修 士 論 文

主成分分析を用いた未登録語のシソーラスへの追加

指導教官　奥村学 助教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

鈴木勝仁

1998 年 2 月 13 日

要 旨

自然言語処理 (NLP) において, シソーラスは, 最も重要な知識源の 1 つである. しかし, NLP システムの視点からみると, シソーラスは, 語彙が不十分である. そのため, シソーラスの拡張が望まれる. 本研究では, 共起データを手掛かりに, シソーラス上に存在しない名詞 (未登録語) の配置先を推定する方法を提案する. 本研究の特色は, 単語の配置先推定に悪い影響を与えるスパースネスの問題と, 多義語の問題に対処するために, 主成分分析とよばれる統計的手法を用いることである. 実験では正解の範囲を緩めることで, 多義語の問題に対して, 提案方法の精度は 73.7%であり, 従来の配置先推定の方法は最大のもので 66.2%であった. また, スパースネスの問題に対して, 従来は扱えなかったが, 提案方法は 32.5%の精度を得た.

目 次

図 目 次

表 目 次

第 1 章

はじめに

1.1 目的と背景

単語間の上位下位関係や同義関係を記述したシソーラスは, 自然言語処理分野における最も重要な知識源の 1 つである. シソーラスは, 多義性の解消や類義語の獲得など, 多くの研究で利用されている. これまでのシソーラスは, 手作業で構築され, かなり広範囲な語彙が登録されている. しかし, 特定領域で扱われる自然言語処理システムの立場からすると, シソーラスは, 不必要となる一般的な単語を多く含み, 必要となる領域特有の単語を含んでいないため, 語彙が不十分であると考えられる. また, シソーラスの構築後は, 新たに単語を加えることはなく, 語彙の変化に対応できないなどの問題がある. これら問題を解決するために, 本研究では, シソーラスに存在しない単語 (未登録語) の配置先を推定する方法を提案する.

未登録語をシソーラスへ配置する方法として, 文字情報に基づくものがある [3]. 漢字は, 表意文字であり, それは語の意味に関わる有用な情報になる. 日本語シソーラスを対象とするとき, 文字 (漢字) の情報を使うことにより, 比較的容易に意味的な情報を得ることができる. しかし, この文字情報は, 未登録語がカタカナや簡略化された語の場合には適用できないため, 有効ではない.

他の方法として, コーパスから得た共起データに基づくものがある. 例えば, 格関係にある名詞と動詞の共起データによって, シソーラス中の名詞と未登録語を動詞要素 (格と動詞の組) で表現でき, それら名詞と未登録語の動詞要素の似かより具合で, 配置先を決めることができる. 共起データに基づく研究には, 浦本 [1] と, 徳永 [2], 中山 [3] のがある. 浦

本の研究では, 人がシソーラスを構築したときに使われた情報 (分類基準) が記されていないために, その分類基準を共起データから抽出し, その分類基準とコサイン距離を用いて, 未登録語を ISAMAP という上位下位シソーラスに配置する方法を提案した. 浦本の方法では, 未登録語は幾つか隣接したノード集合の中に配置される. また, 徳永の研究では, 未登録語がシソーラスのノードに属す確率を推定するために共起データを用いる. 未登録語は, その確率を基にして分類語彙表という分類シソーラスのノードに配置される. 徳永の方法では, 未登録語に幾つかの配置先を提示する. そして, 中山の研究では, 浦本とは異なる方法で, 分類基準を共起データから抽出し, その基準を用いて, 未登録語を分類語彙表に配置する方法を提案した. 中山の方法でも, 未登録語に幾つかの配置先を提示する. 共起データを用いることで, 単語間の類似度計算が可能になり, 未登録語の配置先が推定できる. しかし, 共起データを用いて類似度を計算する場合, 以下の 2 つの基本的な問題がある.

スパースネスの問題 : コーパスから得られる共起データは不完全であることや, 動詞概念の表現方法は普通幾つもあるために, 未登録語の動詞要素と, 配置先の単語の動詞要素とが一致しない可能性がある.

多義語の問題 : 多くの単語[1]は, 複数の意味があるため, 未登録語の動詞要素が, 配置先と関係のない単語の動詞要素と一致する可能性がある.

スパースネスの問題において, 浦本の研究では, ノードに直接属す単語と未登録語との類似度を計算するため, スパースネスの問題が生じやすいと考えられる. 徳永の研究では, 未登録語と単語の類似度ではなく, 未登録語と単語集合との類似度を計算するため, 共起データの不足分を補うことができるが, 従来のスパースネス解消方法と比べると弱い. そのスパースネス解消方法とは, シソーラスの構造を用いて, 意味的に似た動詞を幾つかまとめるというものである. この解消方法を中山の研究では用いている. しかし, この解消方法では, まとめた動詞が複数の意味をもった語 (多義語) になる可能性がある. その語の多義性を解消することは, まとめる前の動詞に戻すことであるためにできない.

多義語の問題においては, 従来の研究では考慮されていない. 単語の語義曖昧性解消の方法は, 辞書で多義語を複数の意味にわけ, それら意味の中から適切なものを選ぶことを行うが, 十分な精度がないため有効な方法でない.

[1] 本研究では, 名詞の多義性を考慮せず, 動詞の多義性だけを考慮する.

本研究では, 従来のスパース性解消の方法や, 多義性解消の方法に問題があるために, 主成分分析とよばれる統計的手法を用いて, スパースネスの問題と, 従来の配置先推定の研究で扱われなかった多義語の問題に対処する.

1.2 論文の構成

本稿は, この章を含め 6 つの章からなる. 第 2 章では, 従来の研究において, シソーラスに存在しない単語をどのような情報を基にして配置するのかを説明する. そして, その情報を手掛かりにして配置先を推定する場合に生じるスパースネスの問題と多義語の問題について述べる. 第 3 章では, スパースの問題と多義語の問題をどのように解決するのかを述べる. 第 4 章では, 未登録語の配置先を同定するモデルについて述べる. 第 5 章では, スパースネスの問題及び多義語の問題にどのくらい対処できたかを検証及び, 過去の研究との精度を比較を行う. 最後に, まとめを行う.

第 2 章

未登録語の配置先推定問題

未登録語の配置先推定問題は, 単語の語義の曖昧性解消問題 [6, 7] の変形とみなすことができる. 配置先推定問題と曖昧性解消問題の違いは, 解空間の広さである. 従来の多義性解消では, あらかじめ語義の候補はわかっているので, (例えば,bank の語義は「土手」か「銀行」である), その語義のうちのどれか (一般的には, 高々10 数個) を決定することになる. 一方, 未登録語の場合, あらかじめ語義はわかっていないので, 解の候補はシソーラス上の全ての単語であり, シソーラス上での, 正確な位置を示すのは非常に困難である. そのため, 従来の未登録語の配置先推定のアルゴリズムでは, 未登録語に対して, 最もらしい語義を幾つか呈示することを行なう. 本研究も従来と同様に最もらしい語義を幾つか呈示する.

2.1節では, シソーラスについて述べ, 2.2節では, まず, 従来の配置先方法を説明し, そして, その方法で用いる知識源に生じる問題について説明する. 2.3節では, 提案方法の概要を述べる.

2.1 シソーラス

シソーラスとは, 単語を意味によって分類配列し, 各単語について同義語や, 類義語, 上位語, 下位語, 反義語, 対義語などを記述した辞書である. シソーラスの作成を最初に試みたのはロジェ(P.Roget) である. ロジェは, まず単語を次の 6 種類のクラスに分類した.

1. abstract relation
2. space

3. matter

4. intellect

5. volition

6. emotion, religion and morality

そして, それらをさらに詳しく細分化するということを行ない, 結果として木構造の分類体系を作成した. 分類された最小の単語のグループは木構造の葉の部分に存在するという構造である. このようにロジェのシソーラスは上位語/下位語の体系というよりは, 同義語関係に中心をおいたものであった. そしてその主たる目的は, 人間が文章を書く場合に自分の考えを最も適切に表現する単語を探すという作業の手助けをすることであった.

シソーラスは, 単語を分類配列したものであるが, 分類配列されたデータというものは, 計算機にとって最も処理しやすい形式のデータである. すなわち, シソーラスという意味表現 (意味定義) の方法は本質的に計算機処理に適したものであるといえる. 情報検索の分野では, 用語のずれによる検索の失敗を同義語関係の情報によって防ぐという目的で早くからシソーラスが利用されてきた. 一方, 最近の自然言語処理では, 単語と単語の間の類似度を計算する上でシソーラスが重要な役割を果たしている. これは, シソーラスの木構造において近くに存在する単語同士は類似度が高いが, 距離が離れるに従って類似度が小さくなると考えることができるからである.

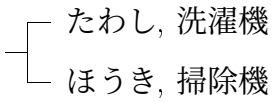
シソーラスには, 分類シソーラスと上位下位シソーラスという 2 種類のタイプがある.

分類シソーラス : 単語の分類が中心であり, 木構造の葉の部分だけに実際の単語が存在するもの

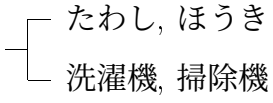
上位下位シソーラス : 単語の上位下位関係を中心において, 木構造の内部のノードにも単語が対応するもの.

現在, シソーラスの作成は, ほとんど手作業で行なわれている. 格フレームを作成する場合には, 動詞がどのような名詞を格要素とするかを実際の言語データによって調査することができるので, 作業は比較的進めやすい. しかし, シソーラスの場合には, 単語の類義関係や上位下位関係が実際の言語表現として示されることはほとんどない. 単語の類似用例を収集して参考することはできても, 同義関係や上位下位関係などの決定は, 最終的には人間の直観に頼るしかないが, これは非常に困難な作業である.

さらに, 単語を分類する際には単語に対する視点の問題を考慮する必要がある. すなわち, 言葉には別の視点から見れば, 別の意味の側面に焦点があたるという性質がある. 例えば, 「たわし」, 「洗濯機」, 「ほうき」, 「掃除機」という 4 つの単語は, 道具という視点からは,



と分類されるが, 動力という視点からは



と分類されるだろう. このように, シソーラスは実際には単純な木構造とはならず, あるノードに複数の親ノードが存在するということが起こりうる. このような視点の問題まで考慮してシソーラスを人手で作成することは, 多大な労力が伴う. そのため, 計算機によるシソーラスを自動構築/拡張が望まれる. 本研究では, 既存のシソーラスへの未登録語の登録による拡張を対象とする.

最後に, 電子化されたシソーラスを挙げる. 分類シソーラスとして, ロジェのシソーラス [12], 分類語彙表 [13] などがある. 上位下位シソーラスとして, プリンストン大学の WordNet[15], EDR の概念体系辞書 [16], 東京工業大学の ISAMAP[14] などがある.

2.2 従来の配置先推定の方法

従来の配置先推定の研究 [1, 2, 3] は, 文字情報や共起情報を手掛かりにして, 未登録語の適切な配置先を推定する. 2.2.1節では, 文字情報について述べ, 2.2.2節では, 共起情報について述べる.

2.2.1 文字情報

日本語の単語の場合, サフィックス (suffix) が上位語になっている場合が多い. 例えば,

経済問題 → 問題　　通貨危機 → 危機

専門用語の場合も同様である.

ＩＩＲフィルター → フィルター　　ＩＥＥＥ規格 → 規格

このような日本語の単語の性質より, 以下のような方法が考えられる.

> ある単語がシソーラスに存在しない場合, その単語のサフィックスのうち, シソーラス中に存在する最も長いもので代用する.

日本語は基本的に head final の言語であり, それは複合語についても成り立つ. 従って, 多くの未登録語の場合, サフィックスに関する検索が行なえる. しかし, 一部の未登録語のおいては, プレフィックス (prefix) に関する検索の方がよい場合がある. 例えば, 「消化器」の配置先を推定する場合, 「器」には “道具” 程度の意味合いしかないが, もし「消化」に関するノード (語義) があるならば, 「器」に関するノードよりも, 「消化」の方が適切な配置先と考れる. サフィックスだけではなくプレフィックスも考慮した方法として中山のもの [3] がある.

文字情報を用いた方法は, 低コストでありながら, 高精度で配置先推定が行なえる長所があるが, 次のような単語に適用できない場合が多いという短所がある.

- カタカナ, ひらがな, アルファベットのみから構成される単語
- 簡略化された単語 (e.g., 住専/住宅金融専門社)
- 慣用的な熟語 (e.g., 伝家の宝刀, 酒池肉林)

2.2.2　共起情報

共起情報とは, ある単語がどのような単語と共起するのかというものである. 共起情報は, 一般的に, テキスト形式のコーパスから獲得される. 未登録語の配置先推定において共起情報を用いる場合, シソーラス中の単語 (名詞) と未登録語 (名詞) に対して, 共起する単語 (共起語) を割り当て, それら共起語の振舞いが似ている度合の高いノードから順に未登録語の配置先の候補とする.

この節では, まず, コーパスから獲得される共起情報のデータ形式について述べ, 次に, 似ている度合を計算する方法 (類似度) について述べる.

共起データ

コーパスから獲得する共起情報のデータ形式は, 2 種類ある. 1 つは, Yarowsky の語義の曖昧性解消の研究 [6] で用いられたような, 単語とある範囲内で共起する単語の集合に着目した形式であり, もう 1 つは, 単語間の係り受け構造に着目した形式である. 従来の配置先推定の研究では, 単語間の係り受け構造に着目した形式で獲得する. 単語の集合ではなく, 単語間の係り受け構造を用いる理由を以下に述べる.

まず, あらかじめいくつかの語義がわかっている単語の場合, 語義の候補は, もちろん語義の数だけあり, この中から正しい語義を決定するわけである. 一方, 未登録語の場合, その単語の意味の候補は, シソーラス上のノードすべてである. 語義の推定を妨げるノイズを除去するために, より制限の強い関係を用いる.

第 2 の理由は, 格関係のような関係要素によって明示される単語間の関係は, 単語分類の視点として用いることができるためである. 例えば, シソーラス上で「飛行機」と「船」が並列ノードのとき, この 2 つの単語およびそれらの下位語を分類するのは, 「飛行機が飛ぶ」というが「船が飛ぶ」とは言わないという現象である. つまり「~が飛ぶ」という (係り受け関係) は, 「飛行機」を分類するという視点になり得る.

最後に, 少しテクニカルな理由だが, 単語集合によって作られる行列は, 同じ共起語でも観測地点が異なる可能性があるために, 係り受け構造より, 行列の等質性が悪いと考れる. 等質性とは, 行列の各列あるいは各行の対比が可能でなければならないことである.

従来の配置先推定の研究で用いられる共起データの形式は, 多くの場合, 格関係にある名詞と動詞の関係がコーパスにどれくらいの出現するのかというものであり, 以下のような形式で獲得する.

共起データの形式 : (名詞, 格助詞, 動詞, 頻度)

以降, 本論文では, 格助詞と動詞の組を**動詞要素**と呼ぶ.

共起データを獲得したことによって, 名詞を動詞要素で表現することができる. 例えば, 以下のような共起データが得られることで, 表 2.1のような行列表現が可能となる.

共起データの例
（連絡船　，が，出る，2）
（連絡船　，に，乗る，3）

（ジープ　，が，走る，1）
（イーグル，が，飛ぶ，1）

名詞 \ 動詞要素	が出る	に乗る	···	が走る	···	が飛ぶ	···
連絡船	2	3	···	0	···	0	···
ジープ	0	0	···	1	···	0	···
···	·	·	···	·	···	·	···
イーグル	0	0	···	0	···	1	···
···	·	·	···	·	···	·	···

表 2.1: データ行列の例

この表において, ジープは, “が走る” という動詞要素がコーパスに 1 回出現し, “に乗る” などの動詞要素は, コーパスに出現しなかったことを示している.

名詞を動詞要素で表現することによって, その動詞要素に基づき, 名詞間の類似度計算が可能となる. そして, シソーラス上の名詞と共起データの名詞との対応関係をとることによって, シソーラス上でも名詞間の類似度計算が可能となる.

共起データを用いた配置先推定の方法

従来の配置先推定では, 共起データを用いて, 単語と, 単語 (又は単語集合) との意味的近さを求めるモデルが提案されている. それらモデルを用いた配置先推定の方法例を簡単に述べる.

浦本の方法 [1] では, まず, 配置先推定の前処理として, シソーラス上の各ノードに分類基準という情報を付与する. その後, 配置先推定の処理として, 未登録語とノードとのコサイン距離 (2.1式) による類似度を計算し, その類似度が或閾値を超えたノードを一時的な配置先の候補とする. そして, それら候補の中から分類基準などの優先情報を用いてさらに候補を絞りこむ. その結果, 未登録語は, 幾つか隣接したノード集合の中に配置される.

$$\mathrm{cosine}(X,Y) \;=\; \frac{\sum_{i=1}^{m} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{m} x_i^2}\sqrt{\sum_{i=1}^{m} y_i^2}} \tag{2.1}$$

ここで, X と Y は単語であり, m は動詞要素の種類である. 浦本が提案する分類基準とは, シソーラス上の各々のノードと他ノードとの差異を計る情報であり, その情報は共起データから抽出される. ノードを nd とすると, そのノードの分類基準は, 以下のように抽出される.

$$\text{typicalness}(\text{nd}, v) \quad = \quad \frac{\sum_{w \in c} f_r(w, v)}{\sum_{w \in c \cup c'} f_r(w, v)} \tag{2.2}$$

ここで, c は nd とそれの子孫の単語集合であり, c' は nd の兄弟とそれら子孫の単語集合である. $f_r(w, v)$ は, 単語 wと動詞要素 vが共起出現している頻度である. tyicalness$>$ 0.5 の条件を満たす動詞要素 vがノード nd の分類基準となる. 分類基準による優先方法は, 未登録語が分類基準を満たす (所有する) ノードほど配置先として優先する.

また, 徳永の方法 [2] では, 共起データを基にして, 未登録語 wが, ノード c に含まれる確率 $P(c|w)$ を計算するモデルを提案した.

$$P(c|w) \quad = \quad P(c) \sum_{v_i} \frac{P(V = v_i|c) P(V = v_i|w)}{P(V = v_i)}. \tag{2.3}$$

確率 $P(c|w)$ の中の全ての確率は, 以下の式に基づいて推定される.

$$\hat{P}(V = v_i|c) \quad = \quad \frac{\sum_{w \in c} f_r(w, v_i)}{\sum_i \sum_{w \in c} f_r(w, v_i)} \tag{2.4}$$

$$\hat{P}(V = v_i|w) \quad = \quad \frac{f_r(w, v_i)}{\sum_i f_r(w, v_i)} \tag{2.5}$$

$$\hat{P}(V = v_i) \quad = \quad \frac{\sum_w f_r(w, v_i)}{\sum_i \sum_w f_r(w, v_i)} \tag{2.6}$$

$$\hat{P}(c) \quad = \quad \frac{\sum_{w \in c} \sum_v f_r(w, v)}{\sum_c \sum_{w \in c} \sum_v f_r(w, v)} \tag{2.7}$$

ここで, $f_r(w, v)$ は, 名詞 wと動詞要素 vが共起出現している頻度である. 未登録語は, その確率を基にして分類語彙表という分類シソーラスのノードに配置される. 徳永の方法では, 未登録語に幾つかの配置先を提示する.

そして, 中山の研究 [3] では, 浦本とは異なる方法で, 分類基準を共起データから抽出し, その基準を用いて, 未登録語を分類語彙表に配置する方法を提案した. 形式的には, 未登録語 n の動詞要素を V_nとし, ノードを Nとすると, 分類基準に基づくモデル $Sim(n, N)$ は 2.8式のようになる.

$$Sim(n, N) \quad = \quad \sum_{v \in V_n(N)} (a \cdot imp(v) + b \cdot \log_2 tr(v))) \tag{2.8}$$

$$imp(v) = -\log_2 \frac{|C_v|}{|C|} \tag{2.9}$$

$$tr(v) = \frac{|N_v|}{|N|} \tag{2.10}$$

ここで, a と b は定数である. N_vはノード Nに属す単語のうち, 動詞要素 vと共起する単語である. そして, Cは全ノードであり, C_vは vと共起するノードである. ノードと他ノードとの差別化をより強くするために, $imp(v) > 3.32$ を満たす動詞要素を用いて, $Sim(n, N)$ を計算する. $Sim(n, N)$ の値が大きいほど, n と Nが意味的に似ている. 中山の方法でも, 徳永の方法と同様に, 未登録語に幾つかの配置先を提示する.

2.2.3 共起データを用いた場合に生じる問題

この節では, 共起データを用いて単語間の類似度計算する場合に生じる問題について述べる. その問題とは, スパースネスの問題と多義語の問題である. それら問題がどのように配置先推定に悪い影響を与えるのかを述べる. そして, それら問題をどのように従来の方法で解消するのかを述べる.

スパースネスの問題 : コーパスから得られる共起データは不完全であることや, 動詞概念の表現方法は普通幾つもあるために, 未登録語の動詞要素と, 配置先の単語の動詞要素とが一致しない可能性がある.

多義語の問題 : 多くの単語[1]は, 複数の意味があるため, 未登録語の動詞要素が, 配置先と関係のない単語の動詞要素と一致する可能性がある.

まず, スパースネスの問題について考える. 例えば, 図 2.1のようなシソーラスに対して, 飛行機のノードに未登録語「ステルス」を配置する状況を考える. この状況において, 表 2.2のような共起データが得られたする. この例において, 未登録語「ステルス」は, 飛行機のノードに属したいのだか, 飛行機のノードに属している「ヘリコプター」や「イーグル」と同じ動詞要素を持っていないために, コサイン距離やタームマッチ法などの類似度では関連を示すことができない. タームマッチ法とは, 名詞間の動詞要素のマッチ数が多いものほど類似性が良いとする方法である. 従来の配置先推定の研究では, このような問題を以下のように扱った. 徳永の研究では, 単語と単語との類似度の計算ではなく, 単語と

[1]本研究では, 名詞の多義性を考慮せず, 動詞の多義性だけを考慮する.

単語集合との類似度を計算することによって, 共起データの不足分を補うようにした. しかし, 中山の研究で用いたような, 従来のスパースネスの解消方法と比べると弱い.

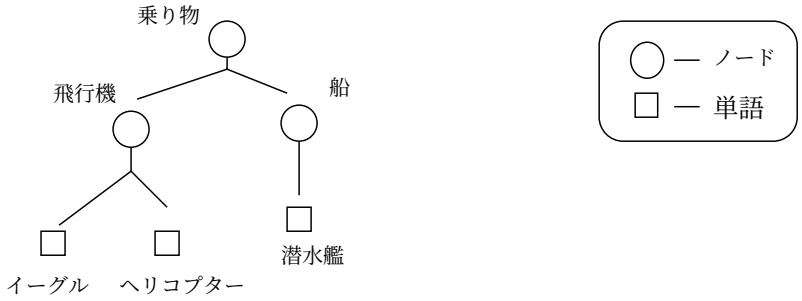


図 2.1: シソーラスの例 1

名詞 \ 動詞要素	を発見する	を探知する	で飛ぶ	を建造する
ヘリコプター	1	0	0	0
イーグル	1	0	1	0
潜水艦	1	1	0	1
ステルス	0	1	0	0

表 2.2: 共起データの例 1

その従来のスパースネスの解消方法とは, 以下のようなものである.

- シソーラスの構造を利用して, 意味的に似た動詞をまとめる.

例えば, 表 2.2の動詞「発見する」と「探知する」が, シソーラス上で類義語であることがわかれば, それら 2 つの動詞を 1 つの動詞と見ることによって, コサインやタームマッチ法でも, 未登録語は, ヘリコプターやイーグルと類似性が得ることができる. しかし, この方法の欠点は, まとめられた動詞要素が, 複数の意味をもつ語 (多義語) になる可能性があるために, 登録精度を下げる原因につながる.

本研究では, スパースネスの問題に対処するために, 統計的手法を用いて, 動詞をまとめるという操作を行う [5]. その統計的手法とは, 主成分分析である. シソーラスの構造では

なく, 主成分分析を用いる理由は, その分析の特性を利用することで, まとめれた動詞に対しても, 多義性を考慮できると考えれるためである.

次に, 多義語の問題について考える. 例えば, 図 2.2のようなシソーラスに対して, 未登録語「ステルス」を配置する状況を考える. この状況において, 表 2.3のような共起データが得られたする. 乗り物に関する対象領域の中では, 動詞「乗る」は多義語であり, 「乗り物にまたがって乗る」と「乗り物の中に入って乗る」という全く異なる意味をもつとすると, この例における「乗る」は, 1 つの意味で扱っているために, タームマッチ法では, 関連したくない単語「バイク」と関連性を示す. 従来の配置先推定の研究では, このような問題を考慮しないで, 未登録語の配置先を推定する.

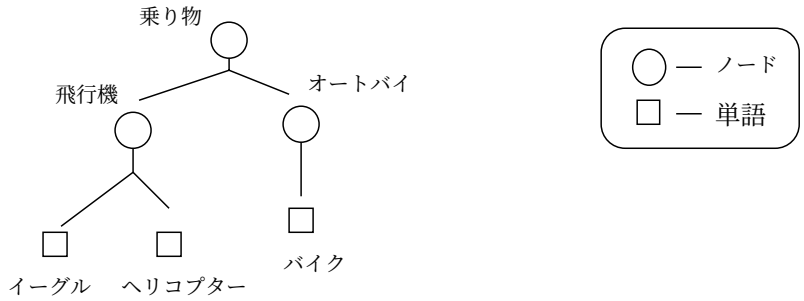


図 2.2: シソーラスの例 2

名詞 \ 動詞要素	に乗る	で飛ぶ	が走る
ヘリコプター	1	0	0
イーグル	1	1	0
バイク	1	0	1
ステルス	1	0	0

表 2.3: 共起データの例 2

多義語の問題に対処するために, 一般的に, 以下のような方法を用いる.

- 単語の語義曖昧性解消を行うことである. その解消方法の多くは, 辞書を利用して, 多義語を複数の意味に分けて, その意味の中から適切なものを選ぶという操作を行う.

例えば, 表2.3中の動詞「乗る」を辞書で調べると, 「乗り物にまたがって乗る」や「乗り物の中に入って乗る」という2つの意味がある. そして, 曖昧性解消のアルゴリズムは, それら意味の中から, 「乗る」に適切な意味を1つ決定する. この例では, バイクと共起する動詞「乗る」は, 「乗り物にまたがって乗る」という意味を選択し, 飛行機のノードに属す単語と共起する動詞「乗る」は, 「乗り物の中に入って乗る」という意味を選択する. その結果, タームマッチ法を用いた場合, 未登録語「ステルス」は, 飛行機のノードに属す単語とだけ関連を示し, バイクと関連しなくなる. 単語の語義曖昧性解消は, 自然言語処理の重要な問題であり, 多くの研究がある [6, 8, 7]. しかし, この方法は, 極めて高価であり, 有効な精度ではない. したがって, 多義語を扱うための有効な方法とはいえない.

本研究では, 従来の解消方法では不十分であるために, 統計的手法を用いて, 動詞を分けるという操作を行う [5]. その統計的手法とは, 主成分分析である. 3章で主成分分析の説明をする.

2.3 本研究の配置先推定の方法

この節では, 本研究が未登録語をどのようにシソーラスへ配置するのかを述べる. 本研究では, 共起情報を手掛かりとして, 未登録語の配置先を推定する. 本研究の特色は, 共起データを用いて類似度計算する場合に生じる2つの問題を解決するために, 共起データを主成分分析することである. その2つの問題に焦点を合わせるために, 従来の配置先推定の研究と異なる. 未登録語の配置先推定の手続きは, 共起データを主成分分析することを除けば, 徳永や中山の方法と同様の手続きを行なう. 未登録語の配置手続きは以下のようになる.

1. コーパスから得た共起データを主成分分析する.
2. 未登録語について, 動詞要素との共起データを取り出す.
3. 未登録語の共起データを (手順1で得た正規直交行列を用いて) 主成分分析する.
4. 未登録語とノードに属する単語集合との類似度をすべてのノードに対して計算する.
5. 類似度が最も大きいノードから順に未登録語の配置先候補とする.

シソーラスのあるノードと未登録語との類似度計算する場合, そのノードが類似度計算に使用する単語は, そのノードの下位にある単語すべてである. このように下位にある単語を用いる利点として, まず, シソーラス上では, 単語が直接属していないノードや, コーパスから共起データを得ることができないノードあるが, そのようなノードに対しても類似度の計算が可能になる. 他には, ノードに直接属す単語だけで類似度計算するよりも, スパースネスの問題が軽減できると考えられる.

第 3 章

共起データの主成分分析

本研究では, 主成分分析を用いて, 共起データを用いた場合に生じるスパースネスの問題と多義語の問題に対処方法にする. 本章では, まず, 主成分分析 [10] について説明する. その後, どのようにスパースネスの問題と, 多義語の問題を解消するのかを述べる.

3.1 主成分分析とは

p 個の変数に関する座標系 $x' = (x_1, x_2, \cdots, x_p)$ を正規直交行列 $L_{(p\times p)}$で回転して新しい座標系 $z' = (z_1, z_2, \cdots, z_p)$ を作る.

データ行列 $X_{(n\times p)}$が与えられたとき, 新しい座標系における値 $Z_{(n\times p)}$は, 次のようになる.

$$Z = XL \tag{3.1}$$

ここで, zについて次の条件をつける[1]

$$\left.\begin{array}{ll} \mathrm{V}[z_k] = \lambda & \\ \mathrm{Cov}[z_k, z_{k'}] = 0 & (k \neq k') \end{array}\right\} \tag{3.2}$$

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0 \tag{3.3}$$

[1] L には, p^2個の要素 (未知数) が含まれているが, 直交行列という条件から $LL' = L'L = I$(単位行列) となり, $p(p+1)/2$ 個の制約が与えれている (LL'は対称行列であることに注意).

こうすると zに関する分散・共分散行列 Λは 3.1式, 3.2式から次のように計算できる.

$$\Lambda = \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_p \end{pmatrix} = \frac{Z'Z}{n-1} = L'\frac{X'X}{n-1}L = L'VL \tag{3.4}$$

3.4式に左から L を掛け直交行列の性質 $LL' = I$を用いると

$$VL = L\Lambda \tag{3.5}$$

となり, 行列 Vの固有値問題に帰せられる. Λの対角要素が固有値, L の各列が固有ベクトルとなる.

通常, もとの変数の尺度依存性をなくすために, 各変数の分散を 1 になるように規準化して固有値・固有ベクトルを求める. この場合 Vは相関行列 Rに一致する.

$$RL = L\Lambda \tag{3.6}$$

3.5式または 3.6式を解いて固有値$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p (\geq 0)$, 対応する固有ベクトル $(l_1, l_2, \cdots, l_p)$ を求めると新しい変数は

$$z_k = l_{k'}x = l_{1k}x_1 + l_{2k}x2 + \cdots + l_{pk}x_p \tag{3.7}$$

となり, この z_kを第 k主成分と呼ぶ.

3.6式から求めたλについては

$$\lambda_1 + \lambda_2 + \cdots \lambda_p = p \tag{3.8}$$

という関係がある.

p はもとの変数の分散の和であり, 3.8式はこれが$\lambda_1, \lambda_2, \cdots, \lambda_p$に分解されていることを表している.

このためもとの変数全体に対する z_kの寄与率 c_kを次の式で定義する.

$$c_k = \frac{\lambda_k}{p} \tag{3.9}$$

z_1から z_mまでの累積寄与率は

$$\sum_{k=1}^{m} c_k = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_m}{p} \tag{3.10}$$

となる.

したがって最初の m 個の z_kを採用して累積寄与率が実用的に十分な程度の大きさであれば残りの $p-m$ 個の変数を分析の対象から外してもよいことがわかる.

主成分分析とはこのように次元を縮少してデータのもつ特徴の見通しをよくするための手法と言える.

3.2 主成分分析の利用

本研究では, 列を動詞要素とし, 行を名詞とする共起データ行列を主成分分析する. 名詞の種類を n とし, 動詞要素の種類を p とすると, 使用する主成分の数 m は, n 行 p 列のデータ行列 Xのランクとする. 主成分の数をランクとする理由は, もとのデータ行列 Xの情報量と同じであるからである [11]. 一般に, 主成分分析は, データの縮約化の手法として用いられるが, 本研究では, 主成分分析を利用する目的は, その分析の特性を利用することなので, もとの情報量を減らす必要はないと考える.

本研究では, 主成分分析の特性を通じて, 従来の解消方法と同様な, 動詞要素をまとめたり, 動詞要素をわけたり, という操作が可能なため, スパースネスの解消と多義性の解消ができると考える.

特性　主成分分析後の空間を張るすべての軸は, もとの軸 (動詞要素) すべての線形結合によって得られる.

以下の節で, スパースネスの問題と多義語の問題の対処方法について述べる.

3.2.1 スパースネスの問題への対応

スパースネスの問題に対して, 従来の解消方法は, シソーラスの構造を用いて動詞要素をまとめることである. 動詞要素をまとめることで, タームマッチ法などでは相関 (類似性) がなかった名詞間に相関があるように試みる. 本研究でも動詞要素をまとめるという操作を行うが, 従来の解消方法と異なる点は, 主成分分析と呼ばれる統計的手法を用いて, 数学的に動詞要素をまとめることである. その統計的手法を用いる理由は, 2.2.3節で述べている.

本研究では, 主成分分析後の新しい動詞 z_kは, 動詞要素 $x_1, x_2, \cdots, x_p$ の線形結合であるので, 動詞要素をまとめていると考える.

$$z_k = l_{1k}x_1 + l_{2k}x_2 + \cdots + l_{pk}x_p$$

そのような考えに基づくと, すべての動詞要素は同じ軸 z_k上で相関が得ることができるので, スパースネスの問題を解消しているとみなせる. 主成分分析によるまとめ方は, 従来の解消方法とは異なって, まとめたい動詞要素を選ばずに, すべての動詞要素をまとめていることに注意してほしい.

3.2.2 多義語の問題への対応

多義語の問題に対して, 従来の解消方法は, 2.2.3節で述べたように, 辞書を調べて, 多義語を複数の意味にわけることである. 本研究では, 多義語を複数の意味にわけるという操作を行うが, 従来の解消方法と異なる点は, 主成分分析と呼ばれる統計的手法を用いて, 数学的に動詞要素をわけることである. 統計的手法を用いる理由は, 語義の曖昧性解消アルゴリズムが, 適切に語義を選ぶことができないためである.

本研究では, 主成分分析することで, (多義性のある) 動詞要素 x_kを主成分分析後の新しい動詞 $z_1, z_2, \cdots, z_m$にわけていると考える. その考えの根拠は, 主成分分析後の新しい動詞 $z_1, z_2, \cdots, z_m$のすべてが動詞要素 x_kを含んでいるからである. 主成分分析は, まさに, 多義性のある動詞要素を辞書で調べると, $z_1, z_2, \cdots, z_m$ の意味があることと一致する.

$$\begin{aligned} z_1 &= l_{11}x_1 + l_{21}x_2 + \cdots + l_{p1}x_p \\ z_2 &= l_{12}x_1 + l_{22}x_2 + \cdots + l_{p2}x_p \\ &\cdot \\ &\cdot \\ z_m &= l_{1m}x_1 + l_{2m}x_2 + \cdots + l_{pm}x_p \end{aligned}$$

スパースネスの問題に対する一般的な方法では, まとめられた動詞要素に多義性が生じる可能性があったが, 主成分分析では, まとめた動詞要素 $x_1 + x_2 + \cdots + x_p$に対しても幾つかの新しい動詞 $z_1, z_2, \cdots, z_m$にわけているので, 多義性がないと考える.

第 4 章

未登録語の配置先を同定するモデル

4.1 区間推定のモデル

本研究では, 未登録語が, ノードの分布区間をどれだけ満たせるのかを調べることによって, 未登録語とノードの意味的な近さを計算するモデルを提案する. そのモデルは以下の仮定を基にして作成する.

本研究の仮定 : シソーラス上のノードに属す単語集合は, 意味的に似た集合である.

主成分分析の性質 : 意味的に似た単語集合は, 主成分軸上で偏った分布をもつ [4]

上記の 2 つのことが成り立つならば, ノードに属す単語集合は, 主成分軸上で偏った分布をもつということが成り立つ. もし主成分軸上で偏った分布を持った単語集合の中に未登録語があるならば, 未登録語は, その単語集合 (ノード) と意味的に似ていると考えられる. この考えを形式的にすると, 次のようになる.

ノード (単語集合) を nd とし, 未登録語を u とすると, 区間推定のモデル range は, 以下のようになる.

$$\begin{aligned} \mathrm{range}(nd, u) &= \sum_{j=1}^{n} \mathrm{get}_j \qquad (4.1) \\ \mathrm{get}_j &= \begin{cases} \mathrm{TR}_j(nd) & if(nd_j - 3 \cdot \mathrm{SD}_j(nd) \leq u_j \leq nd_j + 3 \cdot \mathrm{SD}_j(nd)) \\ 0 & \mathrm{otherwise} \end{cases} \end{aligned}$$

ここで, nd_jは nd の第 j軸の値であり, u_jは, 未登録語 u の第 j軸の値である.

$\mathrm{TR}_j(nd)$ は, 第 j 軸における nd の分布の偏りの度合を表す. $\mathrm{TR}_j(nd)$ の値が大きいほど, 偏った分布をもった軸 jである.

$$\mathrm{TR}_j(nd) = 1 - \frac{\mathrm{SD}_j(nd)}{\sum_{nd} \mathrm{SD}_j(nd)} \tag{4.2}$$

$$\mathrm{SD}_j(nd) = \sqrt{\frac{1}{|nd|} \sum_{w \in nd} (w_j - \bar{w_j})^2} \tag{4.3}$$

$$\bar{w_j} = \frac{\sum_{w \in nd} w_j}{|nd|} \tag{4.4}$$

ここで, n はベクトルの次元数, wは名詞, w_jは名詞の第 j軸の値, SD_jは第 j 軸の標準偏差である. $|nd|$ は, nd に属している単語集合の数である. 類似度 $\mathrm{range}(nd, u)$ は, 軸上の nd の区間に未登録語 u が含まれるほど, nd と意味的に似ている度合が高くなると考えられるので, get_jの合計が大きいほど, 未登録語 u の適切な配置先のノード nd となる. また, TR の値は, ある軸上で, ノードの単語集合の分布が大きくなればなるほど, どのような未登録語でもそのノードの区間に含まれやすいので, 値は小さくなるようにしてある.

第 5 章

評価実験

本章では, 3.2節で述べた問題の対策方法が, 未登録語の分類にどのように影響するのかを調べる. そのために, 主成分分析前と主成分分析後の類似度の比較をスパースの問題と多義語の問題ごとに比較する. また, 未登録語の配置先推定に最も適した方法はなにかを調べるために, 関連研究が用いた類似度を実装し比較する.

本章は, 5.1節で実験に用いるシソーラスや, 実験方法, 評価方法などの実験環境について説明する. そして, 5.2節で, 実験の結果とそれの考察を行う.

5.1 実験環境

5.1.1 シソーラス

本研究では, シソーラスとして, EDR 概念体系辞書 [16] を用いる. 概念体系辞書は, 約 40 万の概念を上位下位の関係で階層的に構造化してある. その構造は, 若干の多重継承があるために, 完全な木ではなくグラフである. 単語は, 40 万概念のどこかに属する.

概念体系辞書は, 概念体系レコードの集合である. 概念体系レコードは, レコード番号, 上位概念識別子, 下位概念識別子, および管理情報からなる. 各フィールドの構造および内容は, 次のとおりである.

＜レコード番号＞ : レコードタイプ (CPC) と識別番号
＜上位概念識別子＞ : 上位概念を表す 16 進整数の概念識別子
＜下位概念識別子＞ : 下位概念を表す 16 進整数の概念識別子

＜管理情報＞　　：更新日付などの管理履歴

概念体系レコードの例を以下に示す.

＜レコード番号＞　　CPC0376089
＜上位概念識別子＞　　0f5354 ：「乗り物」を表す概念識別子
＜下位概念識別子＞　　30f710 ：「飛行機」を表す概念識別子
＜管理情報＞　　DATA="95/6/7"

このような概念レコードを集めることによって, 図 5.1のような体系を作れる.

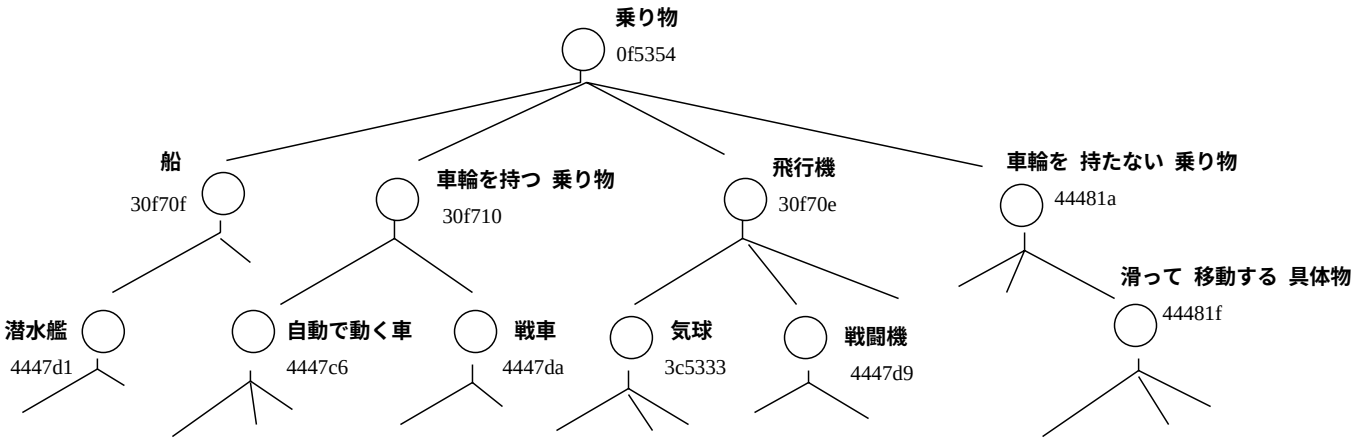


図 5.1: 乗り物の階層

この図は, 「乗り物」という概念 (ノード) の下位分類である. この概念は, 「車輪を持つ乗り物」や, 「船」, 「飛行機」,「車輪を持たない乗り物」といういくつかの概念に分割され, 詳細化されている. さらに, 「飛行機」という概念は, 「気球」や「戦闘機」などの下位概念に詳細化されている. そして, これら概念のどこかに単語が属している.

実験では, 概念体系辞書のうち最上位概念を「**乗り物**」[1]とする部分木を用いる.

[1]概念識別子は 0f5354 である

5.1.2 共起データ

共起辞書

本研究で使用する共起データは, EDR 日本語共起辞書 [18] から獲得する. その共起辞書は, 日本語コーパス [17] の中から係り受け構造を取り出したものである. 日本語共起辞書のレコードは, レコード番号, 見出し情報, 共起句構成要素情報, 構文情報, 意味情報, 共起状況情報, および管理情報から構成される. 各フィールドの構造および内容は, 次のとおりである.

＜レコード番号＞ : レコードタイプ (JCC) と識別番号
＜見出し情報＞ : 共起辞書レコードの見出し
　＜句見出し＞ : 共起句の表記
＜共起構成要素情報＞ : 共起句を構成する形態素列に関する情報
　＜構成要素列＞ : 共起句を構成する個々の形態素
＜構文情報＞ : 共起句の構文構造を示す情報
＜意味情報＞ : 深層の概念関係を示す構文木
　＜部分意味フレーム＞ : 概念関係を示す意味フレーム情報
＜共起状況情報＞ : 日本語コーパスにおける共起状況を示す情報
　＜頻度＞ : 日本語コーパスにおける出現頻度
　＜例文＞ : 共起関係に基づいた文
＜管理情報＞ : 更新日付などの管理履歴

頻度の項目においては, さらに 4 つに分けられている.

＜表層共起頻度＞ : 表層の共起関係の出現頻度
＜共起項目頻度＞ : 深層の概念関係を含む出現頻度
＜受け側形態素頻度＞ : 受け側形態素の頻度
＜係り側形態素頻度＞ : 係り側形態素の頻度

共起辞書レコードの例を以下に示す.

＜レコード番号＞ JCC0227124
＜見出し情報＞

＜句見出し＞　　気球　を　飛ば

＜共起構成要素情報＞

＜要素番号＞＜形態素＞＜かな表記＞＜品詞＞＜概念識別子＞

{ 1　　気球　　キキュウ　　名詞　　3bdb88　　}

{ 2　　を　　ヲ　　助詞　　""　　}

{ 3　　飛ば　　トバ　　動詞　　100880　　}

＜構文情報＞

＜受け側要素＞　　3/飛ば

＜関係要素＞　　2/を/を

＜係り側要素＞　　1/気球

＜意味情報＞

＜受け側概念要素＞　　3/100880/飛ば

＜概念関係子＞　　object

＜係り側概念要素＞　　1/3bdb88/気球

＜共起状況情報＞

＜頻度＞　　1;1;34;20

＜例文＞　　{"＜気球＞を（飛ば）した"}

＜管理情報＞　　DATA="95/6/16"

共起データの形式

この節では, 本研究が使用した共起データの形式とその形式の獲得方法について述べる. 共起データは, 以下のような手順で獲得した.

1. 共起辞書レコードにおいて, 係り側要素の品詞が名詞で, 受け側要素の品詞が動詞で, 関係要素が「を」,「が」, 「で」, 「に」であるようなレコードに対して, 次のような形式で獲得する.

共起データの形式

(係り側要素の概念識別子, 関係要素, 受け側要素, 共起項目頻度)

2. 獲得した共起データの (係り側要素の概念識別子, 関係要素, 受け側要素) について合併し, 共起項目頻度を足し併せる.

以降, 本論文では, 係り側要素の概念識別子を**名詞**, 関係要素と受け側要素の組を**動詞要素**, 共起項目頻度を**頻度**と呼ぶ.

実験では, EDR 共起辞書から「乗り物」に関する名詞の共起データを獲得した. 獲得された名詞は 262 種類であり, 動詞要素 (格と動詞の組) は 1039 種類である. 表 5.1に, その共起データの大きさを格助詞ごとに示す. この表は, 全部で 1,951 レコードあるうち格助詞が「を」であるものは, 516 レコードあり, そのレコード内にある名詞は 136 種類で, 動詞は 277 種類あることを示す. 獲得された 262 種類の名詞を用いて, 最上位概念まで辿ると, 階層の平均の深さは, 約 3.2 であり, 中間ノードの数は 32 種類である. 中間ノードは, 葉ノードを除いたノードのことを意味する.

格助詞	レコード数	名詞の種類	動詞の種類
を	516	136	277
に	438	135	217
が	695	167	377
で	302	91	168
合計	1,951	262	799

表 5.1: 格助詞ごとの共起レコードの大きさ

5.1.3 実験方法

実験方法は, 徳永の研究でも用いられたように, クロスバリデーション (10-fold croos validation) を用いる [9]. その方法は, 262 種類の名詞をランダムに 10 個のグループにわけて, 1 グループを実験データ (未登録語) とし, 残りの 9 グループを訓練データ (シソーラスを構成する名詞) として用いる. そして, その実験データを 10 回繰り返す. そのため, この方法では, ほとんどの名詞を実験データとして使うことができる. しかし, この方法を用いた場合, 実験データとして使えない名詞がいくつかある. それは, 未登録語の動詞

要素すべてが, 訓練データの中にない場合である. 従って, そのような場合を取り除くと, 実験データとして用いる名詞は 235 種類になる.

5.1.4 評価方法

評価方法は, 実験データの各グループに対して, 以下のような配置手続きを行なった後, 各グループの正解数を合計し, 実験データ数で割ったときの精度で評価する.

1. 各未登録語に対して, 全てのノードとの類似度を計算する.
2. 類似度の良い順に, ノードをソートする.
3. 第 N 位までのノードを取り出す ($N = 1, 5, 10$). 同じ類似度がある場合は, 一番最後の順位をを与える. 例えば, 最も類似度が良いものが 11 個あれば, それらの順位はすべて 11 位とする.
4. その中に正解が入っているかを調べる. 正解があれば, その数を数える.

5.1.5 実験データ

本研究の実験では, 実験データ (未登録語) として, 多義語の問題とスパースネスの問題をそれぞれ用意する必要がある. 多義語の問題とスパースの問題は, それぞれ以下のような場合とする.

多義語の問題 : 実験データの動詞要素を EDR 共起辞書で調べて, 概念関係子と動詞の概念識別子のペアが, 乗り物の対象領域の中で, 複数の意味があるかを調べ, 2 つ以上の意味がその対象領域にあれば, 多義語の問題とする.

スパースネスの問題 : 予め用意された正解ノードと実験データとのコサイン距離を計算し, コサインの値が 0 になるものを実験データとする.

表 5.2に, 235 個の実験データにおける多義語の問題 (AM) とスパースネスの問題 (SP) を内訳を示す. 丸 (○) は, 問題である場合で, 罰 (×) は, 問題でないことを示す.

		S	P
		○	×
A	○	62	151
M	×	15	7

表 5.2: 実験データにおける問題の内訳

表 5.2は, 多義語の問題且つ, スパースの問題である実験データが 62 個あることがわかる. 実験で用いた多義語の問題数は 213 題であり, スパースネスの問題数は 77 題である. 未登録語の多くは, EDR 辞書上で, 多義性のある動詞要素をもつことがわかる. また, 多義語の問題 (213 題) と, スパースの問題 (77 題), 全問題 (235 題) それぞれに対して, どれくらいの頻度をもった未登録語を配置するのか表 5.3と, 表 5.4, 表 5.5に示す. 例えば, 表 5.3では, 未登録語と共起する動詞要素の種類が 1 ～ 5 で, その動詞要素の頻度の合計が 1 ～ 5 のものが, 213 問中, 152 題あることを示す.

頻度 \ 種類	1 ～ 5	6 ～ 10	11 ～ 15	16 ～ 20	21 ～	合計
1 ～ 5	152	—–	—–	—–	—–	152
6 ～ 10	6	21	—–	—–	—–	27
11 ～ 15	1	7	7	—–	—–	15
16 ～ 20	0	1	2	4	—–	7
21 ～	0	0	1	1	10	12
合計	159	29	10	5	10	213

表 5.3: 未登録語の頻度と動詞要素の種類ごとの分類 (多義語の問題)

頻度 \ 種類	1 ∼ 5	6 ∼ 10	11 ∼ 15	16 ∼ 20	21 ∼	合計
1 ∼ 5	69	—	—	—	—	69
6 ∼ 10	0	5	—	—	—	5
11 ∼ 15	0	1	0	—	—	1
16 ∼ 20	0	0	0	1	—	1
21 ∼	0	0	0	0	1	1
合計	69	6	0	1	1	77

表 5.4: 未登録語の頻度と動詞要素の種類ごとの分類 (スパースネスの問題)

頻度 \ 種類	1 ∼ 5	6 ∼ 10	11 ∼ 15	16 ∼ 20	21 ∼	合計
1 ∼ 5	174	—	—	—	—	174
6 ∼ 10	6	21	—	—	—	27
11 ∼ 15	1	7	7	—	—	15
16 ∼ 20	0	1	2	4	—	7
21 ∼	0	0	1	1	10	12
合計	181	29	10	5	10	235

表 5.5: 未登録語の頻度と動詞要素の種類ごとの分類 (全問題)

実験データの正解ノードは, 次のようにする.

- 実験データが, もともと概念体系上で, 中間ノードに存在していたならば, その中間ノードを正解とする.
- 未登録語が, もともと葉ノードであったならば, それの親ノードを正解とする. ただし, 親ノードが 2 つ以上ある場合は, それら親のどれか 1 つを選べば正解とする.

5.1.6 類似度

評価に用いる距離尺度び, 未登録語の配置先を同定するモデルを以下に示す.

1. cosine ： コサイン距離 (2.1式)

2. distance ： ユークリッド距離

$$\text{distance}(X, Y) \quad = \quad \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \tag{5.1}$$

3. range ： 提案方法 (4.1式)

4. 徳永 ： 徳永の方法 (2.3式)

5. 中山 ： 中山の方法 (2.8式)

6. baseline ： ランダムに単語を配置するモデル. このモデルは, 問題の難しさを理解するために用意する.

単語と単語集合との類似度計算をする場合, コサイン距離や, ユークリッド距離, 提案方法 range は重心法を用いる.

5.2 結果および考察

本節では, 各実験の結果を示す. 表 5.6と, 表 5.8, 表 5.10の第 1 列目は類似度名であり, 第 2 ～ 4 列目は, 5.1.3節で述べたように, 第 N 位まで配置先の候補を認めた場合である. 各類似度の入れ子の各行は, 第 1 行目は正解数を示し, 第 2 行目は正解率を示す. そして第 3 行目がある場合は, 主成分分析前と後の類似度での改善率を示す.

また, スパースネスの問題と多義語の問題における主成分分析の効果を調べるために, cosine と, distance, range を比較する. 徳永の確率モデルや中山の方法は, 主成分分析後のデータが正と負の値をもつために, 実装できない.

5.2.1 スパースネスの問題への効果について

スパースの問題において, 表 5.6のような結果が得られた.

類似度		正解の順位 (最低 32 位)		
		1 位	5 位	10 位
baseline		2/77 (2.6)	5/77 (6.5)	24/77 (31.2)
中山		0/77 (0.0)	0/77 (0.0)	0/77 (0.0)
徳永		0/77 (0.0)	0/77 (0.0)	0/77 (0.0)
主成分分析前	**cosine**	0/77 (0.0)	0/77 (0.0)	0/77 (0.0)
	distance	5/77 (6.5)	20/77 (26.0)	32/77 (41.6)
	range	0/77 (0.0)	0/77 (0.0)	0/77 (0.0)
主成分分析後	**cosine**	0/77 (0.0) ±0.0	0/77 (0.0) ±0.0	0/77 (0.0) ±0.0
	distance	5/77 (6.5) ±0.0	21/77 (27.3) +1.2	32/77 (41.6) ±0.0
	range	3/77 (3.9) +3.9	13/77 (16.9) +16.9	25/77 (32.5) +32.5

表 5.6: スパースネスの問題 (77 題) における登録精度

主成分分析の影響

主成分分析によって, スパースネスの問題が改善できたかを調べる. 主成分分析前と後の欄の類似度を比較する. その結果, cosine や, distance には影響がほとんどない. しかし, range においては, 順位が 1 位の欄で 3.9%, 5 位の欄で 16.9%, 10 位の欄で 32.5%の改善が見られる.

実験の結果, range のような類似度は, 主成分分析した方がスパースネスの問題を改善できる. 主成分分析前の range において効果がない理由は, 未登録語と同じ動詞要素をもたないノードが複数 (10 個以上) あるため, 正解の順位が悪くなる. 一方, cosine や distance のような類似度では, 主成分分析を行っても影響がない. cosine や distance では, 効果がない原因として, 以下のことが挙げられる.

- 主成分分析は, 空間を張る軸を変化させるが, 対象 (名詞) の空間配置は変化しないため, distance では影響がない.
- 主成分分析後の原点が, 分析前とほとんど変化がないためと, 上記の原因のため, cosine には影響が生じない. (主成分分析後の各原点の平均 : 0.00985)

これら原因から, 主成分分析を利用してスパースネスの問題を改善しようとするなら, range のようなに, ある特定の分布区間 (空間) の中だけで類似度を計算することで効果があるため, distance や range においても同様に特定の空間の中で類似度を計算すれば, 影響がみられる可能性がある.

類似度の比較

スパースの問題において, 最も効果がある類似度は, distance であり, 正解の順位が 1 位の場合の精度は 6.5%であり, 第 5 位の場合は 27.3%で, 第 10 位は 41.6%である. 次に, range であり, 正解の順位が第 1 位の場合は 3.9%であり, 第 5 位の場合は 16.9%, 第 10 位は 32.5%である.

表 5.7は, 正解順位が 10 位の欄における distance と range の正解 (○) と不正解 (×) の内訳を示す.

		range	
		○	×
distance	○	18	14
	×	5	40

表 5.7: 第 10 位における distance と range の正解と不正解の内訳

distance が正解し range が不正解である 14 例の失敗の分類を以下に示す. 各分類の失敗数は, 延べ数である.

1. 上位ノードの方が軸の分布区間が広いために, 正解よりも上位を選んだ失敗 : 6 例 (正解順位が 1 位の欄の結果)
2. 主成分軸上の正解と失敗の各区間の中に, 未登録語が属す数は同じであったが, 軸の重みによって, 適切に選べなかった失敗 : 2 例
3. 未登録語の頻度が, 正解の動詞要素の頻度より多いために, 幾つか主成分軸の区間に入り損う失敗 : 3 例
4. 同じ類似度が複数あるために失敗 : 2 例
5. 未分類 : 2 例 (正解と未登録語の間で動詞要素の一致がないため)

この分類において, 1 番のような失敗例は, 正解ノードとルートの間にあるので, 正解に近いと考えれる. 軸の分布区間は, 一般的にノードに属す単語集合が大きくなるほど, 広くなる. そのために, 上位ノードを選択する傾向が, range の類似度にはある.

5.2.2 多義語の問題への効果について

多義語の問題において, 表 5.8のような結果が得られた.

主成分分析の影響

主成分分析によって, 多義語の問題が改善できたかを調べる. 主成分分析前と後の欄の類似度を比較する. その結果, cosine や, distance には影響がほとんどない. しかし, range

類似度		正解の順位 (最低 32 位)		
		1 位	5 位	10 位
baseline		6/213 (3.0)	33/213 (15.5)	66/213 (31.0)
中山		35/213 (16.4)	105/213 (49.3)	136/213 (63.8)
徳永		29/213 (13.6)	108/213 (50.7)	141/213 (66.2)
主成分分析前	**cosine**	37/213 (17.4)	110/213 (51.6)	138/213 (64.8)
	distance	51/213 (23.9)	114/213 (53.5)	133/213 (62.4)
	range	21/213 (9.9)	91/213 (42.7)	127/213 (59.6)
主成分分析後	**cosine**	37/213 (17.4) ±0.0	110/213 (51.6) ±0.0	137/213 (64.3) -0.5
	distance	51/213 (23.9) ±0.0	115/213 (54.0) +0.5	133/213 (62.4) ±0.0
	range	27/213 (12.7) +2.8	105/213 (49.3) +6.5	157/213 (73.7) +9.4

表 5.8: 多義語の問題 (213 題) における登録精度

においては, 順位が 1 位の欄で 2.8%, 5 位の欄で 6.5%, 10 位の欄で 9.4%の改善が見られる.

cosine や, distance に影響がないのは, 5.2.1節と同様である.

類似度の比較

多義語の問題において, 正解順位が 1 位の欄と 5 位の欄でもっとも良かったのは, distance であり, 精度は 1 位の欄で 23.9%, 5 位の欄で 54.0%である. 一方, その欄において, range の精度は, 1 位の欄で 12.7%, 5 位の欄で 49.3%と劣っている. 以下の表 5.9は, 正解順位が 1 位の欄における distance と range の正解 (○) と失敗 (×) の内訳を示す.

		range	
		○	×
distance	○	12	39
	×	15	147

表 5.9: 第 1 位における distance と range の正解と不正解の内訳

39 例において, 失敗の分類として,

- 上位ノードの方が軸の分布区間が広いために, 正解よりも上位を選んだ失敗
- 未登録語の動詞要素が 1 つで, その要素をもつノードが複数あるために, 正解のノードを適切に選べなかった失敗
- 主成分軸上の正解と失敗の各区間の中に, 未登録語が属す数は同じであったが, 軸の重みによって, 適切に選べなかった失敗
- 未登録語の頻度が, 正解の動詞要素の頻度より多いために, 幾つか主成分軸の区間に入り損う失敗

正解順位を 10 位まで緩めると, 最も良い類似度は, range であり, その精度は 73.7%である. 次によい類似度は, 徳永であり, 精度は 66.2%である. distance は, 10 位の欄では各類似度の中で 5 番目に位置し精度は 62.4%である.

5.2.3 未登録語の配置先推定に適した方法について

スパースネスと多義語の問題を含んだ場合の精度を表 5.10しめす.

類似度の比較

全 235 題において, 正解の順位が 1 位の欄で最も良かったのは, distance の 23.8%である. 次に, cosine の 18.3%である. 5 位の欄で最も良かったのは, distance の 53.6%であり, 次に, cosine の 49.8%である. 10 位の欄で最も良かったのは, range の 71.1%である. 次に徳永と distance の 63.0%である.

正解順位が 1 位の欄において, 最も未登録語に適した類似度は, distance であると考えられる. スパースの問題と多義語の問題でも最も適した類似度である. しかし distance は, 従来の研究は使われていない.

正解順位が 10 位の欄において, 最も適した類似度は, 主成分分析と range の組み合わせである. 従来の研究では, 未登録語の配置先を幾つか提示する方法をとる. 本研究もその方法に従うならば, 正解順位が 5 位の欄において, distance を除けば, ほとんど同等な精度であるが, 10 位の欄では, 従来の精度より約 10%程度の改善があるため, 最も適した類似度と考えれる.

区間推定のモデル range は, 単語集合が多いノードを未登録語の配置先とする傾向がある. その傾向があるため, 正解順位の 1 位の精度があまりよくない. その傾向が生じる理由は, 単語集合が多いノードほど, ノードの分布区間が広くなるためである. このような失敗に対して, 各ノードの区間を一定にする対策が必要となる. 他の対策方法として, ノードの分布区間を計算する際, ノードは子孫の単語ではなく, ノードに直接属す単語だけを利用する. このようにすると, 上位ノードの分布区間が広くならないと考えれるが, 2.3節で述べたように, すべてのノードが未登録語の配置先の候補にならない可能性がある.

類似度の計算において, 浦本の研究のように, ノードは子孫の単語ではなく, ノードに直接属す単語を利用して類似度の計算をした実験結果 (表 5.11) がある. また, ノードに複数の単語が属す場合は, 最近隣法を用いて類似度の計算を行なった. range の分布区間と TR の値は, 浦本の分類規準抽出のように, ノードの子孫の単語集合を利用して求めた. この実験では, 未登録語の配置先の候補は全部で 29 個である. cosine や range では, 全体的に精度が向上している. distance において精度が下がっている理由は, 最近隣法を用いたことで, 同点の類似度が多く存在するためである.

類似度		正解の順位 (最低 32 位)		
		1 位	5 位	10 位
baseline		7/235 (3.0)	37/235 (15.7)	75/235 (31.9)
中山		40/235 (17.0)	112/235 (47.7)	143/235 (60.9)
徳永		33/235 (14.0)	115/235 (48.9)	148/235 (63.0)
主成分分析前	**cosine**	43/235 (18.3)	117/235 (49.8)	145/235 (61.7)
	distance	56/235 (23.8)	126/235 (53.6)	148/235 (63.0)
	range	26/235 (11.1)	98/235 (41.7)	134/235 (57.0)
主成分分析後	**cosine**	42/235 (17.9)	116/235 (49.4)	143/235 (60.9)
	distance	56/235 (23.8)	127/235 (54.0)	148/235 (63.0)
	range	29/235 (12.3)	113/235 (48.1)	167/235 (71.1)

表 5.10: 全問題 (235 題) における登録精度 1

類似度		正解の順位 (最低 29 位)		
		1 位	5 位	10 位
主成分分析前	**cosine**	71/235 (30.2)	144/235 (61.3)	156/235 (66.4)
	distance	47/235 (20.0)	79/235 (33.6)	88/235 (37.4)
	range	47/235 (20.0)	120/235 (51.1)	141/235 (60.0)
主成分分析後	**cosine**	68/235 (28.9)	145/235 (61.7)	161/235 (68.5)
	distance	41/235 (17.4)	71/235 (30.2)	94/235 (40.0)
	range	65/235 (27.7)	153/235 (65.1)	185/235 (78.7)

表 5.11: 全問題 (235 題) における登録精度 2

第 6 章

まとめ

本研究では, 共起データを用いて単語間の類似度を計算する場合に生じるスパースネスの問題と多義語の問題に対処する方法を提案した. その方法とは, 主成分分析を通じて, 動詞要素をまとめたり, わけたりするものである. また, 本研究の仮定と主成分分析の性質に基づいた区間推定のモデル range を提案した. 本研究では, 以下のことが明らかになった.

- 主成分分析は, range の類似度を用いた場合, スパースネスの問題には最大で 32.5%, 多義語の問題には最大で 9.4%の改善があった.
- しかし, cosine や distance の類似度の場合には, 主成分分析の影響がない. その理由として, 全ての情報量を用いたためと, そして, 主成分分析は, 空間を張る軸を変化させるが, 対象 (名詞) の空間配置は変化しないことや, 主成分分析後の原点が分析前とほとんど変化がないためである (重心の平均 : 0.00985).
- 正解順位が 1 位の欄において, 最も貢献した類似度は, distance であった. しかし, 従来の研究では, 未登録語の配置先を幾つか提示する方法をとる. 本研究もその方法に従うならば, 10 位の欄で, 従来の精度より約 10%程度の改善があるため, 最も適した類似度と考えれる.

今後の課題として以下のことを挙げる.

- 区間推定のモデル range は, 単語集合が多いノードを未登録語の配置先とする傾向がある. その傾向が生じる理由は, 単語集合が多いノードほど, ノードの分布区間が広くなるためである. このような失敗に対して, 各ノードの区間を一定にするなどの対策が必要となる.

- 類似度 distance は, cosine や, 徳永の確率モデルではデータスパースネスであっても, distance においては, スパースネスでないという現象がある. スパースネスの問題では distance の 41.6%が最もよく, 多義語の問題では, 主成分分析後の range を除くと, 徳永の 66.2%である. この distance と確率モデルを組み合わすことで, 精度が向上する可能性があるので, 調査する必要がある.
- 最後に, 今回の実験では, シソーラスのサイズが小さいので, 大規模な実験を行う.

謝辞

本研究を進めるにあたり, 島津明教授ならびに奥村学助教授には数多くの御教示を頂きました. また, 本研究に関して多大な助言をしていただいた本田岳夫氏に心から感謝いたします. Thanaruk Threeramunkong 先生には, 終始あたたかいご助言を頂きました. そして, 島津・奥村研究室の皆様がたには研究に関する貴重な支援をして頂きましたことを心から感謝致します.

参考文献

[1] 浦本直彦, 既存のシソーラスとコーパスから抽出した知識の統合, 自然言語処理シンポジウム「大規模資源と自然言語処理」, pp1-9, 1996.

[2] Tokunaga Takenobu, Fujii Atsushi, Iwayama Makoto, Sakurai Naoyuki, Tanaka Hozumi, Extending a thesaurus by classifying words, Proceedings of the ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources, pp16-21, 1997, Jul.

[3] 中山拓也, 松本裕治シソーラスへの未登録語の自動登録自然言語処理 120-16, pp103-108, 1997.

[4] 小島秀樹, 伊藤昭, 辞書にもとづいて語彙をクラスタリングする試み, 言語処理学会第１回年次大会, pp205-208, 1995.

[5] Scott Deerwester, Susan T. Dumais, George W.Furnas,Thomas K. Landauer, and Richard Harshman, Indexing by Latent Semantic Analysis, JOURNAL OF THE AMERICAN SOCIETY FOR INFOMATION SCIENCE, pp391-407, 1990.

[6] D.Yarowsky. Word-sence disambiguation using statistical models of roget's categories trained on large corpora. In Proceedings of COLING'92, pp454-460, 1992.

[7] Philip Resnik Selectional Preference and Sense Disambiguation Tagging Text with Lexical Semantics : why, what, and How? Proceedings of the Workshop, pp52-57, 1997.

[8] Fumiyo Fukumoto, Yoshimi Suzuki, An Automatic Clustering of Articles Using Dictionary Definitions, In Proceedings of COLING'96, pp406-411, 1996.

[9] Sholom M. Weiss and Casimir Kulikowski, Computer systems that learn: classification and prediction methods from statistics, neural nets, machine larning, and expert systems, Morgan Kaufmann, 1991.

[10] 芳賀敏郎, 橋本茂司, 回帰分析と主成分分析, 日科技連出版社, 1980.

[11] 岡太彬訓, 今泉忠, パソコン多次元尺度構成法, 共立出版, 1994.

[12] Chapman, L. R. Roget's International Thesaurus. Thomas Y. Crowell Company Inc., 1977.

[13] 国立国語研究所, 分類語彙表, 秀英出版, 1964.

[14] 田中 and 仁科, 上位/下位関係シソーラス ISAMAP の作成, 情報処理学会自然言語処理研究会, 64-4, pp25-44, 1987.

[15] A. Miller, R.Beckwith, C.Fellbaum, D.Gros, K.Miller, and R.Tengi, Five papers on wordnet, Technical Report CSL Report 43, Cognitive Science Laboratory, Princeton University, 1990.

[16] 日本電子化辞書研究所, EDR 概念体系辞書 1.5 版, 1996.

[17] 日本電子化辞書研究所, EDR 日本語コーパス

[18] 日本電子化辞書研究所, EDR 日本語共起辞書 1.5 版, 1996.