言語資源である概念ベースに加えて，文書データベースとWikipedia を言語資源として活用することで，多義性を有する英字略語に対して意味の推定を行う方法を提案する．

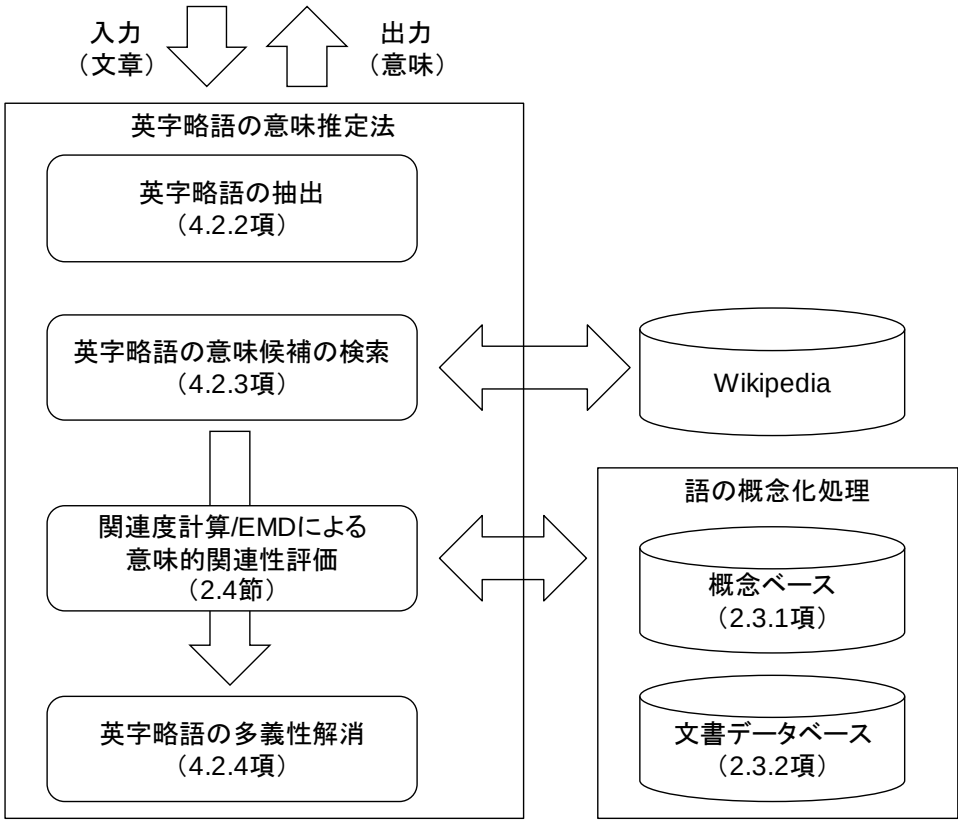


図 4.1: 文書データベースを用いた英字略語の意味推定法の概略図

4.2　提案方法

4.2.1　概要

図 4.1 に文書データベースを用いた英字略語の意味推定法の概略図を示す.

英字略語が含まれる文章を入力として，入力文章から英字略語を抽出する．当該英字略語を 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当該英字略語の意味として出力する．この際，Wikipedia から検索された意味と英字略語（が含まれる文章）の意味的な近さを判断するために，概念ベースと文書データベースを活用することで語の概念化を行う．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，英字略語のもと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.

4.2.2 英字略語の抽出

本章で提案する文書データベースを用いた英字略語の意味推定法の処理対象として扱う英字略語は，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．提案方法は，語彙の意味に着目し，多義性を有する英字略語の意味推定を目的としているため，こういった英字略語は処理対象から除外している．また，英字略語は大文字のアルファベットで構成されることが多いため，本章では，2 文字以上の大文字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 文字以上の大文字アルファベットの羅列を英字略語として抽出する．

4.2.3 Wikipedia による意味候補の検索

4.2.2 項で抽出した英字略語を Wikipedia[21] で検索する．検索の結果，当該英字略語を説明する意味が 1 つであった場合には，その意味を出力する．意味が複数ある場合には，4.2.4 項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．Wikipedia から取得した意味の概念化には，2.3.1 項で述べた概念ベースを用いた概念化と 2.3.2 項で述べた文書データベースを用いた概念化を用いる．一例として，英字略語「IC」を Wikipedia で検索した際の結果を表 4.1 に示す．表 4.1 に示した通り，英字略語「IC」は 14 種類の意味を有している．

Wikipedia から出力した意味には，これらの意味を説明・補助する情報（例えば，表 4.1 の項番 1 における「-」以降の記述）が含まれることがあるが，当該情報は意味の概念化処理において雑音になりえる．そこで，Wikipedia から抽出した意味に対して，表 4.2 に示した規則を上から順に適用することで雑音を削除する．さらに，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アルファベットの羅列（表 4.1 の例では項番 10 と項番 11）が含まれることもあるため，表 4.3 に示したストップワードを適用することで，不要な意味を削除した上で意味候補を取得する．

4.2.4 英字略語の多義性解消

4.2.3 項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，4.2.3 項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全て抽出し，これらの自立語を英字略語の一次属性と見立てる．この処理により，英字略語を擬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登録されていない語（未定義語）が存在する．未定義語については，2.3.2 項で述べた文書データベースを用いた概念化を行う．

表 4.1: 英字略語「IC」を Wikipedia で検索した際の結果

項番	検索結果（英字略語「IC」の意味一覧）
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	インターシティ (ドイツ) - そのうちドイツにおける都市間列車について
9	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)
10	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）
11	イリノイ・セントラル鉄道の報告記号 (Illinois Central railroad)
12	間質性膀胱炎 (Interstitial cystitis) の略
13	インターコンチネンタル (Inter Continental) の略
14	アイドルカレッジ　（IDOL COLLEGE）の略

表 4.2: Wikipedia から英字略語の意味候補を取得する規則

項番	規則
1	意味候補にストップワード（表 4.3 参照）が含まれる場合，意味候補から除外
2	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
3	意味候補内の「など」を削除
4	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
5	意味候補内の「の一つ、」より後ろの語を取得

表 4.3: Wikipedia から英字略語の意味候補を取得する際のストップワード

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

概念化を行うことで，英字略語は概念となるため，一次，二次へと属性を展開できるようになり，意味候補との間の意味的な近さを評価することが可能になる．英字略語を擬似的に概念化する際，一次属性として抽出した自立語が未定義語であった場合，当該一次属性に対する重みは，$tf \cdot idf$ [41, 42] の考え方を応用して算出する．

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 $tfreq(Word_A, A)$ を文章 A 中の全ての自立語の語数 $tnum_A$ で割ることで算出する．算出式は以下の通りである．

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

語の特定性である idf 値は，$SWeb\text{-}idf$ と類似した方法である Statics Article-Inverse Document Frequency（$SA\text{-}idf$）を用いて算出する．$SA\text{-}idf$ を用いる理由は，擬似的な全文章空間の情報として，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ためである．$SA\text{-}idf$ 値の算出式は以下のように定義される．ここで，$N_{SA\text{-}idf}$ は利用する文章集合の全文章数，$df(Word_A)_{SA\text{-}idf}$ はその文章集合の中で自立語 $Word_A$ が出現する文章数である．

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

以上に示した式から算出した tf 値と $SA\text{-}idf$ 値を掛け合わせることで，英字略語の概念化処理において未定義語であった自立語（一次属性）に重みを付与する．

これまでの処理によって概念化された英字略語と意味候補との意味的な近さを 2.4 節で説明した意味的関連性評価方法を用いて評価する．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して出力する．

4.3 評価実験

4.3.1 実験条件

本章では，新聞記事から英字略語を抽出し，当該英字略語が含まれている記事を入力文章とすることで評価を実施した．今回使用した新聞記事は全国紙 1 か月分（約 12,000 記事）であり，2 文字以上の大文字アルファベットの羅列が含まれる記事は約 3,700 記事であった．この約 3,700 記事から表 4.3 に示したストップワードに該当する略語として意味がない文字列を含む記事を人手で削除した上で無作為に 121 記事を抽出し，評価実験データとして使用した．なお，その中で，表記が異なる英字略語の数は 56 個であった．つまり，121 種類の英字略語の意味と 56 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，56 個の英字略語の表記を Wikipedia で検索し，英字略語を説明する意味を出力した．得られた意味の中から文書データベースに存在する意味を抽出した上で，表 4.2 に示した規則と表 4.3 に示したストップワードを適用した結果，453 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 8.1 個，最少で 2 個，最多で 23 個の意味が存在）．

本章の提案方法である文書データベースを用いた英字略語の意味推定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，意味の一致に関する判定は人手で実施しており，Wikipedia から取得した意味候補の中に正解となる候補が複数含まれることもある（例えば，表 4.1 における項番 7 と項番 8 の「インターシティ」は両方ともヨーロッパにおける都市間列車を指しており，どちらを選択しても正解と判定している）．今回は評価の簡略化のため，英字略語として扱う 2 文字以上の大文字アルファベットの羅列が 1 種類のみ含まれる記事を評価対象としている．

4.2.4 項で述べた通り，入力した文章を用いて英字略語を擬似的に概念化する際には，$tf \cdot idf$ の考え方に基づいて属性に重み付けを行う．今回の評価実験における入力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使用した．この 1ヶ月分の新聞記事集合から，概念ベースの収録語数である約 9 万語を超える単語数が得られたことから，当該集合を擬似的な全文章空間の情報とみなしている．

4.3.2　既存方法との比較

本項では，本章における提案方法である文書データベースを用いた英字略語の意味推定法との比較に使用するベクトル空間モデル [8] について述べる．

ベクトル空間モデルは，情報検索の分野で幅広く利用されている検索モデルである．各語の重みから構成されるベクトルとして入力語と文書をそれぞれ表現し，2 つのベクトルの成す角度の余弦によって類似度を計算する点に特徴がある．ベクトル空間モデルにおいて使用される重みにはいくつかの種類があるが，本評価では，情報検索の分野で広く用いられている $tf \cdot idf$ [41, 42] を使用する．4.2.3 項で概念化された英字略語を含む入力文章を q，同様に概念化された意味候補を d_i，両者における自立語の総数（異なり）を M とすれば，入力文章 q と意味候補 d_i はそれぞれ以下のような M 次元ベクトルで表現できる．

$$\begin{aligned} q &= (w_{q_1}, w_{q_2}, \cdots, w_{q_M}) \\ d_i &= (w_{i_1}, w_{i_2}, \cdots, w_{i_M}) \end{aligned}$$

入力文章 q に対する意味候補 d_i の得点 $s_q(d_i)$ は，以下の式に示した通り，2 つのベクトルの余弦により求めることができる．

$$s_q(d_i) = \frac{\sum_{j=1}^{M} w_{i_j} w_{q_j}}{\sqrt{\sum_{j=1}^{M} w_{i_j}^2} \sqrt{\sum_{j=1}^{M} w_{q_j}^2}}$$

以上の処理を実施した結果，最も高い得点を得られた意味候補がベクトル空間モデルを使用した場合の出力結果となる．

4.3.3　評価結果

本章では，文書データベースを用いた英字略語の意味推定法に対して，以下に示す 3 種類の入力文章（A，B，C）および 3 種類の関連性評価方法（1，2，3）を用いて評価を実施した（つまり，合計 9 パターンの評価を実施）．

評価方法（入力文章）

- (A) 英字略語を含む一段落
- (B) 英字略語を含む一文とその前後一文
- (C) 英字略語を含む一文のみ

評価方法（関連性評価方法）

- (1) 関連度計算（関連度）
- (2)Earth Mover's Distance（EMD）
- (3) ベクトル空間モデル（VSM）

評価結果を図 4.2，図 4.3，図 4.4 に示す.

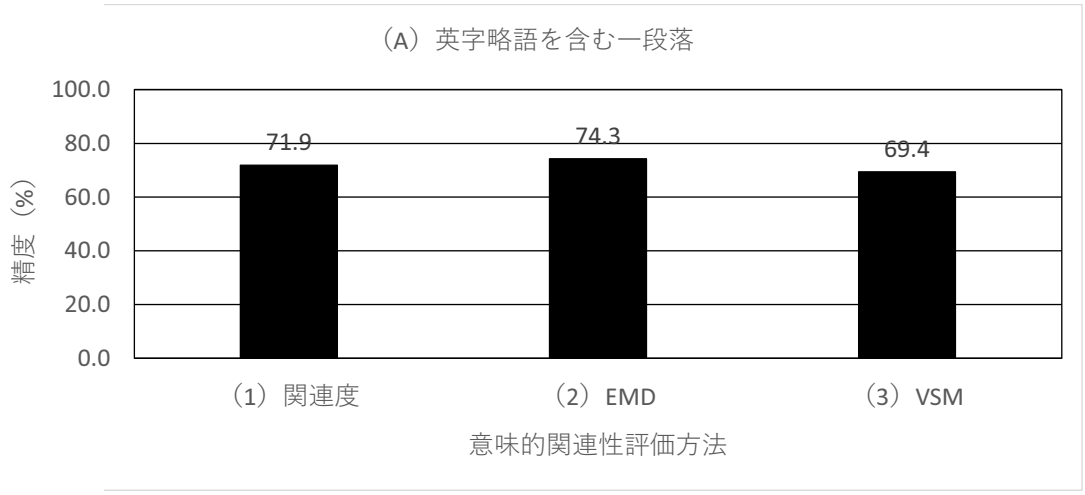


図 4.2: 評価結果その 1（文書データベースを用いた英字略語の意味推定法）

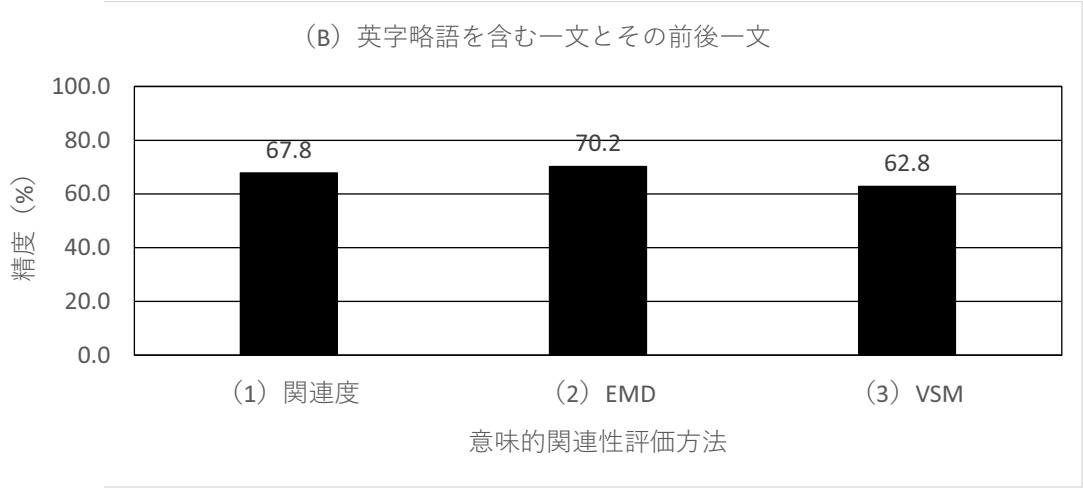


図 4.3: 評価結果その 2（文書データベースを用いた英字略語の意味推定法）

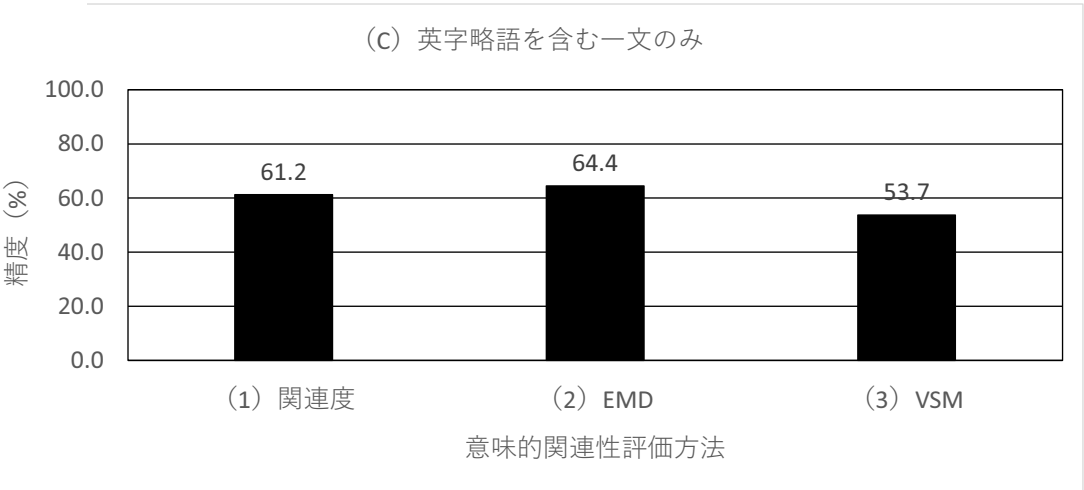


図 4.4: 評価結果その 3（文書データベースを用いた英字略語の意味推定法）

図 4.2，図 4.3，図 4.4 から，全体的な傾向として，関連度計算はベクトル空間モデルより高い精度（2.5%-7.5%）を示した．また，EMD はベクトル空間モデルより精度が高い（4.9%-10.7%）ことに加え，関連度計算よりも若干ではあるが高い精度を獲得した（2.4%-3.2%）．他には，入力文章が長い（入力情報量が多い）ほど高い精度を得ていることが分かる．文書データベースを用いた英字略語の意味推定結果の一例を表 4.4，表 4.5，表 4.6 に示す．

まず，表 4.4 では，英字略語 ACL に対して，Wikipedia から取得した 5 個の意味候補から，9 つの全てのパターンにおいて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，表 4.5 では，英字略語 IMF に対して，4 個の意味候補から，正しい意味を推定できたパターンと推定に失敗したパターンがある．具体的には，ベクトル空間モデルを使用した場合（パターン (A)-(3)，パターン (B)-(3)，パターン (C)-(3)），入力文章の種別によらず意味の推定に失敗している．また，関連度計算を使用し，かつ，入力文章に英字略語を含む一文のみを使用した場合（パターン (C)-(1)）も意味の推定に失敗している．前者（パターン (A)-(3)，パターン (B)-(3)，パターン (C)-(3)）は，単語の表記をもとに関連性を判断するベクトル空間モデルより，単語の意味を考慮して関連性を判断する関連度計算や EMD は有効に機能したためだと考えられる．後者（パターン (C)-(1)）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の関連性の高い属性を除外する可能性がある関連度計算より，全ての属性を計算に利用する EMD が有効に機能したと考えられる．

最後に，表 4.6 では，英字略語 EV に対して，7 個の意味候補から，9 つの全てのパターンにおいて正しい意味を推定することができなかった．これは，入力文章に会社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「エレベーター」より会社に関連する意味候補である「企業価値」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．

表 4.4: 英字略語の意味推定結果一例その 1（文書データベースを用いた英字略語の意味推定法）

英字略語	英字略語の意味	意味候補数	入力文章	入力文章種別	意味的関連性評価方法	推定結果	判定
ACL	AFC チャンピオンズリーグ（AFC Champions League）	5	浦和が今季で契約を満了する元日本代表田中達也（29）と来季の契約を見送る方針であることが18日、分かった。 今季のリーグ戦出場はわずか6試合と出場機会が減り、来季の構想外となった。 クラブ側にとっても苦渋の決断だった。 田中達は帝京高から01年に加入、浦和一筋で通算232戦56点と長くエースに君臨。 06年リーグ初優勝、07年**ACL**初制覇などに貢献し、サポーターの支持も絶大だった。 05年10月には試合中に右足関節脱臼骨折、全治6カ月の重傷を負いながら奇跡のカムバックを遂げ、06年のリーグ優勝に貢献した。 浦和はその功労者でもある田中達の将来を考え、他クラブへ移籍しやすい「戦力外扱い」とする決断に至った。	(A)一段落	(1)関連度	AFC チャンピオンズリーグ	○
					(2)EMD	AFC チャンピオンズリーグ	○
					(3)VSM	AFC チャンピオンズリーグ	○
			田中達は帝京高から01年に加入、浦和一筋で通算232戦56点と長くエースに君臨。 06年リーグ初優勝、07年**ACL**初制覇などに貢献し、サポーターの支持も絶大だった。 05年10月には試合中に右足関節脱臼骨折、全治6カ月の重傷を負いながら奇跡のカムバックを遂げ、06年のリーグ優勝に貢献した。	(B)前後一文	(1)関連度	AFC チャンピオンズリーグ	○
					(2)EMD	AFC チャンピオンズリーグ	○
					(3)VSM	AFC チャンピオンズリーグ	○
			06年リーグ初優勝、07年**ACL**初制覇などに貢献し、サポーターの支持も絶大だった。	(C)一文のみ	(1)関連度	AFC チャンピオンズリーグ	○
					(2)EMD	AFC チャンピオンズリーグ	○
					(3)VSM	AFC チャンピオンズリーグ	○

表 4.5: 英字略語の意味推定結果一例その 2（文書データベースを用いた英字略語の意味推定法）

英字略語	英字略語の意味	意味候補数	入力文章	入力文章種別	意味的関連性評価方法	推定結果	判定
IMF	国際通貨基金（International Monetary Fund）	4	財政危機の発端は、不良債権を抱えた主要銀行の経営悪化である。 政府は巨額の公的資金を投入して銀行を救済した。 その結果、財政赤字が膨らみ、今年の赤字は国内総生産比で32%に悪化する見通しだ。 ユーロ加盟国に義務づけられた3%の財政規律の10倍にも達する異常事態である。 金融市場では、アイルランド国債売りが加速し、金利が急上昇した。 財政赤字を抱えたポルトガルやスペインなどの国債も連鎖して売られ、金利が上昇している。 アイルランドに対する支援策の焦点は、ギリシャ危機の際にIMFが創設した総額7500億ユーロ（約85兆円）の緊急融資制度の活用だ。 アイルランドへの支援額が900億ユーロ（約10兆円）に膨らむ観測も浮上している。 ドイツやフランスが主導して詳細な対策を決め、市場を安心させるべきだ。	（A）一段落	（1）関連度	国際通貨基金	○
					（2）EMD	国際通貨基金	○
					（3）VSM	国際金属労連	×
			財政赤字を抱えたポルトガルやスペインなどの国債も連鎖して売られ、金利が上昇している。 アイルランドに対する支援策の焦点は、ギリシャ危機の際にIMFが創設した総額7500億ユーロ（約85兆円）の緊急融資制度の活用だ。 アイルランドへの支援額が900億ユーロ（約10兆円）に膨らむ観測も浮上している。	（B）前後一文	（1）関連度	国際通貨基金	○
					（2）EMD	国際通貨基金	○
					（3）VSM	インディーズムービー・フェスティバル	×
			アイルランドに対する支援策の焦点は、ギリシャ危機の際にIMFが創設した総額7500億ユーロ（約85兆円）の緊急融資制度の活用だ。	（C）一文のみ	（1）関連度	初期質量関数	×
					（2）EMD	国際通貨基金	○
					（3）VSM	インディーズムービー・フェスティバル	×

表 4.6: 英字略語の意味推定結果一例その 3（文書データベースを用いた英字略語の意味推定法）

英字略語	英字略語の意味	意味候補数	入力文章	入力文章種別	意味的関連性評価方法	推定結果	判定
EV	エレベーター（EleVator）	7	東京都新宿区信濃町の「帝都典礼」本社ビル（地上5階、地下1階）で09年2月、**EV**のかごが来ていない状態で扉が開き、そば店経営、塚田敏雄さん（当時74歳）が転落死した事故で、警視庁捜査1課は20日、ビルの安全管理を担当していた当時の同社常務（60）と、保守点検会社の契約社員（63）を業務上過失致死容疑で書類送検した。 契約社員の送検容疑は、03年1月と07年6月、点検作業で扉の留め金が摩耗し、かごが到着する前に扉が開く故障に気づいたのに部品交換などをせず、そのまま放置したとしている。 元常務は故障情報をきちんと把握しなかったほか、荷物用だったにもかかわらず、社員や出入り業者に使用させていたとしている。 同課によると、契約社員は「故障は分かっていたが、扉の留め金の調整をすれば大丈夫と思った」と供述、元常務は「荷物搬送以外で使わせてはいけないことは分かっていた」と容疑を認めている。 扉は手動タイプ。 塚田さんは09年2月16日午前11時半ごろ、1階で「呼び出しボタン」を押し、扉を開けて乗ろうとしたところ、かごが来ておらず、約4メートル下の地下1階に転落した。 帝都典礼の担当者は「書類送検されたことは真摯（しんし）に受け止める」と話した。	（A）一段落	（1）関連度	企業価値	×
					（2）EMD	企業価値	×
					（3）VSM	企業価値	×
			東京都新宿区信濃町の「帝都典礼」本社ビル（地上5階、地下1階）で09年2月、**EV**のかごが来ていない状態で扉が開き、そば店経営、塚田敏雄さん（当時74歳）が転落死した事故で、警視庁捜査1課は20日、ビルの安全管理を担当していた当時の同社常務（60）と、保守点検会社の契約社員（63）を業務上過失致死容疑で書類送検した。 契約社員の送検容疑は、03年1月と07年6月、点検作業で扉の留め金が摩耗し、かごが到着する前に扉が開く故障に気づいたのに部品交換などをせず、そのまま放置したとしている。	（B）前後一文	（1）関連度	企業価値	×
					（2）EMD	企業価値	×
					（3）VSM	企業価値	×
			東京都新宿区信濃町の「帝都典礼」本社ビル（地上5階、地下1階）で09年2月、**EV**のかごが来ていない状態で扉が開き、そば店経営、塚田敏雄さん（当時74歳）が転落死した事故で、警視庁捜査1課は20日、ビルの安全管理を担当していた当時の同社常務（60）と、保守点検会社の契約社員（63）を業務上過失致死容疑で書類送検した。	（C）一文のみ	（1）関連度	企業価値	×
					（2）EMD	企業価値	×
					（3）VSM	企業価値	×

Wikipedia は現存するインターネット百科事典の中で収録語数が最も多いとされるが，本章の提案方法である文書データベースを用いた英字略語の意味推定法は，処理の拠り所である事典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．これは，事典を使用する以上，避けられない問題である．ただし，評価実験に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 4.5%であった．よって，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照とする場合，Wikipeida に未登録であることに起因する精度の低下は 5%程度であると考えられる．なお，今回実施した評価実験に関しては，Wikipedia に登録されていない英字略語は評価実験データから除外している．

4.4 まとめ

本章では，英字略語の意味推定法として，Wikipedia を情報源として構築した文書データベースを用いて多義性を有する英字略語に対して意味の推定を行う方法を提案した．新聞記事から抽出した英字略語に対して複数パターンの評価実験を行った結果，提案方法は英字略語の意味推定に成功する精度が最高で 70%を超えており，提案方法が英字略語の多義性をある程度解消できていることを示した．さらに，提案方法は比較方法であるベクトル空間モデルよりも良好な結果を示したことから，その有効性を確認することができた．

本章で提案した方法により，多義性を有する未定義語（英字略語）の意味を理解しやすくできるようになるため，機械が言語を理解する能力を実現するための一助となることが期待できる．

第5章　オートフィードバックを用いた固有名詞の意味推定法

5.1　はじめに

本章では，オートフィードバックを用いて固有名詞の意味を推定する方法について取り上げる [56]．第 3 章において文章データベースを用いて未定義語（固有名詞）を概念化し，概念化した未定義語が所属するべき NTT シソーラスのノードを提示することで，未定義語の意味推定を行う方法について述べた．しかし，本提案方法を適用できる未定義語は Wikipedia から構築した文書データベースに存在する語に限定される．

そこで，第 3 章で用いた言語資源である概念ベースと NTT シソーラスに加えて，インターネット情報を言語資源として活用（オートフィードバックを用いた概念化を活用）することで，ほぼあらゆる未定義語に対して意味推定を行うことができる方法を提案する．なお，本章においても，第 3 章と同様，未定義語の代表的な存在である固有名詞を対象として意味推定を行う．

5.2　提案方法

5.2.1　概要

図 5.1 にオートフィードバックを用いた固有名詞の意味推定法の概略図を示す．

まず，入力された未定義語に対して，オートフィードバックを用いて概念化を行う．あわせて，NTT シソーラスのノードの概念化も行う．そして，概念化された未定義語と各ノードに対して関連度計算を用いて意味的な近さを判断し，所属候補となるノード（所属候補ノード）の絞込みを行う．最後に，絞り込んだ所属候補ノードに対して検索ヒット数を活用するノード動詞と共起ヒットを用いて入力された未定義語が所属するべきノードを決定する．

5.2.2　未定義語の概念化及びシソーラスのノードの概念化

入力された未定義語に対して，オートフィードバックを用いて概念化を行う（2.3.3 項参照）．ここで，概念化された未定義語 U を l 個の属性 u_i と重み w_{u_i} を用いて，以下のように表現する．

$$\text{未定義語 } U = \{(u_1, w_{u_1}), (u_2, w_{u_2}), \cdots, (u_l, w_{u_l})\}$$

次に，NTT シソーラスのノードの概念化を行う（2.5.2 項参照）．同様に，概念化されたノード N_a を m 個の属性 n_{a_i} と重み $w_{n_{a_i}}$ を用いて，以下のように表現する．

$$\text{ノード } N_a = \{(n_{a_1}, w_{n_{a_1}}), (n_{a_2}, w_{n_{a_2}}), \cdots, (n_{a_m}, w_{n_{a_m}})\}$$

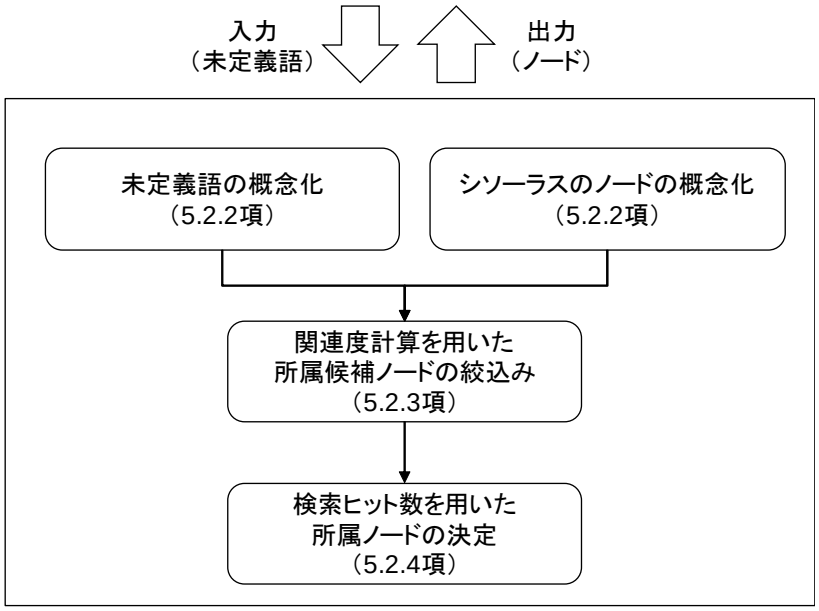


図 5.1: オートフィードバックを用いた固有名詞の意味推定法の概略図

5.2.3　関連度計算を用いた所属候補ノードの絞り込み

5.2.2 節において概念化を行った未定義語 U と，概念化を行った NTT シソーラスの各ノード N_a に対して関連度計算を実施し，関連度 $DoA(U, N_a)$ を求める．そして，0.02 以上の関連度を持つノードを所属候補ノードとして抽出する．なお，関連度の閾値 0.02 は，閾値を 0.0 から 0.05 まで 0.001 毎に変化させて評価実験を行った結果，最も高い精度が得られた値を採用したものである．当該実験については，5.3.1 項で述べる．

本提案方法では，第 3 章で提案した文書データベースを用いた固有名詞の意味推定法と同様，385 個の最下位ノードを使用するが，上述した閾値を用いた絞込みにより所属候補ノード数を 385 個から 10 個程度に絞り込むことが可能となる．その結果，後述する検索ヒット数を用いた所属ノードの決定処理において，処理回数を 20 分の 1 以下に削減することに成功している．

5.2.4　検索ヒット数を用いた所属ノードの決定

5.2.3 節の処理で絞り込んだ所属候補ノードに対して検索ヒット数を用いて未定義語が所属するノードの決定を行う．検索ヒット数を利用する方法としてノード動詞と共起ヒットを使用する（3.4.1 項及び 3.4.2 項を参照）．

未定義語が所属するノードを決定する計算式を以下に示す．未定義語 U が所属するべきノードとしてノード N_a の中で最もノード得点 $NodeValue(U, N_a)$ が高いノードを出力する．ここで，$DoA(U, N_a)$ は未定義語 U とノード N_a の関連度，$VerbHit(U, N_a)$ は未定義語 U にノー

ド N_a のノード動詞を連結したキーワードを Web 検索エンジンで検索した時の検索ヒット数，$CoincidenceHit(U, N_a)$ は未定義語 U とノード N_a の And 検索を Web 検索エンジンで行ったときの検索ヒット数を示す．なお，log の計算は検索ヒット数が 3 件以上のときに実施し，2 件以下の場合は 1 としている．

$$NodeValue(U, N_a) =$$
$$DoA(U, N_a) \cdot \log(VerbHit(U, N_a)) \cdot \log(CoincidenceHit(U, N_a))$$

以下に未定義語「G ショック」及び「クイニーアマン」を例に，所属ノードを決定する処理におけるノード得点上位 5 個までの計算過程を表 5.1，表 5.2，表 5.3，表 5.4，表 5.5，表 5.6 に示す．表 5.1 が未定義語とノード得点上位 5 個のノードとの関連度，表 5.2 と表 5.3 がノード得点上位 5 個のノードが持つノード動詞とノード動詞を用いたときの検索ヒット数，表 5.4 と表 5.5 が共起ヒットを用いたときの検索ヒット数，表 5.6 が未定義語のノード得点上位 5 個のノード得点を表している．

表 5.1: 所属ノード決定処理における計算過程一例（関連度）

G ショック		クイニーアマン	
ノード	関連度	ノード	関連度
時計	0.188	パン	0.196
通信機器	0.085	菓子	0.129
放送局	0.089	調味料	0.147
鉱油	0.053	果物	0.068
石油	0.034	ミルク	0.031

表 5.2: 所属ノード決定処理における計算過程一例（ノード動詞その 1）

G ショック			
ノード	ノード動詞	検索キーワード	検索ヒット数（件）
時計	を購入	G ショックを購入	1,250
通信機器	で通信	G ショックで通信	0
放送局	が放送	G ショックが放送	0
鉱油	を購入	G ショックを購入	1,250
石油	を精製	G ショックを精製	0

表 5.3: 所属ノード決定処理における計算過程一例（ノード動詞その 2）

クイニーアマン			
ノード	ノード動詞	検索キーワード	検索ヒット数（件）
パン	を食べる	クイニーアマンを食べる	3
菓子	を食べる	クイニーアマンを食べる	3
調味料	を加える	クイニーアマンを加える	0
果物	を食べる	クイニーアマンを食べる	3
ミルク	を飲む	クイニーアマンを飲む	0

表 5.4: 所属ノード決定処理における計算過程一例（共起ヒットその 1）

G ショック		
ノード	検索キーワード	ヒット数（件）
時計	G ショック & 時計	210,000
通信機器	G ショック & 通信機器	754
放送局	G ショック & 放送局	536
鉱油	G ショック & 鉱油	0
石油	G ショック & 石油	12,000

表 5.5: 所属ノード決定処理における計算過程一例（共起ヒットその 2）

クイニーアマン		
ノード	検索キーワード	ヒット数（件）
パン	クイニーアマン & パン	947
菓子	クイニーアマン & 菓子	604
調味料	クイニーアマン & 調味料	278
果物	クイニーアマン & 果物	178
ミルク	クイニーアマン & ミルク	9,150

表 5.6: 所属ノード決定処理における計算過程一例（ノード得点）

Gショック		クイニーアマン	
ノード	ノード得点	ノード	ノード得点
時計	16.409	パン	1.478
通信機器	0.564	菓子	0.906
放送局	0.558	調味料	0.829
鉱油	0.379	果物	0.388
石油	0.315	ミルク	0.284

表 5.7: 評価セットの一例（オートフィードバックを用いた固有名詞の意味推定法）

未定義語	所属ノード（正解ノード）
G ショック	時計
クイニーアマン	パン
マイルドセブン	たばこ
アインシュタイン	学者・研究者
新島襄	教師

5.3 評価実験

本章では 3.3.1 項で述べた理由と同様の理由により，未定義語に関する表現として固有名詞を扱う．評価実験に使用する固有名詞として，20 名の被験者から各 10 個ずつ固有名詞と当該固有名詞が所属するべき NTT シソーラスのノードを提供してもらうことで，200 語の評価セットを作成した．なお，多義性を持つ語の意味推定は第 6 章で扱うため，本章で扱う評価セットは一意に意味を判断できる固有名詞から構成されている．作成した評価セットの一例を表 5.7 に示す．

評価セットにおける各未定義語の入力に対して，本章が提案するオートフィードバックを用いて固有名詞の意味を推定する方法を適用して出力した結果として，正解ノードを得た未定義語を正解，正解ノードを得られなかった未定義語を不正解として精度を算出する．

5.3.1 閾値調査

5.2.3 節で述べた所属候補ノードの絞込みにおいて関連度の閾値を決定するために，閾値を 0.0 から 0.05 まで 0.001 毎に変化させて未定義語の所属ノードの決定を行ったときの実験結果を図 5.2 に示す．図 5.2 より，関連度の閾値が 0.014 から 0.02 の間で，最も高い 66.0%の精度が得られている．そこで，最高精度が得られている範囲で所属候補ノードを最も絞り込むことができる 0.02 が関連度の閾値として適当であると考えられる．この値を所属候補ノードの絞込みを行

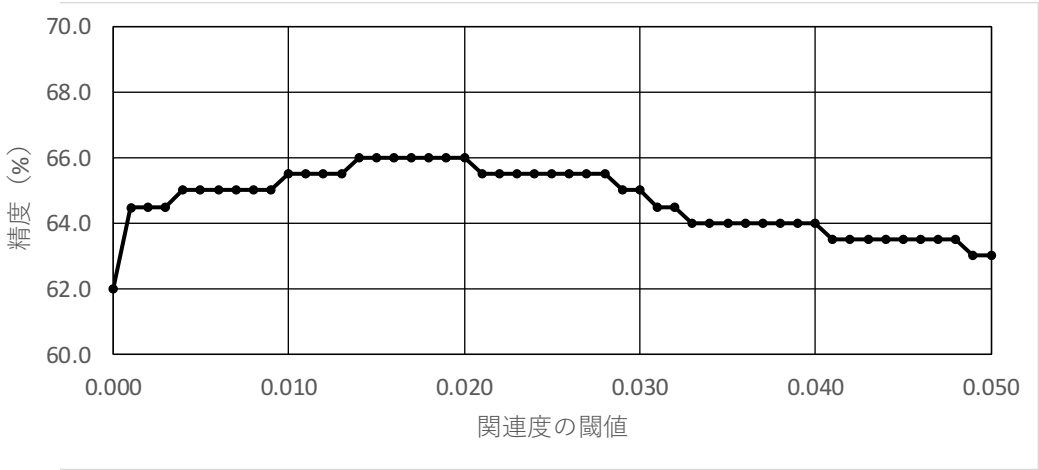


図 5.2: 関連度の閾値と精度

う際に用いる閾値とした．

5.3.2　評価結果

オートフィードバックを用いた固有名詞の意味推定法を用いて，未定義語を分類するべきノードを出力する．なお，本評価では，3.3.2 項と同様，順位をつけて複数のノードを出力する．つまり，5.2.4 項に示した式に基づきノード得点を算出し，最もノード得点が高いノードが 1 位，2 番目にノード得点が多いノードが 2 位となる．本評価では，10 位のノードまで出力を行った．

評価結果を図 5.3 に示す．図 5.3 は出力したノードに正解ノードが含まれる精度を示している．図 5.3 より，第 1 位のみを出力した場合の精度が 66.0%であり，第 10 位まで出力することで 9 割を超える精度が得られていることから，全体的に未定義語と関連度があると考えらえるノードを獲得していると判断できる．また，第 1 位に関連が強いと考えらえるノードが得られた場合，上位のノードに正解ノードが得られる傾向にあった．例えば，正解ノードが「教師」である未定義語「新島襄」を入力した場合，第 1 位に正解ノードである「教師」と関連が強い「教育」が得られ，第 2 位に正解ノードである「教師」が得られた．

5.4　既存方法との比較

本節では，既存方法として，ベクトル空間法に基づく方法について説明する．本節で用いるベクトル空間法にはシソーラスと共起辞書を必要とし，本方法ではシソーラスに NTT シソーラス [5] を使用し，共起辞書に EDR コーパス [58] を使用している．EDR コーパスは 22 万文からなる文章のデータベースであり，係り受け関係にある単語対を抽出した共起辞書を用いている．

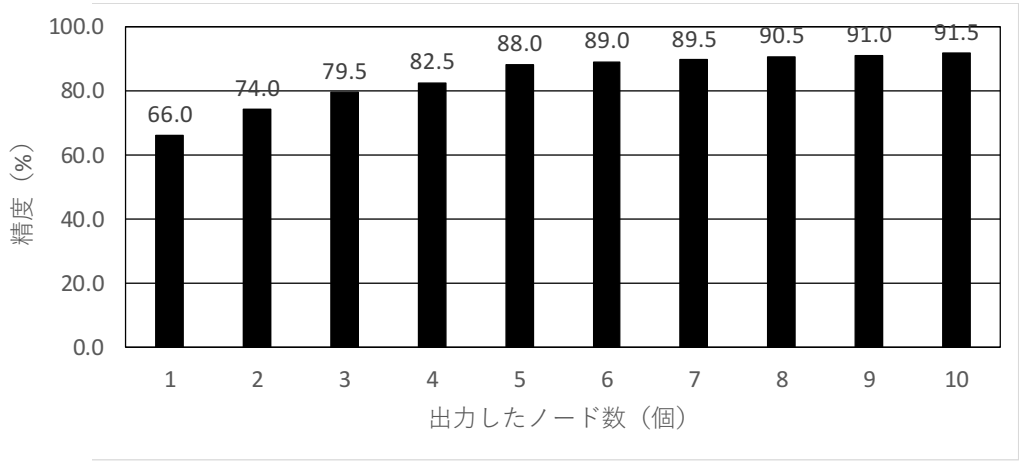


図 5.3: 精度評価結果（オートフィードバックを用いた固有名詞の意味推定法）

5.4.1 ベクトル空間法

ベクトル空間法はシソーラスにおける各ノードの特徴ベクトルと未定義語の特徴ベクトルの類似度をベクトル間の余弦を用いて算出し，類似度の高いノードに未定義語を分類する．最も単純なベクトル空間法では，特徴ベクトルは名詞と動詞との共起頻度によるベクトルである．ノードの特徴ベクトルの各要素は，そのノードに属する名詞と動詞との共起頻度を足し合わせたものである．また，未定義語の特徴ベクトルの各要素は，未定義語と動詞の共起頻度そのものである．以下に，ベクトル空間法の詳細を説明する．

まず，シソーラスに既に分類されている名詞（リーフ）$noun_i$ の集合 $NOUN$，シソーラスのノード $node_j$ の集合 $NODE$，共起を考慮する動詞 $verb_k$ の集合 $VERB$ を以下に定義する．なお，I はシソーラスにおける名詞（リーフ）の数，J はシソーラスにおけるノードの数，K は共起を考慮する動詞の数である．

$$NOUN = \{noun_1, noun_2, \cdots, noun_I\}$$
$$NODE = \{node_1, node_2, \cdots, node_J\}$$
$$VERB = \{verb_1, verb_2, \cdots, verb_K\}$$

次に，ノード w と動詞 z が共起したことを表す 1 つの学習データを以下に定義する．

$$\{(w, z) \,|\, w \in NODE, z \in VERB\}$$

$(w, z)^N$ は N 個の学習データからなる系列である．学習データを生成するために用いる文章の中では，名詞 $noun_i$ と動詞 $verb_k$ が共起しているが，学習データを生成する時点で名詞と動詞の二項組 $(noun_i, verb_k)$ をノードと動詞の二項組 $(node_j, verb_k)$ に変換する．なお，ノード $node_j$ は名詞 $noun_i$ が属するノードであり，複数のノードに属する場合は複数の二項組に変換する．したがって，未定義語 U が属するノード $node^*$ と未定義語 U と共起した動詞 y の系列 y^M は，

以下のように表すことができる．

$$\{(node^*, y^M) \,|\, node^* \in NODE, y \in VERB\}$$

しかし，$node^*$ は未知であり，実際に観測される未定義語に関連する情報（未定義語データ）は未定義語 U と共起した動詞 y の系列 y^M の二項組 (U, y^M) である．よって，ベクトル空間法を用いた未定義語分類問題は学習データ $(w, z)^N$ と未定義語データ (U, y^M) を観測したもとで未定義語 U が属する $node^*$ を推定する問題となる．

上記を考慮すると，ベクトル空間法では以下の式を用いて未定義語 U を分類するノードを決定することができる．これらの式に登場する表現について説明する．$d_{\cos}\{(w, z)^N, (U, y^M)\}$ は学習データ $(w, z)^N$ と未定義語データ (U, y^M) を引数に取り，未定義語 U を分類するべきノードを決定する関数を表す．$vec(node_j)$ はノード $node_j$ の特徴ベクトル，$vec(U)$ は未定義語 U の特徴ベクトルである．$co((node_j, verb_k) \,|\, (w, z)^N)$ は学習データ $(w, z)^N$ 中の $(node_j, verb_k)$ の数，つまり，ノード $node_j$ と動詞 $verb_k$ が共起した回数を表す．同様に，$co(verb_k \,|\, y^M)$ は未定義語データ (U, y^M) の y^M 中の $verb_k$ の数，つまり，未定義語 U と動詞 $verb_k$ が共起した回数を表す．cos はベクトル間の余弦の値を求める関数，$vec_A \cdot vec_B$ はベクトル vec_A と vec_B 間の内積，$||vec||$ はベクトル vec のノルムである．

$$\begin{aligned} d_{\cos}\{(w, z)^N, (U, y^M)\} &= \arg\max_{node_j}\{\cos(vec(node_j), vec(U))\} \\ &= \arg\max_{node_j}\left\{\frac{vec(node_j) \cdot vec(U)}{||vec(node_j)||\,||vec(U)||}\right\} \end{aligned}$$

$$\begin{aligned} vec(node_j) = \quad & \{co((node_j, verb_1) \,|\, (w, z)^N), co((node_j, verb_2) \,|\, (w, z)^N), \cdots, \\ & co((node_j, verb_K) \,|\, (w, z)^N)\} \end{aligned}$$

$$vec(U) = \{co(verb_1) \,|\, y^M), co(verb_2) \,|\, y^M), \cdots, co(verb_K) \,|\, y^M)\}$$

1番目の式より，未定義語 U の特徴ベクトルである $vec(U)$ と余弦の値が最高になる特徴ベクトル $vec(node_j)$ に対応する $node_j$ に未定義語 U を分類していることが分かる．

なお，上記のような単純に共起頻度を用いるベクトル空間法以外に，各共起頻度に重み付けを行う $tf \cdot idf$ を導入したベクトル空間法も提案されている．情報検索などの分野において実用化されている方法は，この $tf \cdot idf$ を導入したベクトル空間法 [59] である．$tf \cdot idf$ を導入したベクトル空間法では，上述の2番目と3番目の式において，特徴ベクトルの第 k 要素に $\log \frac{J}{a(verb_k)}$ を掛け合わせたものを特徴ベクトルとして採用し，その上で1番目の式を用いて未定義語の分類を行う．ただし，$a(verb_k)$ は動詞 $verb_k$ と共起頻度が1以上のノードの数である．

表 5.8: 未定義語と仮定して NTT シソーラスから抽出したリーフの一例

未定義語	分類された所属ノード
装置	道具，機械
義侠心	人情，善悪，同情
侮辱	軽蔑，失礼
オンエア	放送，再生
大みそか	節気，荒れ地，土

5.4.2 比較評価

オートフィードバックを用いた固有名詞の意味推定法に対する比較実験の方法を以下に示す [13].

○ステップ 1
NTT シソーラスに分類されている名詞（リーフ）の中で概念ベースに存在する単語から 1000 語を未定義語と仮定して抽出する.

○ステップ 2
NTT シソーラスに属している残りのリーフと EDR コーパス頻出動詞上位 500 語との共起回数を算出し，学習データを作成する.

○ステップ 3
さらに，NTT シソーラスから取り出しておいた 1000 語の未定義語について，ステップ 2 で作成した学習データと同様に EDR コーパス頻出動詞上位 500 語との共起回数を共起辞書から算出し，1000 個の未定義語データを作成する.

○ステップ 4
作成した学習データと未定義語データをもとに 5.4.1 項で述べたベクトル空間法を用いて，各未定義語を分類するべきノードを出力する.

抽出された未定義語と当該未定義語が分類された所属ノードの例を表 5.8 に示す.
図 5.4 に実験結果を示す．図 5.4 における *Cos* は共起頻度のみを用いたベクトル空間法，$tf \cdot idf$ は $tf \cdot idf$ を用いたベクトル空間法，提案方法がオートフィードバックを用いた固有名詞の意味推定法に対応している.
本実験において，未定義語（未定義語と仮定して NTT シソーラスから抽出したリーフ）が元々所属していたノードに分類できた場合を正解と判定する．また，未定義語が複数のノードに所属していた場合には，分類したノードがその中のどれか 1 つと一致すれば正解と判定している．なお，本実験においても，図 5.3 と同様に，10 位のノードまで出力している.

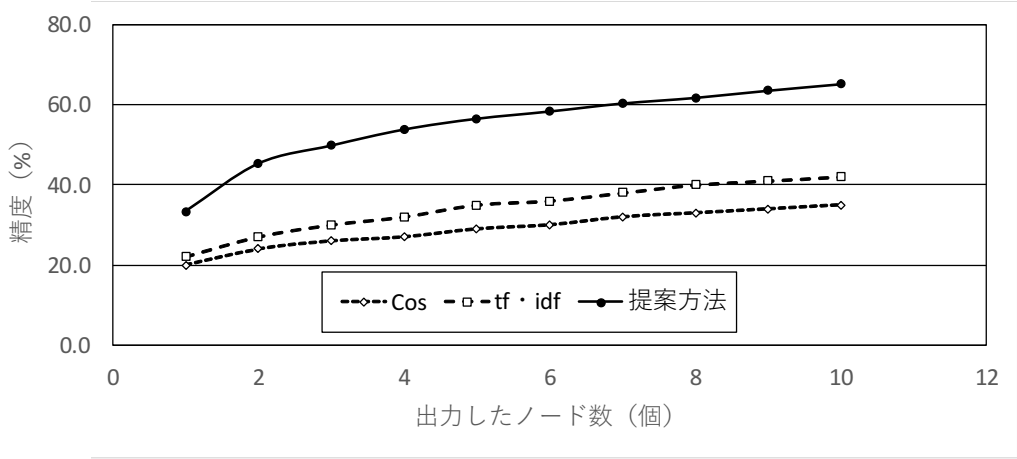


図 5.4: 比較評価結果（オートフィードバックを用いた固有名詞の意味推定法）

図 5.4 より，提案方法の精度は共起頻度によるベクトル空間法（*Cos*）より 13〜30%程度高く，$tf \cdot idf$ を用いたベクトル空間法（$tf \cdot idf$）に対しても 11〜23%程度高くなっており，提案方法がベクトル空間法に基づく方法よりも優れた結果を示していることが分かる．

本来，本章における提案方法であるオートフィードバックを用いた固有名詞の意味推定法は，既存のシソーラスに分類されていない未定義語（固有名詞など）に対して有効な方法である．その一方で，本実験で用いた既存のシソーラス（NTT シソーラス）から抽出した仮想的な未定義語の実体は一般的な単語である．一般的な単語は多くの文書で使用されるため，5.2 節で説明した方法を用いると，様々なページから広範囲な属性を獲得することになる．その結果，獲得できる属性にばらつきが生じるため，適切な属性を獲得することが難しくなる．そのため，本章における提案方法であるオートフィードバックを用いた固有名詞の意味推定法は，本実験に対しては不利な部分があるといえる．

上述の内容を踏まえると，本実験において不利な部分を有しているにも関わらず，オートフィードバックを用いた固有名詞の意味推定法は良好な結果を得ることに成功している．従って，オートフィードバックを用いた固有名詞の意味推定法が未定義語に限らず，一般的な単語に対しても柔軟に機能することを示していると言える．

5.5　まとめ

本章では，固有名詞の意味推定法として，オートフィードバックを用いて未定義語を概念化し，概念化した未定義語が所属するべきシソーラスのノードを提示する方法を提案した．未定義語として固有名詞を取り上げ，評価実験を行った結果，第 1 位のノードのみを出力した場合の精度が 66.0%であり，第 10 位のノードまでを出力した場合の精度が 90%を超えることを示した．この結果より，提案方法が未定義語とある程度関連があるノードを出力できているといえる．さらに，提案方法は比較方法であるベクトル空間法に基づく方法よりも良好な結果を示し

たことから，その有効性を確認した．

本章で提案した方法により，未定義語（固有名詞）の意味を理解しやすくできるようになる．提案方法はインターネット上に登場する日本語（≒ほぼあらゆる日本語）に適用することができるため，機械が言語を理解する能力を実現するための一助となることが期待できる．

第6章　オートフィードバックを用いた英字略語の意味推定法

6.1　はじめに

本章では，オートフィードバックを用いて英字略語の意味を推定する方法について取り上げる [60]．第 4 章において文書データベースを用いて英字略語の意味推定を行う方法について述べた．本提案方法は，文書データベースを用いて Wikipedia から取得した英字略語の意味候補を概念化し，英字略語と概念化した各意味候補の意味的な近さを評価することで，英字略語の多義性を解消した上で英字略語の意味推定を行う方法である．しかし，本提案方法を適用できる英字略語は Wikipedia から構築した文書データベースに存在する語に限定される．

そこで，第 4 章で用いた言語資源である概念ベースと Wikipedia に加えて，インターネット情報を言語資源として活用（オートフィードバックを用いた概念化を活用）することで，より広範な英字略語に対して意味推定を行うことができる方法を提案する．

6.2　提案方法

6.2.1　概要

図 6.1 にオートフィードバックを用いた英字略語の意味推定法の概略図を示す．

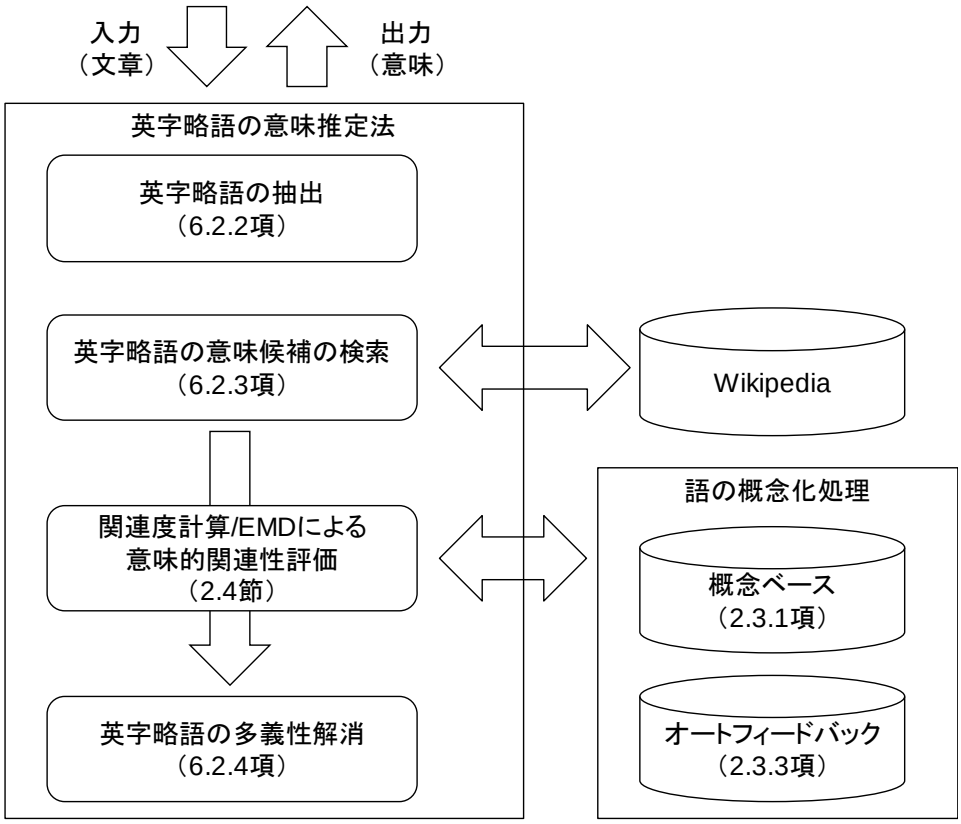


図 6.1: オートフィードバックを用いた英字略語の意味推定法の概略図

英字略語が含まれる文章を入力として，入力文章から英字略語を抽出する．当該英字略語を 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当該英字略語の意味として出力する．この際，Wikipedia から検索された意味と英字略語（が含まれる文章）の意味的な近さを判断するために，概念ベースとオートフィードバックを活用することで語の概念化を行う．

6.2.2　英字略語の抽出

本章で提案するオートフィードバックを用いた英字略語の意味推定法の処理対象として扱う英字略語は，第 4 章と同様，英単語の頭文字から構成される表記とする．提案方法は，語彙の意味に着目し，多義性を有する英字略語の意味推定を目的としているため，2 文字以上の大文字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 文字以上の大文字アルファベットの羅列を英字略語として抽出する．

6.2.3 Wikipedia による意味候補の検索

6.2.2 項で抽出した英字略語を Wikipedia[21] で検索する．検索の結果，当該英字略語を説明する意味が 1 つであった場合には，その意味を出力する．意味が複数ある場合には，6.2.4 項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．Wikipedia から取得した意味の概念化には，2.3.1 項で述べた概念ベースを用いた概念化と 2.3.3 項で述べたオートフィードバックを用いた概念化を用いる．6.2.4 項と 6.3 節に述べるように，本章における英字略語の概念化は新聞記事を用いて実施しているため，Wikipedia から取得した意味候補の概念化も新聞記事を利用して実施したいところであるが，各意味候補が新聞記事に含まれているとは限らない．そこで，新聞記事と同様に，雑多な情報から構成されるインターネット情報を利用して意味候補の概念化を実施する．

Wikipedia から出力した意味候補に対して，4.2.3 項と同様，表 4.2 に示した規則を上から順に適用することで雑音を削除し，表 4.3 に示したストップワードを含む意味候補を除外することで，英字略語の意味候補を取得する．

6.2.4 英字略語の多義性解消

6.2.3 項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，6.2.3 項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全て抽出し，これらの自立語を英字略語の一次属性と見立てる．この処理により，英字略語を擬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登録されていない語（未定義語）が存在する．未定義語については，2.3.3 項で述べたオートフィードバックを用いた概念化を行う．概念化を行うことで，英字略語は概念となるため，一次，二次へと属性を展開できるようになり，意味候補との間の意味的な近さを評価することが可能になる．英字略語を擬似的に概念化する際，一次属性として抽出した自立語が未定義語であった場合，当該一次属性に対する重みは，$tf \cdot idf$ [41, 42] の考え方を応用して算出する．

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 $tfreq(Word_A, A)$ を文章 A 中の全ての自立語の語数 $tnum_A$ で割ることで算出する．算出式は以下の通りである．

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

語の特定性である idf 値は，第 4 章と同様，$SA\text{-}idf$ を用いて算出する．$SA\text{-}idf$ 値の算出式は以下のように定義される．ここで，$N_{SA\text{-}idf}$ は利用する文章集合の全文章数，$df(Word_A)_{SA\text{-}idf}$ はその文章集合の中で自立語 $Word_A$ が出現する文章数である．

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

以上に示した式から算出した tf 値と $SA\text{-}idf$ 値を掛け合わせることで，英字略語の概念化処理において未定義語であった自立語（一次属性）に重みを付与する．

これまでの処理によって概念化された英字略語と意味候補との意味的な近さを 2.4 節で説明した意味的関連性評価方法を用いて評価する．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して出力する．

6.3　評価実験

6.3.1　実験条件

本章では，新聞記事から英字略語を抽出し，当該英字略語が含まれている記事を入力文章とすることで評価を実施した．今回使用した新聞記事は全国紙 1 か月分（約 12,000 記事）であり，2 文字以上の大文字アルファベットの羅列が含まれる記事は約 3,700 記事であった．この約 3,700 記事から表 4.3 に示したストップワードに該当する略語として意味がない文字列を含む記事を人手で削除した上で無作為に 124 記事を抽出し，評価実験データとして使用した．なお，その中で，表記が異なる英字略語の数は 57 個であった．つまり，124 種類の英字略語の意味と 57 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，57 個の英字略語の表記を Wikipedia で検索し，英字略語を説明する意味を出力した．得られた意味に対して，表 4.2 に示した規則と表 4.3 に示したストップワードを適用した結果，696 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 12.2 個，最少で 2 個，最多で 29 個の意味が存在）．

本章の提案方法であるオートフィードバックを用いた英字略語の意味推定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，意味の一致に関する判定は人手で実施しており，Wikipedia から取得した意味候補の中に正解となる候補が複数含まれることもある（例えば，表 4.1 における項番 7 と項番 8 の「インターシティ」は両方ともヨーロッパにおける都市間列車を指しており，どちらを選択しても正解と判定している）．今回は評価の簡略化のため，英字略語として扱う 2 文字以上の大文字アルファベットの羅列が 1 種類のみ含まれる記事を評価対象としている．

6.2.4 項で述べた通り，入力した文章を用いて英字略語を擬似的に概念化する際には，$tf \cdot idf$ の考え方に基づいて属性に重み付けを行う．今回の評価実験における入力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使用した．この 1ヶ月分の新聞記事集合から，概念ベースの収録語数である約 9 万語を超える単語数が得られたことから，当該集合を擬似的な全文章空間の情報とみなしている．

6.3.2　評価結果

本章では，オートフィードバックを用いた英字略語の意味推定法に対して，以下に示す 3 種類の入力文章（A，B，C）および 3 種類の関連性評価方法（1，2，3）を用いて評価を実施した（つまり，合計 9 パターンの評価を実施）．なお，ベクトル空間モデル（VSM）は提案方法であるオートフィードバックを用いた英字略語の意味推定法と比較するために使用している（ベクトル空間モデルについては 4.3.2 項参照）．

評価方法（入力文章）

- (A) 英字略語を含む一段落
- (B) 英字略語を含む一文とその前後一文
- (C) 英字略語を含む一文のみ

評価方法（関連性評価方法）

- (1) 関連度計算（関連度）
- (2)Earth Mover's Distance（EMD）
- (3) ベクトル空間モデル（VSM）

評価結果を図 6.2，図 6.3，図 6.4 に示す．なお，ベクトル空間モデル以外に英字略語の意味推定を行う方法として，[61] や [62] が報告されている．前者は新聞記事などの文章において，英字略語の意味が括弧書きで併記されることに着目し，英字略語の意味推定を実現している．後者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ことで英字略語の意味推定を実現している．これらの方法は適用可能な英字略語に制約条件（英字略語の意味が括弧書きで併記されていることが必要，片仮名表記が必要）がある．今回の評価で使用した 124 記事に対して，当該方法が適用できる英字略語はそれぞれ 35%程度であったため，これらの方法の精度は最高でも 35%程度となる．

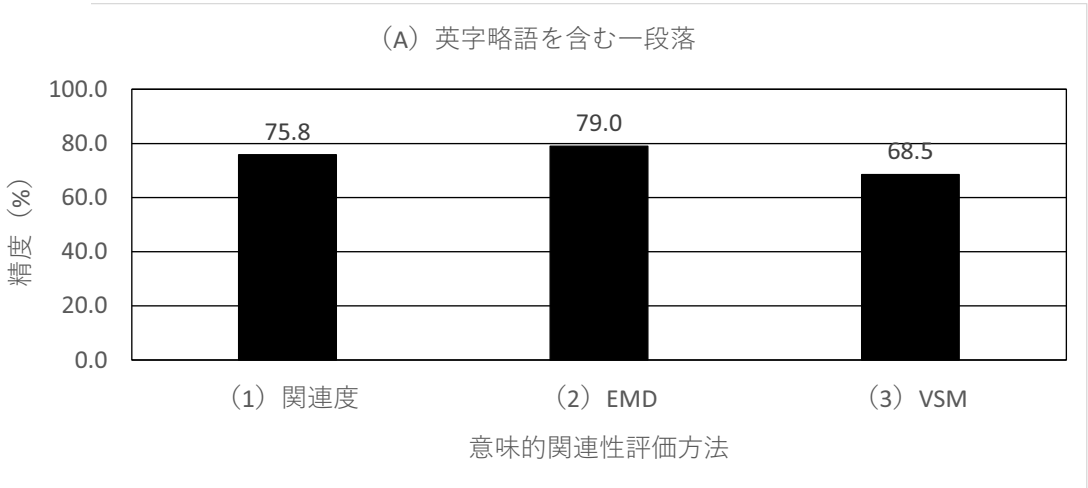


図 6.2: 評価結果その 1（オートフィードバックを用いた英字略語の意味推定法）

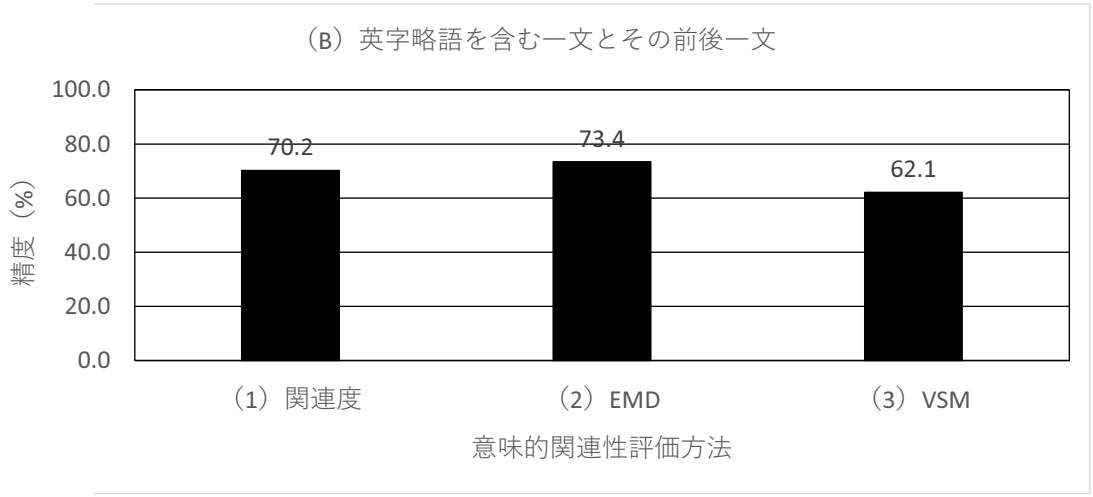


図 6.3: 評価結果その 2（オートフィードバックを用いた英字略語の意味推定法）

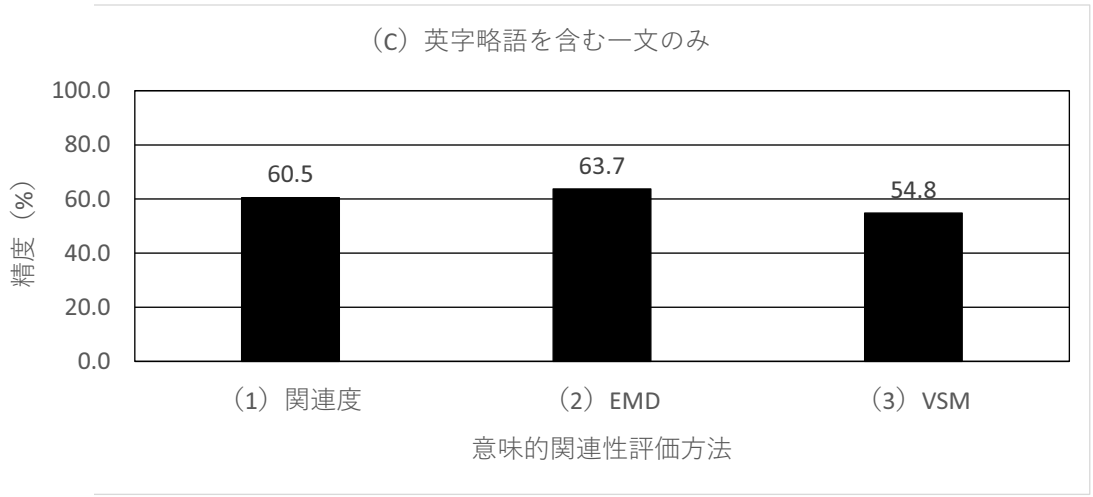


図 6.4: 評価結果その 3（オートフィードバックを用いた英字略語の意味推定法）

図 6.2，図 6.3，図 6.4 から，全体的な傾向として，関連度計算はベクトル空間モデルより高い精度（5.7%-8.1%）を示した．また，EMD はベクトル空間モデルより精度が高い（8.9%-11.3%）ことに加え，関連度計算よりも若干ではあるが高い精度を獲得した（3.2%）．他には，入力文章が長い（入力情報量が多い）ほど高い精度を得ていることが分かる．オートフィードバックを用いた英字略語の意味推定結果の一例を表 6.1，表 6.2，表 6.3 に示す．

まず，表 6.1 では，英字略語 AFC に対して，Wikipedia から取得した 13 個の意味候補から，9 つの全てのパターンにおいて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，表 6.2 では，英字略語 HD に対して，13 個の意味候補から，正しい意味を推定できたパターンと推定に失敗したパターンがある．具体的には，ベクトル空間モデルを使用し，かつ，入力文章に英字略語を含む一文とその前後一文または英字略語を含む一文のみを使用した場合（パターン (B)-(3)，パターン (C)-(3)），意味の推定に失敗している．また，関連度計算を使用し，かつ，入力文章に英字略語を含む一文のみを使用した場合（パターン (C)-(1)）も意味の推定に失敗している．これは，意味候補「ハーレーダビッドソン」を構成する属性に「メーカー」など入力文章に登場する語が含まれており，意味的に近いと判定されたためだと考えられる．

最後に，表 6.3 では，英字略語 FB に対して，15 個の意味候補から，9 つの全てのパターンにおいて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．

表 6.1: 英字略語の意味推定結果一例その 1（オートフィードバックを用いた英字略語の意味推定法）

英字略語	英字略語の意味	意味候補数	入力文章	入力文章種別	意味的関連性評価方法	推定結果	判定
AFC	アジアサッカー連盟（Asian Football Confederation）	13	国際サッカー連盟は17日、ハマム元理事を永久活動停止処分としたと発表した。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。ハマム元理事は15日付でサッカーに関わる全ての職を辞すると、文書で国際サッカー連盟に通達していた。ハマム元理事は昨年、会長選に絡む買収疑惑で国際サッカー連盟から永久資格停止処分を受けていた。今年7月、スポーツ仲裁裁判所は証拠不十分として元理事側の異議申し立てを認めたが、国際サッカー連盟はさらに倫理委員会が調査していた。	(A)一段落	(1)関連度	アジアサッカー連盟	○
					(2)EMD	アジアサッカー連盟	○
					(3)VSM	アジアサッカー連盟	○
			国際サッカー連盟は17日、ハマム元理事を永久活動停止処分としたと発表した。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。ハマム元理事は15日付でサッカーに関わる全ての職を辞すると、文書で国際サッカー連盟に通達していた。	(B)前後一文	(1)関連度	アジアサッカー連盟	○
					(2)EMD	アジアサッカー連盟	○
					(3)VSM	アジアサッカー連盟	○
			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。	(C)一文のみ	(1)関連度	アジアサッカー連盟	○
					(2)EMD	アジアサッカー連盟	○
					(3)VSM	アジアサッカー連盟	○

表 6.2: 英字略語の意味推定結果一例その 2（オートフィードバックを用いた英字略語の意味推定法）

英字略語	英字略語の意味	意味候補数	入力文章	入力文章種別	意味的関連性評価方法	推定結果	判定
HD	高精細度（High Definition）	13	米紙ウォール・ストリート・ジャーナルは12日、米国のアップルが日本のシャープ、台湾の鴻海精密工業などアジアの電子メーカーと**HD**大画面テレビの設計を共同で進めていると報じた。 アップルはスマートフォン（多機能携帯電話端末）に続き、テレビ市場でもサムスン電子と激突する可能性が高まった。 同紙は消息筋の話として「まだ正式なプロジェクトではなく、テストの初期段階だ」と伝えた。 シャープは現在、アップルに液晶パネルを供給。 鴻海は中国でiPhoneを受託生産する富士康国際（フォックスコン・インターナショナル）の親会社。 アップルが両社と組んでテレビを発売するとの見方は昨年からあった。 その可能性が最近になって高まった格好だ。 米国のベンチャー投資家、マーク・アンデルセン氏は「アップルがテレビを発売するのは確実だ。 時期は2014年が有力だが、早ければ来年の発売も可能だ」と指摘した。	（A）一段落	（1）関連度	高精細度	○
					（2）EMD	高精細度	○
					（3）VSM	高精細度	○
			米紙ウォール・ストリート・ジャーナルは12日、米国のアップルが日本のシャープ、台湾の鴻海精密工業などアジアの電子メーカーと**HD**大画面テレビの設計を共同で進めていると報じた。 アップルはスマートフォン（多機能携帯電話端末）に続き、テレビ市場でもサムスン電子と激突する可能性が高まった。	（B）前後一文	（1）関連度	高精細度	○
					（2）EMD	高精細度	○
					（3）VSM	ハーレー ダビッドソン	×
			米紙ウォール・ストリート・ジャーナルは12日、米国のアップルが日本のシャープ、台湾の鴻海精密工業などアジアの電子メーカーと**HD**大画面テレビの設計を共同で進めていると報じた。	（C）一文のみ	（1）関連度	ハーレー ダビッドソン	×
					（2）EMD	高精細度	○
					（3）VSM	ハーレー ダビッドソン	×

表 6.3: 英字略語の意味推定結果一例その 3（オートフィードバックを用いた英字略語の意味推定法）

英字略語	英字略語の意味	意味候補数	入力文章	入力文章種別	意味的関連性評価方法	推定結果	判定
FB	フェイスブック（FaceBook）	15	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。 州当局によると、今年5月の上場の前に、業績見通しに関し、モルガン・スタンレーの担当者に相談した。 そのうえで、証券アナリストらには説明した内容を米証券取引委員会への提出資料には記載しなかった。	(A)一段落	(1)関連度	政府短期証券	×
					(2)EMD	政府短期証券	×
					(3)VSM	政府短期証券	×
			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。 州当局によると、今年5月の上場の前に、業績見通しに関し、モルガン・スタンレーの担当者に相談した。	(B)前後一文	(1)関連度	政府短期証券	×
					(2)EMD	政府短期証券	×
					(3)VSM	政府短期証券	×
			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。	(C)一文のみ	(1)関連度	政府短期証券	×
					(2)EMD	政府短期証券	×
					(3)VSM	政府短期証券	×

6.4 まとめ

本章では，英字略語の意味推定法として，オートフィードバックを用いて多義性を有する英字略語に対して意味の推定を行う方法を提案した．新聞記事から抽出した英字略語に対して複数パターンの評価実験を行った結果，提案方法は英字略語の意味推定に成功する精度が最高で70%を超えており，提案方法が英字略語の多義性をある程度解消できていることを示した．さらに，提案方法は比較方法であるベクトル空間モデルよりも良好な結果を示したことから，その有効性を確認することができた．

本章で提案した方法により，多義性を有する未定義語（英字略語）の意味を理解しやすくできるようになる．提案方法はインターネット上に登場する多くの英字略語に適用することができるため，機械が言語を理解する能力を実現するための一助となることが期待できる．

第7章　結論

本研究では，機械が人間と自然な会話を行うために言語を理解する方法を実現することを目指し，未定義語の意味を推定・取得する方法を提案した．未定義語を基準となる言語資源である概念ベースに登録されていない単語と定義し，いくつかの言語資源を活用して未定義語の意味推定を実施し，その有効性を確認した．

第 2 章では，機械に人間が行っているような連想能力や意味理解能力を持たせることを目的とした連想メカニズムについて述べた．本論文における連想メカニズムは，単語に属性と重みの集合を与える語の概念化方法，概念化された語と語の間の関連を定量的に評価することで意味の近さを測ることを可能とする意味的関連性評価方法，および，単語を意味的に分類した分類体系であるシソーラスから構成される．

語の概念化方法として，電子化国語辞書などから構築した概念ベースを用いる方法，インターネット百科事典である Wikipedia から構築した文書データベースを用いる方法，および，インターネット情報である Web 検索エンジンの検索結果を利用するオートフィードバックを用いる方法を説明した．概念ベースを用いる方法は一般名詞を中心に約 9 万語を概念化することができ，その精度は 80%を超える高品質な概念化方法である．文書データベースを用いる方法は固有名詞を中心に約 100 万語を概念化することができ，その精度は 70%を超える概念化方法である．オートフィードバックを用いる方法はインターネット上に登場する語を概念化することができ，その精度は 70%程度とやや品質は低いものの適用範囲が広い概念化方法である．したがって，上述した 3 つの概念化方法を使い分けることで，語の特性に応じて概念化を実施することが可能となる．

意味的関連性評価方法として，関連度計算と Earth Mover's Distance を説明した．どちらの方法も，概念化された 2 つの語を入力することで語間の意味の近さ（関連）を定量的に表現した値を出力することができる．両者は，意味の近さを計算する際に，入力された語が持つ属性の対応の取り方が異なる．関連度計算は関連性が高い順に属性を対応付ける 1 対 1 で対応を取る計算方法であることに対して，Earth Mover's Distance は全ての属性を対応付ける M 対 N で対応とる計算方法である．

シソーラスとして，木構造を持つ名詞シソーラスであり，上位下位シソーラスの 1 つである NTT シソーラスを説明した．さらに，概念ベースを用いて NTT シソーラスのノードを概念化する方法について述べ，NTT シソーラスに対しても意味的関連性評価方法を適用することを可能にした．

本研究においては，上述した連想メカニズムを用いることで，連想能力や意味理解能力に基づいて未定義語の意味を推定・取得する方法を提案した．

第 3 章では，文書データベースを用いて固有名詞の意味を推定する方法について述べた．本章では，基準となる言語資源である概念ベースに加えて，文書データベースと Wikipedia を言語資源として活用している．具体的には，Wikipedia を情報源として構築した文書データベースを用いて未定義語（固有名詞）の概念化を実施し，概念化した未定義語と NTT シソーラスの各ノードに対して関連度計算を適用することで，未定義語が所属するべきノードを提示する方法を提案した．未定義語として代表的な存在である固有名詞を対象に評価実験を行った結果，第 1 位のノードのみを出力した場合の精度が約 41%，第 10 位のノードまで出力した場合の精度が 80%を超えることを示した．さらに，未定義語と各ノードの関連性の評価に検索ヒット数を組み込んで評価することで，第 1 位のノードのみを出力した場合の精度が約 60%となることを示し，提案方法の有効性を示した．

第 4 章では，文書データベースを用いて英字略語の意味を推定する方法について述べた．本章では，基準となる言語資源である概念ベースに加えて，文書データベースと Wikipedia を言語資源として活用し，よく利用されるものの多義性により意味の理解が困難な英字略語の意味を推定する方法について提案した．具体的には，英字略語を含む入力文章に基づいて概念化を実施した英字略語と，Wikipedia から取得した意味候補に対して意味的関連性評価方法を適用することで，英字略語の多義性を解消し，意味の推定を実現している．評価実験により，提案方法の精度は最高で 70%を超えていることに加え，比較方法であるベクトル空間モデルよりも良好な結果を示したことから，その有効性を示した．

第 5 章では，オートフィードバックを用いて固有名詞の意味を推定する方法について述べた．本章では，基準となる言語資源である概念ベースに加えて，インターネット情報と NTT シソーラスを言語資源として活用している．具体的には，インターネット上に登場する語を概念化することが可能なオートフィードバックを活用することで，より多くの未定義語（固有名詞）に対して所属するべき NTT シソーラスのノードを提示する方法を提案した．固有名詞を対象に評価実験を行った結果，第 1 位のノードのみを出力した場合の精度が 60%を超え，第 10 位のノードまで出力した場合の精度が 90%を超えることを示した．さらに，提案方法は比較方法であるベクトル空間法に基づく 2 種類の比較方法よりも優れた精度を有することを示し，その有効性を確認した．

第 6 章では，オートフィードバックを用いて英字略語の意味を推定する方法について述べた．本章では，基準となる言語資源である概念ベースに加えて，インターネット情報と Wikipedia を言語資源として活用している．具体的には，オートフィードバックを用いた概念化を活用することで，より広範な英字略語の意味を推定する方法について提案した．評価実験により，提案方法の精度は最高で 70%を超えており，さらに，比較方法であるベクトル空間モデルよりも良好な結果を示したことで，その有効性を確認した．

近年の情報処理技術の進展に伴い，今後，高度化・知的化されたコンピュータなど様々な機械が人間と共存するようになる状況が到来すると考えられる．本論文では，人間と共存する機械に与える仕組みの一環として，人間と自然な会話を行うために重要な能力である言語の理解を目指し，語彙の意味を推定する方法について述べた．人間が曖昧さや柔軟さを持って語の意

味を理解するために活用している「連想」という能力を実現するために構築された概念ベースを基準となる言語資源と考え，概念ベースに存在しない語を未定義語と定義し，いくつかの言語資源を活用して未定義語の意味を推定・取得する方法を提案した．本論文では，未定義語として，代表的な存在である固有名詞，および，多義性を持つために意味の推定が困難な英字略語を取り上げ，それぞれの意味を推定する方法を提案し，連想に基づく意味推定の有効性を示した．今後，本研究で培った言語を扱う技術である自然言語処理技術を発展させ，人間が持つ連想能力を機械上に実現し，人間と共存できる機械する研究・開発に取り組んでいきたいと考えている．

謝辞

本論文は，私が同志社大学大学院理工学研究科（旧工学研究科）知識情報処理研究室に在籍していた期間に行った研究をまとめたものである．本研究を遂行するにあたり，多くの方々に多大なるご指導，ご支援を賜りましたことを心から感謝いたします．

河岡司教授，渡部広一教授，土屋誠司教授には，研究の立ち上げ方やその進め方，研究成果のまとめ方や研究者としての考え方など研究者としての基礎的能力と共に，研究成果を如何に日常の問題解決へ応用するかなどの応用能力について懇切丁寧にご指導していただきました．また，社会人からの博士後期課程への入学も快く受け入れていただき，多大なご配慮をいただきましたこと心から感謝いたします．

下原勝憲教授には，本論文の審査をしていただきました．丁寧に論文を読んでいただいたことに加え，研究内容の意義や研究成果のアピールなどについて議論させていただき，貴重な助言や示唆をいただきましたことを感謝いたします．

本論文で提案した，連想メカニズムを活用した語彙の意味推定法の検討において，関係諸氏には貴重な助言をいただくと共に，実験データの収集・解析にご支援をいただきました．第3章の文書データベースを用いた固有名詞の意味推定法については惠村日南子氏，第4章の文書データベースを用いた英字略語の意味推定法については斎木淳氏，第5章のオートフィードバックを用いた固有名詞の意味推定法については伊藤俊介氏，辻泰希氏，第6章のオートフィードバックを用いた英字略語の意味推定法については田邊僚氏に深く感謝いたします．また，日頃から研究についての討論や発表練習などに協力いただいた知識情報処理研究室の諸氏にはお礼を申し上げます．

最後に，あらゆる面で援助していただき，大学で研究を行うという貴重な場を私に与えてくださった両親に感謝いたします．

参考文献

[1] 西田豊明：視点　人工知能スキーマ：人々は人工知能をどうとらえているか, 情報管理, Vol. 60, No. 1, pp. 50–55 (2017).

[2] 早稲田大学ヒューマノイドプロジェクト編著：人間型ロボットのはなし, 日本工業新聞社 (1999).

[3] 日経メカニカル, 日経デザイン共同編集：RoBolution(ロボリューション)-人型二足歩行タイプが開くロボット産業革命, 日経 BP 社 (2001).

[4] 海野裕也：人と機械の言語獲得, 認知科学, Vol. 24, No. 1, pp. 16–22 (2017).

[5] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦：日本語語彙体系, 岩波書店 (1997).

[6] Miller, G. A.: WordNet: A Lexical Database for English, *Communications of the ACM*, Vol. 38, No. 11, pp. 39–41 (1995).

[7] Collins, A. M. and Quillian, M. R.: Retrieval time from semantic memory, *Journal of Verbal Learning and Verbal Behavior*, Vol. 8, No. 2, pp. 240–247 (1969).

[8] Salton, G., Wong, A. and Yang, C. S.: A Vector Space Model for Automatic Indexing, *Communications of the ACM*, Vol. 18, No. 11, pp. 613–620 (1975).

[9] Rychener, M. D.: Control requirements for the design of production system architectures, in *Proceedings of the 1977 symposium on Artificial intelligence and programming languages*, pp. 37–44 (1977).

[10] 浦本直彦：コーパスに基づくシソーラス-統計情報を用いた既存のシソーラスへの未知語の配置, 情報処理学会論文誌, Vol. 37, No. 12, pp. 2182–2189 (1996).

[11] 田中穂積, 仁科喜久子：上位／下位関係シソーラス ISAMAP1 の作成［I］, 情報処理学会研究報告自然言語処理, Vol. 84, No. 1987-NL-064, pp. 25–34 (1987).

[12] 田中穂積, 仁科喜久子：上位／下位関係シソーラス ISAMAP の作成［II］, 情報処理学会研究報告自然言語処理, Vol. 84, No. 1987-NL-064, pp. 35–44 (1987).

[13] 前田康成：統計的決定理論に基づく既存名詞シソーラスへの未知語登録方法に関する一考察, 電子情報通信学会論文誌, Vol. J83-A, No. 6, pp. 702–710 (2000).

[14] 榊剛史, 松尾豊, 内山幸樹, 石塚満：Web 上の情報を用いた関連語のシソーラス構築について, 自然言語処理, Vol. 14, No. 2, pp. 3–31 (2007).

[15] 別所克人, 内山俊郎, 内山匠, 片岡良治, 奥雅博：単語・意味属性間共起に基づくコーパス概念ベースの生成方式, 情報処理学会論文誌, Vol. 49, No. 12, pp. 3997–4006 (2008).

[16] Mihalcea, R.: Using Wikipedia for AutomaticWord Sense Disambiguation, in *Proceedings of NAACL HLT*, pp. 196–203 (2007).

[17] Sun, Y., Lin, L., Tang, D., Yang, N., Ji, Z. and Wang, X.: Modeling Mention, Context and Entity with Neural Networks for Entity Disambiguation, in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pp. 1333–1339 (2015).

[18] 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).

[19] 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, 電子情報通信学会技術研究報告, Vol. 101, No. 712, pp. 109–116 (2002).

[20] 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).

[21] Wikimedia Foundation, : ウィキペディアフリー百科事典, https://jp.wikipedia.org.

[22] Hofmann, T.: Probabilistic Latent Semantic Indexing, in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57 (1999).

[23] 入江毅, 渡部広一, 河岡司, 松澤和光：知的判断メカニズムのための概念間の類似度評価モデル, 情報処理学会研究報告知能と複雑系（ICS）, No. 1(1998-ICS-115), pp. 93–98 (1999).

[24] 笠原要, 松澤和光, 石川勉：国語辞書を利用した日常語の類似性判別, 情報処理学会論文誌, Vol. 38, No. 7, pp. 1272–1283 (1997).

[25] 大野晋, 浜西正人：類語国語辞典, 角川書店 (1985).

[26] 武部良明（編）：必携類語実用辞典, 三省堂 (1977).

[27] 三省堂編集所（編）：必携用字用語辞典　第四版, 三省堂 (1992).

[28] 見坊豪紀：三省堂国語辞典　第四版, 三省堂 (1992).

[29] 松村明, 三省堂編修所（編）：大辞林, 三省堂 (1988).

[30] 新村出（編）：広辞苑　第四版, 岩波書店 (1991).

[31] 長尾真：岩波講座　ソフトウェア科学　15　自然言語処理, 岩波書店 (1996).

[32] 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).

[33] 眞鍋康人, 小島一秀, 渡部広一, 河岡司：概念間の関連度やシソーラスを用いた概念ベースの自動精錬手法, 同志社大学理工学研究報告, Vol. 42, No. 1, pp. 9–20 (2001).

[34] 橋本隆志, 渡部広一, 河岡司：新聞記事等の文書を用いた概念自動学習による概念ベース構築方式, 情報処理学会研究報告自然言語処理, Vol. 2002, No. 20(2001-NL-148), pp. 89–96 (2002).

[35] 工藤拓：MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://taku910.github.io/mecab/.

[36] 中川裕志, 前田朗：専門用語（キーワード）自動抽出用 Perl モジュール”TermExtract” の解説, http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html.

[37] 中川裕志, 湯本紘彰, 森辰則：出現頻度と連接頻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).

[38] はてな：はてなキーワード, http://d.hatena.ne.jp/keyword/.

[39] 奥野陽：Social IME, 2016 年 9 月サービス終了.

[40] Google LLC, : Google, http://www.google.co.jp/.

[41] Salton, G. and Buckley, C.: Term-weighting approaches in automatic text retrieval, *Information Processing & Management*, Vol. 24, No. 5, pp. 513–523 (1988).

[42] 徳永健伸：情報検索と言語処理, 東京大学出版会 (1999).

[43] 佐藤敏紀：mecab-ipadic-NEologd, https://github.com/neologd/mecab-ipadic-neologd/.

[44] Google LLC, : Google Trends, https://trends.google.co.jp/trends/?geo=JP.

[45] 自由国民社： 「現代用語の基礎知識」選 ユーキャン新語・流行語大賞, https://www.jiyu.co.jp/singo/.

[46] 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, Vol. JSAI04, pp. 1–4 (2004).

[47] 奈良先端科学技術大学院大学：ChaSen – 形態素解析器, http://chasen-legacy.osdn.jp/.

[48] 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).

[49] 荒木孝允, 奥村紀之, 渡部広一, 河岡司：比較対象概念の共通属性を重視する動的関連度計算方式, 同志社大学理工学研究報告, Vol. 48, No. 3, pp. 140–150 (2007).

[50] Rubner, Y., Tomasi, C. and Guibas, L. J.: The Earth Mover's Distance as a Metric for Image Retrieval, *International Journal of Computer Vision*, Vol. 40, No. 2, pp. 99–121 (2000).

[51] Hoffman, A. J.: On simple linear programming problems, in *Proceedings of Symposia in Pure Mathematics*, pp. 317–327 (1963).

[52] 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).

[53] 伊藤俊介, 渡部広一, 河岡司：情報検索における未知語理解支援方式～未知語のシソーラスノードへの分類～, 情報処理学会研究報告自然言語処理, Vol. 2004-NL-159, No. 1, pp. 61–66 (2004).

[54] 永田昌明：統計的言語モデルと N-best 探索を用いた日本語形態素解析法, 情報処理学会論文誌, Vol. 40, No. 9, pp. 3420–3431 (1999).

[55] 趙國, 宮山章子, 山下洋一：N-best 音声認識における認識スコアを利用した候補提示数の決定, 電子情報通信学会論文誌, Vol. J88-D-II, No. 6, pp. 1003–1011 (2005).

[56] 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付け手法, 自然言語処理, Vol. 15, No. 3, pp. 91–113 (2008).

[57] Harris, Z. S.: *Mathematical Structures of Language*, Interscience Publishers (1968).

[58] 日本電子化辞書研究所：EDR 電子化辞書利用マニュアル第 2.1 版 (1994).

[59] Witten, I. H., Moffat, A. and Bell, T. C.: *Managing Gigabytes: Compressing and Indexing Documents and Images, Second Edition*, Morgan Kaufmann (1999).

[60] 後藤和人, 土屋誠司, 渡部広一：語彙の概念化と Wikipedia を用いた英字略語の意味推定方法, 自然言語処理, Vol. 24, No. 3, pp. 351–369 (2017).

[61] 岡崎直観, 石塚満：日本語新聞記事からの略語抽出, 人工知能学会全国大会論文集, Vol. JSAI07, pp. 1–3 (2007).

[62] 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).

研究業績一覧

以下に記載した「○」は筆頭の研究業績であることを示す.

項目	題名	年月日	発表した方法	著者
修士論文				
1 ○	Web を用いた未定義語のシソーラスノードへの自動割付け	2008.3	同志社大学修士論文	後藤　和人
論文				
2 ○	Understanding Support Method of Unknown Words Using Robot Type Search Engine	2007.9	Lecture Notes in Artificial Intelligence 4692, Springer-Verlag, pp.631-638	Kazuto Goto, Noriyuki Okumura, Hirokazu Watabe, Tsukasa Kawaoka
3 ○	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付け手法	2008.7	自然言語処理, vol.15, no.3, pp.91-113,	後藤　和人，土屋　誠司，渡部　広一，河岡　司
4 ○	Meaning Estimation Method of Alphabetical Abbreviation Using Lexical Conceptualization and Wikipedia	2017.6	International Journal of Future Computer and Communication, vol.6, no.2, pp.53-57	Kazuto Goto, Seiji Tsuchiya, Hirokazu Watabe

項目	題名	年月日	発表した方法	著者
5 ○	語彙の概念化と Wikipedia を用いた英字略語の意味推定方法	2017.6	自然言語処理, vol.24, no.3, pp.351-369	後藤　和人，土屋　誠司，渡部　広一
6	Experimental Verification of 1-tap Time Domain Beamforming for P-MP relay system via 75 GHz band Measured CSI	2019.8	IEICE Transactions on Communications, E102-B, no.8, pp.1751-1762	Mizuki SUGA, Atsushi OHTA, Kazuto GOTO, Takahiro TSUCHIYA, Nobuaki OTSUKI, Yushi SHIRATO, Naoki KITA, Takeshi ONIZAWA
研究発表 (国際会議)				
7 ○	Understanding Support Method of Unknown Words Using Robot Type Search Engine	2007.9	Knowledge-Based Intelligent Information and Engineering Systems: KES 2007 - WIRN 2007, 11th International Conference, Salerno, Italy	Kazuto Goto, Noriyuki Okumura, Hirokazu Watabe, Tsukasa Kawaoka
8	Experimental Evaluation of 1-Tap Time Domain Beamforming based on 75 GHz Indoor CSI	2017.7	Proc. 2017 IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting (AP-S/URSI), California, USA, pp.1425-1426	Mizuki Suga, Kazuto Goto, Takahiro Tsuchiya, Hideyuki Tsuboi, Chunhsiang Huang, Kazuki Maruta, Atsushi Ohta

項目	題名	年月日	発表した方法	著者
9	Study on Orthogonalization Spacing of Antenna Arrays for Spatial Multiplexing in a millimeter-wave massive MIMO System	2018.4	2018 Wireless Telecommunications Symposium, Arizona, USA	Chun-Hsiang Huang, Atsushi Ohta, Kazuto Goto, Yushi Shirato, Yutaka Imaizumi, Naoki Kita
10	Experimental Study of Irradiation Range of Directional Antenna for NLOS Area - Toward High Accuracy Evaluation of Interference Area-	2018.7	2018 Asian Workshop on Antennas and Propagation, Pattaya, Thailand	Naoki Kita, Yushi Shirato, Kazuto Goto, Hideyuki Tsuboi, Yutaka Imaizumi, Hiroyuki Nakamura
研究発表(研究会)				
11 ○	ロボット型検索エンジンを用いた未知語の理解支援手法	2007.7	電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, vol.107, no.158, pp.85-90	後藤　和人，渡部　広一，河岡　司
12 ○	語彙の概念化とWikipediaを用いた英字略語の意味推定手法	2016.3	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, SIG-AM-12-01, pp.1-8	後藤　和人，土屋　誠司，渡部　広一
13 ○	概念ベースとWikipediaを用いた英字略語の意味推定	2017.3	電子情報通信学会技術研究報告, vol. 116, no. 489, AI2016-45, pp. 25-30	後藤　和人，土屋　誠司，渡部　広一

項目	題名	年月日	発表した方法	著者
14	75GHz 帯伝搬測定によるストリートスモールセル基地局向け無線エントランス環境の実験的評価	2017.6	電子情報通信学会技術研究報告, vol. 117, no. 103, RCS2017-74, pp. 143-148	土屋　貴寛，後藤　和人，菅　瑞紀，黄　俊翔，坪井　秀幸，黒崎　聰，太田　厚，飯塚　正孝
15 ○	ストリートスモールセル基地局への無線エントランスにおけるビームフォーミング技術 ～ 75GHz 帯屋外/屋内伝搬測定による特性評価 ～	2017.6	電子情報通信学会技術研究報告, vol. 117, no. 103, RCS2017-75, pp. 149-154	後藤　和人，土屋　貴寛，菅　瑞紀，太田　厚，飯塚　正孝
16	高周波数帯を用いた無線エントランスシステムにおける低相関指向性形成のためのアンテナ構成法	2017.7	電子情報通信学会技術研究報告，vol. 117, no. 132, RCS2017-136, pp. 221-226	太田　厚，田中　健，白戸　裕史，菅　瑞紀，後藤　和人，飯塚　正孝
17	5G スモールセル基地局向け無線エントランスの超低遅延 ARQ の提案	2017.11	電子情報通信学会技術研究報告，vol. 117, no. 284, RCS2017-218, pp. 75-80	太田　厚，黒崎　聰，後藤　和人，今泉　豊，北　直樹
18	Massive MIMO を用いたマルチビーム・スタジアム Wi-Fi の一検討	2018.1	電子情報通信学会技術研究報告，vol. 117, no. 396, RCS2017-293, pp. 143-148	太田　厚，田中　健，後藤　和人，北　直樹

項目	題名	年月日	発表した方法	著者
研究発表(全国大会)				
19 ○	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付け手法	2006.3	情報処理学会第 68 回全国大会, 4N-3, pp.447-448	後藤　和人，渡部　広一，河岡　司
20	住宅地における到来角度特性の高さ依存性	2009.3	2009 年電子情報通信学会総合大会, 通信講演論文集 1, B-1-10, p.10	伊藤　俊夫，北　直樹，山田　渉，後藤　和人，横山 信治
21 ○	広域ユビキタスワイヤレスネットワークにおける低消費電力呼出方式の検討	2009.3	2009 年電子情報通信学会総合大会, 通信講演論文集 1, B-5-136, p.569	後藤　和人，布　房夫，清水　芳孝，渡辺　和二
22	広域ユビキタスネットワークにおける優先制御方式の検討 (1) ― マルコフ近似モデルを用いたランダムアクセスウィンドウ制御方式の提案 ―	2010.3	2010 年電子情報通信学会総合大会, 通信講演論文集 2, B-8-16, p.279	布　房夫，後藤　和人，清水　芳孝，加々見　修，吉野　修一
23 ○	広域ユビキタスネットワークにおける中継伝送方式の検討	2010.3	2010 年電子情報通信学会総合大会, 通信講演論文集 2, B-8-18, p.281	後藤　和人，布　房夫，清水　芳孝，加々見　修，吉野　修一
24 ○	広域ユビキタスネットワークにおける中継伝送方式の検討 (その 2)	2010.9	2010 年電子情報通信学会ソサイエティ大会, 通信講演論文集 2, B-8-29, p.177	後藤　和人，布　房夫，清水　芳孝，吉野　修一

項目	題名	年月日	発表した方法	著者
25 ○	センサネットワークにおける中継伝送アクセス方式の検討	2011.3	2011 年電子情報通信学会総合大会, 通信講演論文集 2, B-8-14, p.263	後藤　和人，清水　芳孝，吉野　修一
26 ○	伝送効率改善のための無線アクセス方式切替方法の検討	2011.9	2011 年電子情報通信学会ソサイエティ大会, 通信講演論文集 2, B-8-43, p.201	後藤　和人，清水　芳孝，吉野　修一
27	ユーザセントリックワイヤレスホームネットワークにおける無線信号処理プロトコルの検討（1）	2012.3	2012 年電子情報通信学会総合大会, 通信講演論文集 1, B-17-23, p.621	清水　芳孝，後藤　和人，白戸　裕史，藤田　隆史，藤野　洋輔，吉野　修一
28 ○	ユーザセントリックワイヤレスホームネットワークにおける無線信号処理プロトコルの検討（2）	2012.3	2012 年電子情報通信学会総合大会, 通信講演論文集 1, B-17-24, p.622	後藤　和人，清水　芳孝，白戸　裕史，藤田　隆史，藤野　洋輔，吉野　修一
29	広域センサーネットワークを用いた無線 LAN アクセスポイント遠隔制御技術の検討	2013.3	2013 年電子情報通信学会総合大会, 通信講演論文集 1, B-5-150, p.560	清水　芳孝，後藤　和人，水野　晃平，熊谷　智明，吉野　修一
30 ○	近傍配置された無線基地局における近接チャネル利用方法の一検討	2013.3	2013 年電子情報通信学会総合大会, 通信講演論文集 1, B-5-199, p.609	後藤　和人，白戸　裕史，鈴木　康夫，片山　穣，小林　守，吉野　修一
31	可搬型 ICT 基盤における無線アクセスネットワーク技術	2013.3	2013 年電子情報通信学会総合大会, TK-3-5, pp.SSS26-SSS27	熊谷　智明，清水　芳孝，後藤　和人，水野　晃平，吉野　修一

項目	題名	年月日	発表した方法	著者
32 ○	マルチホップ無線アクセスネットワーク構成方法に関する一検討	2013.9	2013 年電子情報通信学会ソサイエティ大会, 通信講演論文集 1, B-5-98, p.461	後藤　和人，清水　芳孝，熊谷　智明，吉野　修一
33	個人情報を重視した時事情報提供手法	2015.9	FIT2015（第 14 回情報科学技術フォーラム）, 第 2 分冊, D-007, pp.75-76	津田　健太郎，後藤　和人，土屋　誠司，渡部　広一
34	Wikipedia 記事の文章群を用いた多義を有する英字略語の意味判断システム	2016.9	FIT2016（第 15 回情報科学技術フォーラム）, 第 2 分冊, F-034, pp.249-250	後藤　大介，後藤　和人，土屋　誠司，渡部　広一
35	大規模アンテナ無線エントランスシステムの 75GHz 帯屋内伝搬測定	2017.3	2017 年電子情報通信学会総合大会, 通信講演論文集 1, B-5-109, p.445	土屋　貴寛，後藤　和人，菅　瑞紀，黄　俊翔，坪井　秀幸，丸田　一輝，太田　厚
36	75GHz 帯屋内実測 CSI を用いた 1 タップ時間領域ビームフォーミングの特性評価	2017.3	2017 年電子情報通信学会総合大会, 通信講演論文集 1, B-5-110, p.446	菅　瑞紀，後藤　和人，土屋　貴寛，坪井　秀幸，黄　俊翔，丸田　一輝，太田　厚
37	75GHz 帯屋内実測 CSI 情報を用いたデジタルアシスト型アナログビームフォーミングの特性評価	2017.3	2017 年電子情報通信学会総合大会, 通信講演論文集 1, B-5-111, p.447	黄　俊翔，丸田　一輝，菅　瑞紀，後藤　和人，土屋　貴寛，坪井　秀幸，太田　厚
38 ○	75GHz 帯屋内実測 CSI を用いた大規模アンテナの空間多重特性評価	2017.3	2017 年電子情報通信学会総合大会, 通信講演論文集 1, B-5-112, p.448	後藤　和人，黄　俊翔，丸田　一輝，菅　瑞紀，土屋　貴寛，坪井　秀幸，太田　厚

項目	題名	年月日	発表した方法	著者
39	超密集地における Massive MIMO を用いたマルチビームアクセス技術	2017.9	2017 年電子情報通信学会ソサイエティ大会, 通信講演論文集 1, B-5-73, p.322	太田　厚, 菅　瑞紀, 田中　健, 後藤　和人, 北　直樹
40 ○	5G スモールセル基地局への無線エントランスにおけるデジタルアシスト型アナログビームフォーミング技術の拡張	2017.9	2017 年電子情報通信学会ソサイエティ大会, 通信講演論文集 1, B-5-75, p.324	後藤　和人, 田中　健, 太田　厚, 飯塚　正孝, 北　直樹
41	DAABF を実装した 40GHz 帯 Massive MIMO 実ビット伝送実験装置による伝送特性	2018.3	2018 年電子情報通信学会総合大会, 通信講演論文集 1, B-5-115, p.406	田中 健, 太田 厚, 後藤 和人, 白戸 裕史, 竹厚 善生, 赤堀 耕一郎, 谷口 徹, 北 直樹
42	RoF を適用したミリ波 FWA システムの提案	2019.3	2019 年電子情報通信学会総合大会, 通信講演論文集 1, B-5-112, p.378	白戸 裕史, 伊藤 耕大, 菅 瑞紀, 後藤 和人, 俊長 秀紀, 北 直樹
43 ○	ミリ波 RoF-FWA システムにおけるスループット特性改善方法	2019.3	2019 年電子情報通信学会総合大会, 通信講演論文集 1, B-5-115, p.381	後藤 和人, 俊長 秀紀, 白戸 裕史, 北 直樹
44	5G モバイルフロントホールへの無線適用に関する一検討　～　アンテナ小型化に伴う同一チャネル間干渉の影響　～	2019.3	2019 年電子情報通信学会総合大会, 通信講演論文集 1, B-5-119, p.385	俊長 秀紀, 白戸 裕史, 黄 俊翔, 伊藤 耕大, 後藤 和人, 北 直樹
45 ○	ミリ波 RoF-FWA システムにおける送信順序制御方法のスループット特性評価	2019.9	2019 年電子情報通信学会ソサイエティ大会, 通信講演論文集 1, B-5-67, p.321	後藤 和人, 俊長 秀紀, 白戸 裕史, 北 直樹

項目	題名	年月日	発表した方法	著者
その他(特許)				
46 ○	無線端末呼出方法及び無線アクセスシステム	2009.2	特開 2010-199901	後藤　和人，布　房夫，清水　芳孝
47	無線通信方法、無線基地局、無線アクセスシステム	2010.2	特開 2011-176498	布　房夫，清水　芳孝，後藤　和人
48 ○	無線通信方法、無線基地局、無線アクセスシステム	2010.2	特開 2011-176507	後藤　和人，布　房夫，清水　芳孝
49 ○	無線基地局装置	2011.2	特開 2012-165301	後藤　和人，清水　芳孝
50 ○	無線通信システム、及び無線方式設定方法	2011.11	特開 2013-102380	後藤　和人，望月 伸晃，清水 達也
51	基地局装置、及び無線通信方法	2011.12	特開 2013-120977	藤田　隆史，清水　芳孝，後藤　和人，藤野　洋輔，白戸　裕史，吉野　修一
52	再送パケット検出回路及び再送パケット検出方法	2011.12	特開 2013-121123	藤野　洋輔，白戸　裕史，清水　芳孝，藤田　隆史，後藤　和人，吉野　修一
53	無線通信システム、及び無線通信方法	2011.12	特開 2013-123089	清水　芳孝，白戸　裕史，藤田　隆史，藤野　洋輔，後藤　和人，清水　達也，吉野　修一

項目	題名	年月日	発表した方法	著者
54	ヘッダ情報識別装置及びヘッダ情報識別方法	2011.12	特開 2013-126005	藤野　洋輔，白戸 裕史，清水　芳孝，藤田　隆史，後藤 和人，吉野　修一
55 ○	通信方法、無線アクセスシステム、及びプログラム	2011.12	特開 2013-126009	後藤　和人，清水 芳孝，藤野　洋輔，藤田　隆史，白戸 裕史，吉野　修一
56 ○	無線通信システム及び無線通信方法	2011.12	特開 2013-126010	後藤　和人，望月 伸晃，清水　達也，吉野　修一
57 ○	制御方法、無線システム及び無線基地局	2012.5	特開 2013-243524	後藤　和人，清水 芳孝，清水　達也，吉野　修一
58 ○	無線通信装置及び無線通信方法	2012.11	特開 2014-093688	後藤　和人，白戸 裕史，鈴木　康夫，片山　穣，小林　守，内田　大誠，吉野 修一
59	無線通信システム、遠隔基地局及び無線通信方法	2013.9	特開 2015-070409	清水　達也，後藤 和人，小林　守，吉野　修一
60	無線ネットワーク構築装置、無線ネットワークシステム、および、無線ネットワーク構築方法	2014.3	特開 2015-186074	清水　芳孝，熊谷 智明，後藤　和人

項目	題名	年月日	発表した方法	著者
61	アクセスポイント制御装置、アクセスポイント制御方法及びアクセスポイント制御プログラム	2014.9	特開 2016-054444	鈴木　康夫，清水　芳孝，後藤　和人，熊谷　智明
62 ○	無線基地局、サーバ装置、送信レート選択方法及び送信レート選択プログラム	2014.9	特開 2016-058904	後藤　和人，鈴木　康夫，清水　芳孝，熊谷　智明
63	無線通信システム、無線ＬＡＮアクセスポイント、サーバ装置、無線通信方法及び無線通信プログラム	2014.9	特開 2016-058915	清水　芳孝，熊谷　智明，鈴木　康夫，後藤　和人
64	無線通信装置及び無線通信方法	2017.2	特開 2018-142941	太田　厚，丸田　一輝，白戸　裕史，田中　健，黒崎　聡，後藤　和人
65	無線通信装置及び無線通信方法	2017.2	特開 2018-142942	太田　厚，丸田　一輝，白戸　裕史，田中　健，黒崎　聡，後藤　和人
66	無線通信装置及び無線通信方法	2017.2	特開 2018-142943	太田　厚，丸田　一輝，白戸　裕史，田中　健，黒崎　聡，後藤　和人
67 ○	無線通信装置及び再送制御方法	2017.4	特開 2018-182660	後藤　和人，太田　厚，黒崎　聡，飯塚　正孝

項目	題名	年月日	発表した方法	著者
68	無線通信装置及び無線通信システム	2017.6	特開 2019-9744	太田　厚，田中　健，白戸　裕史，菅　瑞紀，後藤　和人
69	無線通信装置及び回転量算出方法	2017.7	特開 2019-29706	太田　厚，田中　健，白戸　裕史，菅　瑞紀，後藤　和人
70	無線通信装置及び無線通信方法	2017.8	特開 2019-36853	太田　厚，後藤　和人，田中　健，菅　瑞紀
71	基地局装置及び無線通信システム	2017.10	特開 2019-75734	太田　厚，後藤　和人，北　直樹
72	無線通信装置及び無線通信方法	2018.3	特開 2019-176435	太田　厚，北　直樹，白戸　裕史，後藤　和人，伊藤　耕大，黄　俊翔
73	再送制御方法	2018.3	特開 2019-176440	太田　厚，北　直樹，今泉　豊，黒崎　聰，後藤　和人，伊藤　耕大
74	無線通信方法、無線通信システム及び無線局装置	2018.3	特開 2019-180065	太田　厚，北　直樹，白戸　裕史，今泉　豊，後藤　和人，黄　俊翔
75	無線通信装置及び無線通信方法	2018.7	特開 2020-10124	太田　厚，伊藤　耕大，後藤　和人，白戸　裕史，北　直樹

項目	題名	年月日	発表した方法	著者
76	干渉電力推定方法、干渉電力推定装置及びプログラム	2018.9	特願 2018-183559（出願中）	坪井　秀幸，今泉　豊，後藤　和人，伊藤　耕大，北　直樹，中村　宏之
77 ○	置局設計方法、置局設計装置、及び置局設計プログラム	2019.1	特願 2018-242831（出願中）	後藤　和人，俊長　秀紀，坪井　秀幸，白戸　裕史，北　直樹
78	設置候補提示方法、設置候補提示装置及びプログラム	2019.1	特願 2019-001401（出願中）	坪井　秀幸，後藤　和人，俊長　秀紀，白戸　裕史，北　直樹
79	見通し検出方法、見通し検出装置、及び見通し検出プログラム	2019.1	特願 2019-004727（出願中）	坪井　秀幸，俊長　秀紀，後藤　和人，白戸　裕史，北　直樹，鬼沢　武
80	置局設計方法、置局設計装置、及びプログラム	2019.1	特願 2019-007230（出願中）	俊長　秀紀，北　直樹，白戸　裕史，坪井　秀幸，後藤　和人
81 ○	割当方法及び信号処理装置	2019. 2	特願 2019-027783（出願中）	後藤　和人，俊長　秀紀，白戸　裕史，菅　瑞紀，伊藤　耕大，北　直樹
82	干渉評価方法、干渉評価装置、及び干渉評価プログラム	2019. 3	特願 2019-048821（出願中）	坪井　秀幸，俊長　秀紀，後藤　和人，白戸　裕史，北　直樹