社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－ＦＩ－86 (4)
2007－ＤＤ－60 (4)
2007／3／27

技術文書からの用語知識の自動獲得方式の検討

今村 誠　高山 泰博　三上 崇志　岡田 康裕
三菱電機（株）情報技術総合研究所 音声・言語処理技術部

用語の意味分類は、情報検索、テキストマイニング、業務依存の文書チェック/ガイダンスなどの自然言語処理の応用システムにおける重要な知識源である。本稿では、新聞記事等を中心に検証されてきた単語共起分布に基づく意味分類推定方式を、約1.5 万文規模の FA(Factory Automation)の設計仕様書に適用した実験結果を報告する。実験の結果、部品名や装置名のような体言的意味分類では、共起関係として「係り受け先」と「文節内の後方」が有効であり、作用や現象のような用言的な意味分類では、「係り受け元」が有効であることを確認した。属性は、両者の間の性質をもつが、特性,材質,状態など種々の用語を含むため、抽出が難しかった。また、「係り受け共起における表層格の区別の有無の効果」、「文節内共起の前方・後方の区別の有無の効果」、および、「複数種別の共起関係を併用した方式」についても考察した。

Word Knowledge Acquisition from Engineering Document

Makoto IMAMURA　Yasuhiro TAKAYAMA　Takashi MIKAMI　Yasuhiro OKADA
Mitsubishi Electric Corporation Information Technology R & D Center Human Media Technology Dept.

The words with classified senses are the important knowledge resource for application systems using natural language processing, such as information retrieval, textmining, and the work-dependent document check and guidance systems. The word sense disambiguation methods based on the cooccurrence distribution are traditionally examined for sentences in the newspaper articles. This paper describes an experimental result of the word sense disambiguation methods applied to about fifteen thousand sentences contained in the design specification documents in the factory automation domain.

In the result of the experiment, the governing words and the forward words in the bunsetsu segment support the sense estimation for the nominal words such as part names and abstract objects. The governed words support estimation for the verbal words such as actions and phenomena. The estimation for the attributive words is difficult due to various sub sense variation in the class such as characteristics, material, status. We also report the effect of the surface cases in the cooccurrence dependency, and the issues of quantity and quality of the training data.

1 はじめに

用語の意味分類や、用語間の上位下位関係や同義関係を記述した辞書(シソーラス)は、自然言語処理の応用システムにおける重要な知識源である。例えば、情報検索では、分析精度を向上させるために同義語辞書が用いられ、テキストマイニングでは分析観点や比較対象物の候補提示するために、対象分野毎のシソーラスが必要になる。設計書をチェックリストや不具合事例と照合して、関連不具合情報や注意事項を自動提示する設計品質向上支援システムでは、チェックルールを作成する際に、対象製品の仕様書や不具合報告書中に記載される部品名や不具合現象などの専門用語辞書が必要になる。また、情報機器の操作説明のガイダンスシステムでは、操作名、対象物、および機能などの用語辞書が必要になる。

日本語の既存シソーラスとしては、分類語彙表[1]やEDR 概念辞書[2]があるが、一般用語を主な対象としているため、上記の応用システムに適用するには、「専門用語の語彙が不足している」や「分類体系に応用面からの観点が欠けている」という場合があり、シソーラスを応用毎にカスタマイズする必要が生じる。

シソーラスを自動構築する研究は、獲得対象のデータの種別により分類すると、「人間用の辞書を用いる方法」と「テキストコーパスを利用する方法」とに大別される([3])。前者は、人間用に編集された辞書の機械可読版の語釈文を解析して、語と語の関係を抽出する方法である。後者は、「同じ文脈に出現する語は意味的にも似ている」という分布仮説に基づくものである(以下、「共起分布による方法」と呼ぶ)。

本報告では、FA(Factory Automation)分野の設計書をコーパスとして、部品名や属性などの意味分類の獲得方式について検討する。

本稿の構成は、以下の通りである。2章では、意味分類の応用ニーズと自動獲得の既存技術、および、本稿での考察範囲について述べる。3章では、本稿での実験の選択肢やパラメータを明確にするために、「共起分

布による方法」を、単語と素性が作る空間が各々共役関係にあるという観点から定式化する。4章では、実験の対象データ、推定対象の意味分類、および実験方法について述べる。5章では、実験結果と考察結果を述べる。6章では、まとめと今後の課題を述べる。

2 課題

本章では、意味分類の自動獲得に対する応用ニーズと、分布仮説に基づく意味分類自動獲得の技術課題、および、本稿で実施する実験の目的について述べる。

2.1 意味分類の自動獲得に対する応用ニーズ

本節では、情報検索、テキストマイニング、設計品質向上における意味分類の応用ニーズを述べる。

2.1.1 情報検索

ユーザが入力した検索式を拡張する際に、同義語や上位下位関係が用いられる。いずれも、検索もれの削減(再現率の向上)に寄与する。

2.1.2 テキストマイニング

分析観点や比較対象物の候補提示するために、対象分野毎のシソーラスが有用である。

例えば、不具合分析では、故障部位、故障部位の修飾表現、故障現象、故障対処操作などに用語を分類して、分類毎の用語の頻度情報を提示することにより、分析の観点を見つけやすくすることができる。

また、FAQ(Frequently asked questions)の作成支援では、FAQ を分類する上で参考になりそうな用語を作成したいというニーズがある。

2.1.3 設計品質向上支援システム

チェックルールを作成する際に、対象製品の仕様書や不具合報告書中に記載される部品名や不具合現象などの専門用語辞書が必要になる。

不具合未然防止システム[4]で利用される意味分類の階層例を図 1に示す。 不具合未然防止システムのチェックルールで用いられる用語は、「対象物」、「属性」、および「動作」に分類される。そして、「対象物」は、「部品」や「装置」などに細分化される。コンデンサやキャパシタなどは、部品に属する。 また「対象物」、「部品」などのシソーラスの木上のノードにあたるものが意味分類であり、木の末端に具体的な単語が入る。

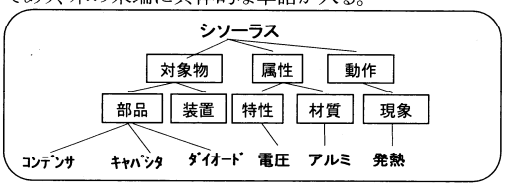

図 1シソーラスの例

2.1.4 情報機器の操作説明のガイダンスシステム

操作名や操作対象物などの用語辞書が必要になる。DVD レコーダであれば、操作名とは、録画、再生、編集などであり、操作対象物とは、リモコン、プレイリスト、テレビ番組などである。

2.2 分布仮説に基づく意味分類の自動獲得

2.2.1 既存技術

1 章で述べた分布仮説に基づく意味分類の獲得は、下記の3ステップからなる。

1. コーパスから、単語毎に、ある文脈において共起する言語表現を抽出する。
2. 単語毎の言語表現共起分布の差異から、単語と単語間の距離を定義する。
3. 単語間の距離に基づいて、クラスタリングする。

各々のステップで、どのような技術を使うかで様々な組み合わせがあるが、上流での重要な選択肢は、1の文脈(共起関係の種別)である。例えば、Hindle[5]では、同じ動詞の主語あるいは目的語にならない名詞同士は似ているという観察に基づいて、「ある動詞の主語あるいは目的語になる」という共起関係を用いている。また、Tokunaga[6]では、EDR コーパス(新聞、雑誌等)を対象として、日本語の「が」、「を」、「に」、「で」の各々の格の各要素の名詞と同士の共起関係から、格ごとに抽出したクラスタリング結果を比較検討している。

どのような共起関係を用いるかは、獲得しようとする意味分類の細かさにも依存する。一般には、獲得する知識が細かくなるほど、より細かい共起関係を用いる。例えば、鳥澤[7]では、「本を買う」と「本を読む」のような「用途とその用途の準備表現」の抽出を対象としており、並列動詞句、関係代名詞名詞句、接続詞「ため」などのパターンを扱っている。一方、Gale[8]のように、「drug は医療と犯罪など異なった意味に用いられる」といった曖昧性を解消する場合には、共起関係としては、その単語の周辺に出現する単語(文内共起、パラグラフ内共起)を扱っている。

現実的な応用では、意味分類辞書を最初から作るのではなく、既存の意味分類辞書を併用する場合も多い。浦本[9]や Tokunaga[10]では、ステップ 2 で得られた距離を用いて、その意味分類辞書に登録されていない単語を適切な階層に配置する方式を検討している。

2.3 本稿での考察範囲

本稿で実施する実験の問題意識、目的、および、方式について述べる。

2.3.1 問題意識

本稿では、2.1 節で述べた応用を想定して、以下の

問題意識に基づいた検討を実施する。

(1) 対象分野が、技術分野である。また、対象文書は、設計仕様書、製品マニュアル、FAQ、障害報告書などである。訓練セットとしては、既存研究の多くが対象とする新聞や辞典と比較して、「量が十分でない(1 万文程度)」、また、「テーブルやタイトルなどの完全な文でなく単語の羅列など、表層格をもった共起対は多くない」などの質の問題もある。

(2)意味分類の推定時に中間的に得られる言語知識が、情報検索、テキストマイニング、設計品質向上、および、ガイダンスなどに役立つようにしたい。

2.3.2 実験の目的

訓練データの規模は約 1 万文と小さいが、扱う素性数は約5万であることから、過学習になることが予想される。訓練データを十分に集めることが難しいという応用上の要請もあるので、「言語表現のパターン」や「既存の辞書知識」など、コーパスから獲得した知識を人手で補正・精査していくような方式を検討していきたい。

本稿では、上記の準備として、設計仕様書を対象として、意味分類推定方式の「分野依存性」、「スパース性」、および「訓練データへの依存性」を検討するための実験を行う。

(1)分野依存性

FA(Factory Automation)という分野を例にして、その分野での意味分類毎にその獲得に有効な素性、精度を落とす原因となる素性を分析する。

(2)スパース性

スパース性に起因する過学習について考察するため、素性のカウントの詳細度を変えた実験を行う。

(3)訓練データへの依存性

訓練データへの依存性をみるために、訓練データを 10 分割した交差検定により、精度の差のばらつきを分析する。

2.3.3 実験のベースとなる獲得アルゴリズム

実験に用いる推定アルゴリズムとしては、推定結果を解釈しやすいという以下にあげる特徴に注目して、Hindle[5]や Gale[8]らの方式を用いた。

(1)単語、文、文書間に陽に距離が定義できる。そのために、 多値分類や検索などに応用しやすい。

(2)素性を基底ベクトルとするユークリッド空間として、単語の特徴を直感的に解釈しやすい。

(3)低頻度の場合には、共起頻度統計処理からパターンマッチングに連続的に移行する解釈がとりやすい。

また、2.2.1 で述べたように、通常の共起分布による方法は3ステップ目でクラスタリングを行うが、本稿では、既存の意味分類階層への配置方式への適用を念頭において、ステップ3をクラスの推定問題として扱った。具体的には、推定対象の意味分類がもつ素性の共起分布から、単語と意味分類の距離を定義して、最も近い意味分類を推定結果とした。

3 共起分布に基づく単語の意味分類推定方式

本章では、「共起分布による方法」がもつ選択肢やパラメータを明確にするために、単語と素性が作る空間が各々共役関係にあるという観点から、単語間の距離を定式化する。 この定式化では、共起分布による単語間距離の計算プログラムにおける基本的なデータ構造と関数を整理することもねらっている。

Hindle[5]や Gale[8]では、単語と共起する素性の頻度分布により、単語間の距離を計算する。この距離の構成方法は、数学の関数解析での共役空間を用いた弱位相の構成方法[11]と似ている。というのは、単語と共起する素性の集合を、「単語上の関数の集合(単語空間の共役空間)」とみなして、単語空間に距離を定義しているからである。

以下では、この類似関係に従って、共起分布による方法の基本構成要素として、素性空間と単語空間を定義する。ここでは、数学的な厳密性にはこだわらず、類似度やダイバージェンスも距離と呼ぶ。

3.1 素性空間

3.1.1 素性の集合(素性タイプ)

素性集合は、どのような共起関係を想定するかを決めるもので、単語や素性の確率密度関数を定義する際の全体集合を規定する。

以下、代表的な共起関係の種別と、素性をどの程度の詳細度で見るかについて述べる。

(1) 共起関係の種別

共起関係は、文章の解析結果として得られる構造に応じて定義できる。表層的なテキスト解析を前提とすると、代表的な共起関係としては、文内共起(パラグラフ内などを含む)、文節内共起、係り受け共起がある。共起する素性の構成要素は、文内共起と文節内共起では単語(または、句)である。係り受けでは、直接的な関係としては、表層格、係、文節内の単語である。さらに、間接的な係り受け関係、例えば、兄弟文節(係り先文節に係る自分以外の文節)、孫文節(係り先が係る、また、係り元の係り元の文節など)を考えることもできる。

共起素性を特徴づける文法的な情報(辞書の情報、解析の結果得られる情報)をどの程度の詳細度で扱うかには任意性がある。以下では、本稿で扱う素性を列挙する。

表 1　共起関係毎の文法的な素性

共起関係	構成要素	素性の特徴付け
文内共起	単語	近接関係(n 文内、文内、n 単語以内など)、品詞、意味分類
文節内	単語	出現位置(前/後)、品詞、意味分類
係り受け	表層格 文節 文節内の単語	表層格、係り先/元/兄弟/孫、単語の品詞、単語の意味分類

(2)　素性集合の類別(商)、和、削除(射影)

直感的には、推定対象とする意味分類がより詳細になればなるほど、素性もより詳細にする必要があるはずである。この観点からは、詳細な解析結果情報そのものを訓練データにすればいいということになる。しかし、素性を詳細にすると個々の素性としての識別能力はあがるはずが、素性集合が大きくなり共起頻度はよりスパースになるので、過学習になりやすいというトレードオフがある。そこで、言語現象と抽出目標に即した抽象度の素性を選択することが重要になる。以下では、このトレードオフを念頭において、プリミティブな素性集合(解析結果の詳細情報)の商、直和、射影により新たな素性集合が構成できることを述べる。

(i)素性の同一視(商集合)と情報抽出パターン

共起関係を決めると、次に重要なことは、どの程度の抽象度で素性を数えるかを決めることである。例えば、係り受け関係で、「表層格は区別せずに、係り先か元かだけを扱う」や「同じ意味分類をもつ単語は、同じものとして扱う」などを決めることである。集合としては、商集合を扱うことに相当する。

「[部品名]+「の」+X の場合は、X は属性名の可能性が高い」などの情報抽出のパターンルールは、素性を定義する際に、表層格は区別するが、係り元の単語の意味分類が同じものを同一視していると解釈できる。

(i) 素性集合の直和

文節内共起と係り受け共起のように、互いに重複しないものは、集合の直和をとることにより、新たな素性集合を構成できる。ただし、和をとると、全体の頻度数を計算する際に、異なった観点からの集計頻度の和をとるので、確率分布としては直感的に解釈しにくくなる。

(ii) 素性の削除(射影)

訓練データが十分にないことに起因して、特定の素性が過学習の原因になることがある。そこで、「統計的な検定に基づいく低頻度」、あるいは、「あきらかに推定対象とは無相関の用語」などを、素性集合から削除する。

3.1.2 共起頻度行列、確率密度関数、素性関数

単語間に距離を入れるために、素性を単語(意味分類)から実数への関数として定義する。関数としての素性(素性関数)には種々の定義方法があり、具体例は次節で扱うが、いずれの場合も、基本となる量は、単語 wi と素性 fj の共起頻度行列 c(wi, fj)である。

素性集合(何を要素とするか、何を同じとするか)を決めると、共起の全体頻度が定まるので、単語 w の確率密度関数 p(w)、素性 f の確率密度関数 p(f)、データと素性の共起の確率密度関数 p(w,f)が定義できる。共起頻度行列の計算は時間がかかるが、商集合における確率密度で参照する共起行列は、同一視する元になる集合での共起頻度行列を流用できる。

素性関数は、これらの確率密度関数を用いて定義される。

3.2 単語空間

単語空間上の距離の構成方法を述べる。

3.2.1 素性関数によるデータ空間上の距離の定義

距離の構成方法の概略と、代表的な距離である Hindle の類似度と KL(Kullback-Leibler)擬距離について述べる。

(1) 構成方法の概略

素性を単語(意味分類)w から実数への関数 f(w)とすると、素性空間 F の共役空間としての単語間距離は、重み付のベクトル間の内積 K を用いて、下記のような式で定義される。

$$dist_{(F,K)}(w_1, w_2) = \sum_{f \in Fk} K(f_k(w_1), f_k(w_2))$$

上記の式をより直感的に理解するため、単語 d を素性を基底ベクトルとした素性空間 F 上のベクトルとみなして、素性関数 f(d)を内積<d,f>と表記すると、ベクトルの重み付内積 K が分離できる場合には、上記の式は下記のように書ける。

$$< w, g_s >= \sum_{fk \in F} K(s,k) < d, f_k >$$

$$dist_{(F,K)}(w_1, w_2) = \sum_{gx \in G} < w_1, g_s > \bullet < w_2, g_s >$$

上記の式では、核 K により d を素性空間 F 上の点から素性空間 G に上の点に写像し、素性空間 G 上の内積をとることで、単語間の距離を定義していることになる。例えば、Deerwester[12]らの潜在的意味解析(LSI : Latent Semantic Analysis)による単語間距離では、核 K を構成する際に、特異値分解を用いていると解釈できる。

(2) 距離定義の具体例(Hindle 類似度と KL 擬距離)

(i) Hindle の類似度

$$fk(w) = \log_2(p(w, fk) / p(w) \cdot p(fk))$$

$$dist_{(F,\ hindle)}(w1, w2) = \sum_{fk \in F} u(fk(w1), fk(w2)) \cdot \min(|fk(w1)|, |fk(w2)|)$$

但し、$u(x, y) =$ if$(x * y > 0)$ then 1 else 0

(ii) KL 擬距離

ナイブベイズ法による推定で、事前確率なしの場合は、KL 距離を用いることと同じである([13])。

$$fk(w) = p(fk \mid w) = \frac{p(fk, w)}{p(w)}$$

$$dist_{(F,\ kl)}(w1, w2) = \sum_{fk \in F} fk(w1) \cdot \log_2\left(\frac{fk(w1)}{fk(w2)}\right)$$

(3) 複数の素性タイプを反映した距離

(2)により、素性空間 F 毎に距離が定義できる。単語の意味分類の推定というタスクでは、どの素性の距離を重視するか、また、異なる素性空間 F の距離を組み合わせるかという課題がある。

複数の素性タイプを反映した距離を定義するには、3.1.1 で述べた素性集合の直和(⊕)を用いる方法と、本項の距離の合成(和や最小値をとるなど)がある。記号で書くと、$d_{(F \oplus G,K)}$と $d_{(F,K)} + d_{(G,K)}$という選択肢がある。

3.2.2 単語空間の構造の素性空間への反映

「単語空間と素性空間が互いに共役である」、さらに、「素性の主な構成要素は単語である」という性質を用いて、前項で定義した単語間の距離を素性空間の構造に反映させるような Co-Training ができる。

(1) 単語の意味分類による素性クラスの導入

素性の主構成要素は単語なので、単語空間で、単語の意味分類が推定できると、素性を特徴づける情報として利用することができる。例えば、3.1 の情報抽出パターンを構成できる。

(2) 素性間の距離

データ空間を素性空間の共役空間としてとらえると、素性間にも距離が定義できる

4 実験

4.1 対象データ

実験には、設計品質向上支援システムへの適用を想定し、実際に FA 分野の設計業務で使用した設計仕様書 8 冊から抽出した 15,302 文(延べ単語数 180,320 語)を使用した。学習に用いた共起種別毎の素性の異なり数、頻度を表 2に示す。

表 2 学習に利用した素性の異なり数と頻度

分類	素性タイプ	異なり数	頻度
文節内共起	文節内共起単語	10,892	113,862
係り受け共起:表層格(*)区別	係り受け先文節中の単語(係り受け元)	22,971	180,320
	単語(係り受け先)		
	係り受け元文節中の単語(係り受け元)	22,904	180,320

(*) 表層格: ガ格,ヲ格,ニ格,ト格,デ格,カラ格,ヨリ格,ヘ格, マデ格,ノ格,連体,連体:ヤ,連用,同格連体,文末, 未格,同格未格,無格,無格従属:-1,括弧並列, 中断線,NONE

4.2 想定するシソーラス

実験で想定するFA分野の意味分類の階層図を図 2 意味分類の階層図に示す。

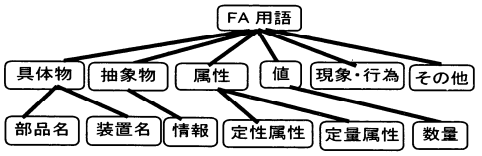

図 2 意味分類の階層図

図 2で、「装置名」、「部品名」、「定量属性」、「現象・行為」は、それぞれ表 3に示す詳細な意味分類を含んでいる。

表 3 詳細な意味分類

意味分類	詳細な意味分類
部品名	部品名,回路名,端子,部位
装置名	システム名,製品名,装置名,機器名,型番
情報	信号,データ
定性属性	特性,材質,状態,機能,用途,要因,条件,形式・方式
定量属性	物理量,時間,期間
現象・行為	動作,現象,操作

今回の実験では、あらかじめ対象データの形態素解析結果を複合語解析して抽出した用語辞書を評価対象に用いた。用語辞書の総数は 2,106 語である。用語辞書の意味カテゴリごとの語数を表 4に示す。

表 4 実験対象の用語辞書の語数

意味分類	語数	出現頻度 2 以上
部品名	528	207
装置名	162	55
情報	101	40
定性属性	417	148
定量属性	367	164
現象・行為	531	232
合計	2,106	846

4.3 実験方法

実験では、形態素解析器として Juman、係り受け解析器として KNP を用いる。なお、複合語解析はシソーラス作成において重要な課題であるが、今回は意味分類推

定の課題を扱うため、4.2 節で述べた用語辞書を Juman のユーザ辞書に登録して実験する。

データ量が約 1.5 万文と少なく偏りが予想されるため、10 分割交差検定法(用語辞書を 10 等分に分割し、訓練 9、テスト1の比率で交差検定)により意味分類推定を行う。素性タイプおよび意味分類ごとの意味分類推定結果の適合率と再現率の平均値を精度比較に用いる。

なお、比較対象の意味分類は、表 4の意味分類のうち、ある程度語数の単語多い【部品名】、【装置名】、【情報】、【定量属性】、【定性属性】、および、【現象・行為】とした。

5 実験結果と考察

Hindle 類似度による推定結果を表 5(Hindle 類似度)に示し、KL 擬距離による推定結果を表 6(KL 擬距離)示す。ただし、頻度 2 以上の単語がある程度少ない【装置名】と【情報】は、交差データの精度のばらつきが大きいため、表中からは省いた。また、意味分類毎に有効な素性を知るために、単独素性による距離と、複数素性を反映させた距離毎に、最も精度がよいセルに斜をかけた。

Hindle 類似度と KL 距離では、おおまかな傾向は似ているため、以下では、Hindle 類似度をとりあげて、「素性タイプ/意味分類毎の推定結果」、「素性集合の直和による距離」、「素性集合の商による距離(係り受け共起における表層格の区別有無、文節内共起の前方後方の区別有無)」について考察する。

5.1 素性タイプ・意味分類毎の推定結果

(1) 結果の概略

表 5によると、以下の傾向がある。

・体言的な性質をもつ【部品名】は、素性タイプ「文節内後」と「係り先」の抽出精度がよい。

・用言的な性質をもつ【現象・行為】は、素性タイプ「係り元」の精度がよい。

・体言的な性質と用言的な性質を併せ持つ【定性属性】や【定量属性】は、【部品名】と【現象・行為】の中間的な位置づけである。

・【定量属性】は、単位などの手がかりがあり、素性タイプ「文節内後」の抽出精度がよい。【定性属性】は、種々の用語が含まれるために、いずれの共起関係も精度がよくない。

(2) 素性タイプ毎の推定有効な素性

表 7に、【部品名】、【定量属性】、【定性属性】、および、【現象・行為】毎に、推定に有効な素性と表現例を示す。

(3) 推定の失敗例とその考察

抽出漏れ(再現率の低下要因)は、共通する素性がないものが多いためである。

抽出ゴミ(適合率の低下)の多くは、人手で見ると意味分類を特徴づけると思われない素性にひきづられたものであり、スパースさに起因する過学習によるものと思われる。訓練データを増やすのが重要だが、以下では、素性の追加と選択により、精度をあげる可能性について考察する。

(i) 素性の追加

信号線「DWN」は、【定性属性】だが【部品名】と推定していた。「アドレスに DWN を使用する」という文の「を使用する」という素性から、【部品名】と推定している。この例では、直接の係り受け関係にある「使用する」だけ用いたのでは、意味分類の区別は難しい。しかし、「アドレスに」の部分を参照すると、【属性】であることを推定できる可能性がある。 すなわち、1段の係り受けだけでなく、係り先文節に係る自分以外の文節(兄弟文節)を考慮することにより、精度を向上できる可能性がある。この素性を導入することは、言語現象としては、係り先を中心とする係り受けの組(格フレーム)としての類似性を考慮することである。言語処理のための語彙情報として動詞の格フレームや選択制限を扱う方式を、意味分類推定に導入することは今後の課題である([14][15][16])。

(ii)素性の削除

今回は、文節 A が文節 B に係る際に、文節 A を構成するすべての単語が、文節 B に係るという共起関係をもつとしていた。そのため、「DIN レール取り付けの際は」という文から、「DIN レール」が「際は」に係るという共起のため、「部品名」が正解のところを「定性属性」と誤った。「過電圧の際は」のように、「【属性】の際は」という共起が訓練データにあるためである。「DIN レール」は「取り付け」に係っているのであって、「際は」に係っているわけではない。したがって、この例からは、係り元としては「文節の主辞となる単語」に制限する方がように見える。

5.2 素性集合の直和と距離の合成との比較

表 5によると、「素性集合の直和の距離」も「素性集合毎の距離の和」は、最も精度のよい単独の素性集合の距離よりも精度がよかった。単独の場合は、不得意な意味分類により、得意な意味分類の精度も落ちているので、他の素性を参照することにより、全体の精度があがったためと思われる。

また、類似度の最大値の精度はよくない。類似度の絶対値にあまり意味がなく、同じ意味分類での相対的な値だけが意味をもつためだと思われる。

異なった素性タイプを反映した距離の定義については、今後の課題である。

表 5　素性タイプ/意味分類毎の推定結果(hindle 類似度)

共起種別		素性タイプ	部品名		定量属性		定性属性		現象・行為	
			適合率	再現率	適合率	再現率	適合率	再現率	適合率	再現率
単独素性	文節内共起	文節内前 Br $d(Br)$	40.6	36.7	47.8	28.2	18.1	6.7	34.9	23.5
		文節内後 Bs $d(Bs)$	40.8	30.9	59.0	32.9	27.1	16.1	50.6	31.7
	係り受け共起(表層格区別)	係り元 Db $d(Db)$	38.3	25.6	41.6	37.6	27.3	16.4	57.5	52.5
		係り先 Df $d(Df)$	54.3	46.8	27.4	14.9	27.4	14.9	56.7	43.4
複数素性	素性の直和 $d(Br \oplus Bs \oplus Df \oplus Db)$		58.1	58.0	52.4	54.7	27.4	16.2	51.6	59.9
	類似度の和 $d(Br)+d(Bs)+d(Df)+d(Db)$		60.6	51.3	65.7	52.4	37.6	17.5	42.0	72.3
	類似度の最大 $\max(d(Br), d(Bs), d(Df), d(Db))$		39.6	29.8	57.2	30.5	27.6	10.8	35.9	69.9

表 6　素性タイプ/意味分類毎の推定結果(KL 擬距離)

共起種別		素性タイプ	部品名		定量属性		定性属性		現象・行為	
			適合率	再現率	適合率	再現率	適合率	再現率	適合率	再現率
単独素性	文節内共起	文節内前 Br $d(Br)$	28.7	62.0	40.3	23.2	16.4	8.1	39.9	24.4
		文節内後 Bs $d(Bs)$	31.4	62.0	56.9	37.8	26.7	13.4	50.3	35.0
	係り受け共起(表層格区別)	係り元 Db $d(Db)$	38.3	51.3	42.9	46.1	24.1	16.4	65.6	47.4
		係り先 Df $d(Df)$	37.9	64.6	61.2	51.2	35.0	21.5	62.7	41.8
複数素性	素性の直和 $d(Br \oplus Bs \oplus Df \oplus Db)$		54.5	61.0	55.9	63.1	26.4	17.5	58.2	58.2
	類似度の和 $d(Br)+d(Bs)+d(Df)+d(Db)$		52.4	46.5	56.6	54.6	26.4	17.6	45.9	59.0
	類似度の最大 $\max(d(Br), d(Bs), d(Df), d(Db))$		47.6	36.9	46.9	47.5	14.4	11.0	44.3	48.4

表 7　意味分類毎の推定に有効な素性(例)

意味分類	単語	素性 (素性タイプ)	例文
部品名	計装アンプ	の-電源　(係り先)	計装アンプの電源
	ツェナーダイオード	を-設ける　(係り先)	ツェナーダイオードを設ける
	5V 電源	ASIC　(文節内後)	5V 電源 ASIC メモリコネクタ
	DIN レール	経由　(文節内後)	DIN レール経由でノイズが回り込む
定量属性	G-S 間電圧	が-降下する　(係り先)	G-S 間電圧が 1.5V まで降下する
	応答速度	が-遅い　(係り先)	応答速度が遅い
	許容損失	の-確認　(係り元)	ダイオードの許容損失の確認
	インピーダンス	Ω　(文節内後)	インピーダンス 35Ω
	ツェナー電圧	VZ　(文節内後)	最小ツェナー電圧 VZ=11.38V
	コンデンサ環境温度	℃　(文節内後)	実使用時のコンデンサ環境温度(℃)
定性属性	実装条件	明確だ　(係り先)	使用部品、実装条件が明確になった
	経年変化	の-影響　(係り先)	経年変化の影響が小さい
	低消費電流タイプ	の-発振器　(係り先)	低消費電流タイプの発振器を採用する
	過電圧異常	を-検出　(係り先)	A 系で過電圧異常を検出する
	空き	ポート　(文節内後)	この現象が空きポート全てに発生した
現象・行為	ON	TR-が　(係り元)	TR が ON する
	逆接続	極性-を　(係り元)	誤って電源の極性を逆接続する
	ドライブ	フォトカプラ-を　(係り元)	AD コンバータはフォトカプラをドライブできない
	遮断	回路-を　(係り元)	過電流時に回路を遮断する

5.3 素性の詳細度(商集合)の比較

(1)係り先共起における表層格の区別

抽出精度の比較的よかった「係り先共起」による意味分類【部品】の推定において、表層格を区別する場合としない場合との精度差異(hindle 類似度の場合)を表 8 に示す。同様に、「係り先共起」による意味分類【現象・行為】の推定結果における差異を表 9に示す。

表 8 係り先共起における表層格の区別有無の比較

	【部品】	
	適合率	再現率
区別する	54.3	46.8
区別しない	48.0	46.1

表 9 係り元共起における表層格の区別有無の比較

	【現象・行為】	
	適合率	再現率
区別する	57.5	52.3
区別しない	40.5	58.2

いずれも、表層格を区別した方が適合率がよくなり、再現率は同じか悪くなる傾向にあるように見える。再現率の差については、データがスパースなためか、あるいは、意味分類と推定方式の性質なのかについての分析は、今後の課題である。

(2) 文節内共起における前方と後方出現の区別

抽出精度の比較的よかった「文節内の後方共起」による意味分類【定量属性】の推定において、前方と後方を区別しない文節内共起との精度差異を表 10に示す。

表 10 文節内共起における前方・後方の区別

	【定量属性】	
	適合率	再現率
文節内共起 (後方)	59.0	32.9
前方/後方区別せず	50.8	36.5

傾向は、係り受けの表層格の区別ありなしの場合と似たような傾向にある。

6 おわりに

新聞記事等を中心に検証されてきた単語共起分布に基づく意味分類推定方式を、約1.5 万文規模の FA 設計仕様書に適用した実験結果を報告した。実験の結果、部品名や装置名のような体言的意味分類では、共起関係として「係り受け先」が有効であり、作用や現象のような用言的な意味分類では、「係り受け元」が有効であることを確認した。

今後の課題としては、応用、自然言語処理、および、機械学習アルゴリズム適用毎に下記がある。

(1)応用上の問題

・人手修正も含めたシソーラス構築の評価方式

・既存の電子化された言語知識の活用方式

・意味分類辞書の網羅度/正確さの応用への影響

(2)自然言語処理の問題

・複合語処理や未知語処理を用いた専門用語の同定

・格フレーム辞書の獲得方式の併用

・構造をもつ素性の導入

(3)機械学習アルゴリズム適用の問題

・SVM などの他の学習アルゴリズムの比較

・統計的な検定に基づく素性の削除

【参考文献】

[1] 国立国語研究所: 分類語彙表, 秀英出版 (1964)

[2] (株)日本電子化辞書研究所:EDR 電子化辞書仕様説明書, (1993)

[3] 徳永 健伸:『単語と辞書』 4 章 辞書と情報処理, 岩波書店, pp.185-190 (2004)

[4] K.Tanigaki, T.Hirano, Y. Okada: "Push-Style Guidance System for Technical Document Writing". Proc.8th International Conferance on Document Analysis and Recognition, pp. 725-729 (2005).

[5] D.Hindle: "Noun Classification from Predicate-Argument Structures", Proc.21th Annual Meeting of ACL, pp.268-275 (1990).

[6] T.Tokunaga, M.Iwayama, and H.Tanaka:"Automatic thesaurus construction based on grammatical relations", Proc.14th IJCAI, pp.1308-1313 (1995).

[7]鳥澤 健太郎: 対象の用途と準備を表す表現の自動獲得, 自然言語処理 Vol.13 No.2 pp.125-144 (2006).

[8] W.Gale, K.W.Church, and D.Yarowsky: A method for disambiguating word senses in a large corpus. Computers and the Humanities 26: pp.415-439(1992).

[9] 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 Vol37 No.12 pp.2182-2189 (1996).

[10] T.Tokunaga, A. Fujii, M.Iwayama, and H.Tanaka: "Extending a thesaurus by classifying words". Proc.ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources, pp. 16-21 (1997).

[11] コルモゴロフ,フォミーン:『函数解析の基礎 上』 第 4 章 線形汎函数と線形作用素, 岩波書店, pp118-175 (1976).

[12] S.Deerwester, S.Dumais,G.W.Funas: T.K.Landauer, R.Harshman: "Indexing by latent semantic analysis". Journal of the American Society for Infomation Science, 41(6) pp.391-407 (1990).

[13] 渡辺澄夫:『データ学習アルゴリズム』 2.1 節 データと学習, pp.23-pp35 (2001)

[14] 松本 裕治:『意味』 4 章 意味と計算, 岩波書店, pp.126-167 (2004)

[15] 大石 亨, 松本 裕治: 格パターン分析に基づく動詞の語彙知識獲得, 情報処理学会論文誌, Vol.36, No.11, pp.2597-2610 (1995).

[16] 河原大輔, 黒橋禎夫:格フレーム辞書の漸次的自動構築, 自然言語処理, Vol.12, No.2, pp.109-131 (2005).