修 士 論 文

# 電子化辞書を利用した,
# 概念に基づくクエリーの拡張に関する研究

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

太田千晶

平成10年 2月 13日

## 要 旨

情報検索に関する研究としてクエリー拡張の研究がある. 情報検索は 計算機の実用化が始まった頃から 重要な課題として様々な研究がなされてきた. これまでの研究において, 既存データベースを用いたクエリー拡張の研究は, 人手で 各タームの語義や ターム間 もしくは 語義間の意味的な関連性を付された信頼性のあるデータベースに基づくクエリー拡張を行えるという利点を持つが, これまではそのデータベースを充分に利用したクエリー拡張の研究は行われていない.

そこで, 本研究では データベースとして EDR 電子化辞書中に記述されている様々な “概念” および “概念関係” に基づいたクエリー拡張手法の提案を行い, 各手法による関連語そのものの評価, そしてそれらの関連語より生成される拡張クエリーの検索精度における有効性の検証を行う.

# 目 次

# 図 目 次

# 表 目 次

# 第 1 章

# 序論

本論文では, 情報検索における クエリー (検索質問) の拡張に関する研究について述べる.

クエリー拡張とは, 入力されたクエリーから 検索に有効な拡張クエリーを生成することによって, 検索精度の向上 または ユーザの欲する情報までの効果的な誘導等を実現しようとするものである.

本研究では, 従来の研究で使用されてきた「表記」だけではなく, そのクエリーが持つ意味的な情報である “概念” に着目し, 様々な概念関係・概念記述による拡張を行うことによって関連性の高いクエリーを自動的に生成することを 1 つの目的とする. そしてさらに, 提案手法によって得られた拡張クエリーを用いることによる検索精度の向上を目指す. 具体的には, 本研究では EDR 電子化辞書中の概念記述を用い [1], ‘概念” を利用することによって実現可能となる複数クエリー拡張の手法を提案・実装し, 評価実験によって その効果の検証を行う.

まず本章では, はじめに 情報検索におけるクエリー拡張の有用性について述べ, 次に, 従来の研究を通して本研究の目的・その位置付けを明確にする.

## 1.1 情報検索とクエリー拡張

計算機による情報検索技術は, 特許事例の検索や図書の検索など計算機の実用化が始まった初期の段階から重要な課題の 1 つとして注目を浴び, これまで様々な研究がなされて来た [2][3]. しかし, 近年 インターネットの普及により新聞, 手紙 から 個人情報に至るまであらゆる情報の電子化が急速に進み, 膨大な数の情報の中から必要な情報をいかに効

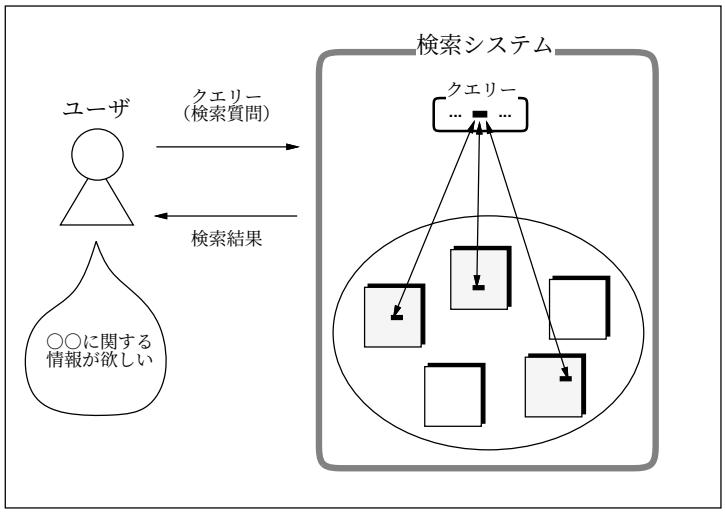



図 1.1: ユーザと検索システム

率よく, かつ正確に発見するかという情報検索の技術の向上が強く望まれるようになってきた. また, ここ数年で 情報機器が一般家庭に急速に普及したことにより, 利用者が電子情報にアクセスする機会も増加している. このような検索に不慣れな一般ユーザの出現によって, 情報検索においては 検索本来の難しさがあらためて認識されるところとなり, 検索技術に並んで「検索支援」技術の必要性も同時に高まってきていると言える.

情報検索の過程において, 図 1.1のように, ユーザは 検索システムに対しクエリー (検索質問) を入力する. このクエリーはユーザの欲する情報についての要求を単語またはフレーズで表したものであるが, 全てのユーザが自分の欲する要求を検索に適切な形で表現できるとは限らない. また, 図 1.1(右側) のように, 一般の検索システムでは 基本的にクエリーに含まれるタームと検索文書中に出現するタームとのキーワードマッチングによって正解文書を導き出しているため, 検索精度がユーザのクエリーに関する知識レベルに依存してしまうという問題が生じる.

そこでこのような問題に対処する一手法として, クエリー拡張が有効であると考える. クエリー拡張は, ユーザがシステムに入力したクエリーを基に, 検索に有効なクエリーを生成し, それによって検索精度の向上を導くことを目的とする. 特に, このクエリー拡張を行うことによって, 検索文書集合内に潜在するクエリーに関連する情報をより多く発見する効果を得ることができる.

## 1.2 従来の研究

クエリー拡張の研究には, 大きく分けると, 特にユーザとシステムとのインタラクションに注目したユーザインターフェースに関する研究と, ユーザが入力したクエリーから自動的に検索に有効な拡張タームを生成する点に注目した研究がある. 前者は Relavance Feedback を用いた研究が盛んである [4][5].

一方 後者に関しては, さらに利用するデータを基準として, 表層的な情報に基づく研究と深層的な情報に基づく方法の 2 種に分けることができる.

前者の表層的な情報に基づく研究としては, まず 西村らによる Association Rule を用いた関連語の発見手法が挙げられる [6]. この研究では, 検索文書中に共起するタームの中から特に関連性のあるタームの関係を発見し, それを Association Rule として構築して拡張に用いるものである.

また一方で, Qiu&Frei による研究がある [7]. この研究では, idf(文書におけるタームの出現頻度の逆数) と tf(タームが出現する文書頻度) を用いた独自の類似尺度により類似性シソーラスを作成し, 拡張に用いている.

これらの方法は, 文書中に共起するタームや各タームの検索文書集合におけるふるまいを利用して関連語の生成を行っており, 特に未定義語にも対応できる点が利点であると考える.

それに対し, 後者の深層的な情報に基づいた方法では, 未登録語には対応できないという欠点を持つが, 一方で 既存の概念を記述した辞書やシソーラス等を用いることによって, 意味的に関連性の保障された概念によって拡張タームを獲得できるという利点を持つ.

このような研究の 1 つに Voorhees らによる WordNet を用いた研究がある [8][9].

Voorheers らは, WordNet における synset と呼ばれる同義語概念を持つタームを基に, 概念シソーラスの上位・同義・下位の概念を持つ synset を獲得し, それら synset に含まれるタームを関連タームとしている.

しかし, このような既存データベースを用いた概念に基づく方法として, このようなシソーラスにおける上下関係のみを用いただけでは, 概念を利用したクエリー拡張の可能性について充分な検討がなされているとは言えないと考えられる.

## 1.3 本研究の目的と位置付け

本研究では これまで充分に研究がなされてきていない “概念” を用いたクエリー拡張の有効性と検索精度への効果について, その可能性を探るために, 既存シソーラス “EDR 電子化辞書” を利用して, 概念に基づいたクエリー拡張を行う.

クエリー拡張においては, 「拡張」と「フィルタリング」という 2 通りの技術が必要とされると考えられる.

そこで本研究では, このような「拡張」と「フィルタリング」の効果を持つ拡張クエリーを生成する手法を各々提案する.

まず 「フィルタリング」の効果を持つクエリーの獲得手法として, 名詞・動詞間の概念関係を用いた拡張手法を提案する. クエリータームと それに共起しやすい名詞 もしくは動詞の共起によって, 正解文書をより精度よく獲得できるようになると期待される.

また, 「拡張」の効果を持つクエリー獲得手法としては, 概念説明文を利用した拡張手法を提案する. 本手法によって, 初期クエリータームと関連があり, かつ 従来の方法では獲得できなかったようなタイプの関連語を得ることができると考えられ, それによって より多くの正解文書を正解として得ることができるようになると考えられる.

ここで、本研究において用いてられる “概念” についての定義を行う。

“概念” の定義

- 言葉が持つ情報を「表層的な情報」と「意味的な情報」に分けた時、その後者を “概念” とする。
- また、本研究では、EDR 電子化辞書における概念識別子を “概念” と定義する。

## 1.4 本論文の構成

本論文は 本章を含め 7 章から構成される. 第 2 章では, 本研究で試作したクエリー拡張システムの概要と提案手法の説明を行う. 第 3 章, 第 4 章では今回提案するクエリー拡張手法についてその目的とクエリー生成についてのアルゴリズム等について述べる.

また 5 章では クエリー拡張の効果をあげるための一手法として, クエリータームの語義曖昧性の解消を概念を利用して行う方法について 3 種類の方法を挙げ, 説明する.

第 6 章では, 提案手法について実験とその評価を行い, 概念を利用したクエリー拡張の情報検索における可能性について考察する.

そして最後に 第 7 章では 本研究のまとめと今後の課題を述べる.

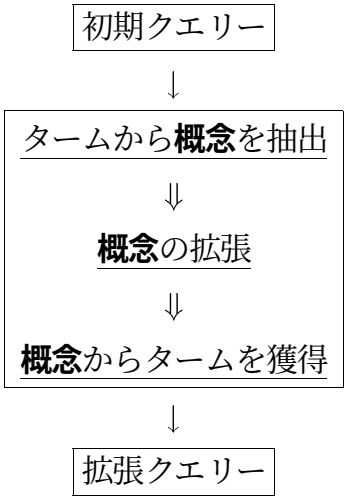
# 第 2 章

# クエリー拡張システムの試作

本研究では, 提案するクエリー拡張の手法等を取り入れたクエリー拡張システムを試作した. 本章ではこのシステムについての説明と, 基本的なクエリー拡張過程における 辞書レコードや概念の処理について説明を行う.

## 2.1 本システムの構成

本システムは, 概念に基づく拡張を行うため, 以下のような基本的な枠組を持つ.




これら一連の基本的な処理を行うことによって, EDR 電子化辞書を使って従来の方法と同様に, シソーラスを利用した { 上位・同義・下位 } 語の獲得が行える.

なお, 本システムでは, AND や OR といった Boolean 演算子を使った検索が可能である.

また本研究では, 以下の 3 手法を実装し, 概念説明文や 様々な概念間関係を利用したクエリー拡張および多義性解消が行えるように この基本システムを拡張している. 1–3 によって拡張されたシステムの概観を 図 2.1に示す.

| | |
|---|---|
| 1 | **名詞・動詞間の概念関係を利用したクエリー拡張** |
| 2 | **概念説明文を利用したクエリー拡張** |
| 3 | **概念を利用した クエリータームの語義の曖昧性解消手法** |

本システム (図 2.1) における処理の手順を以下に順に述べる.

1. 初期クエリーの入力
2. 初期クエリーを形態素解析[1]し [10], 必要な自立語のみを抽出し, さらにそこからストップワードを除いて, 残ったものをクエリータームとて抽出する.

   なお, ここで言う「必要な自立語」には今回 動詞, 形容詞, { 普通・固有・サ変・時相 } 名詞を採用した.
3. 2 で得られた全てのクエリータームに対し, 「関連ターム生成部」の処理を行う

   (a) (左側) クエリーターム概念説明文を利用した拡張を実施

   (b) (右側)

   i. 日本語単語辞書を用いて, 各タームの概念識別子を獲得する.

   ii. タームが多義語である場合, 語義の曖昧性解消手法を適用し, 概念識別子を一に意決定する (決定できなかった場合には全ての語義により拡張を行う)

   iii. 従来法による拡張を行う (なお, 本研究では ここで「同義語による拡張」を行っているため, この部分での入力と出力は同じ概念識別子となる)
4. 各方法で拡張された概念識別子から タームを抽出する.

| | | |
|---|---|---|
| 1 | 単語見出し | 各々 形態素解析を行い, |
| 2 | 概念見出し | 必要な自立語が 2 つ以上得られた場合は, |
| 3 | 概念説明文 | “ターム 1 AND ターム 2” というクエリータームを作る |

[1] 本研究における形態素解析は, 全て 茶筌 version1.5 を用いた.

なお, 3 の概念説明文については, 独自に作成したルールに基づき,
平均で 最後〜最後から 2 番目の自立語を関連タームとして獲得し,
“ターム a AND ターム b” のような AND 項クエリーを作成し,
それをクエリータームとする.

5. 4 で得たクエリータームは そのまま検索に用いる場合もあるが, ここで 必要に応じて 各クエリータームから AND 項を生成したり, クエリーに重みを加えたりする処理を行う.

6. 拡張クエリーの出力
   なお, 拡張クエリーは 5 で生成した全てのクエリータームを OR 演算子で結んだものである.

## 2.2 文書検索システム

本研究では, 1 節で説明したクエリー拡張システムで 各提案手法による拡張クエリーを作った後, それをクエリーとした検索を行い, その精度を調査する必要がある.

そこで, 本研究ではそのための検索システムを構築した. 本節では この検索システムについて説明する.

### 2.2.1 前処理

検索を行うためには, 検索文書をあらかじめ処理して, 検索に適した形式に加工しておかなくてはならない.

本研究では 以下のような処理によって, 検索用データベースを作成する.

1. 全検索文書を形態素解析器にかけ, 必要な自立語のみを獲得する[2].

2. 1 で獲得した 文書に出現する全てのターム (形態素) について, tf.idf 法に基づき 出現文書による重みを算出し, 各文書ベクトルを作成する (詳しくは 6 章で説明する).

[2] ストップワード処理も行う

3. 2 の結果を検索用データベースとして DBM ファイルに変換し, 検索システムに組み込む.

### 2.2.2 検索システムの処理手順

検索システムの概観を 図 2.2に示し, その手順を以下に述べる.

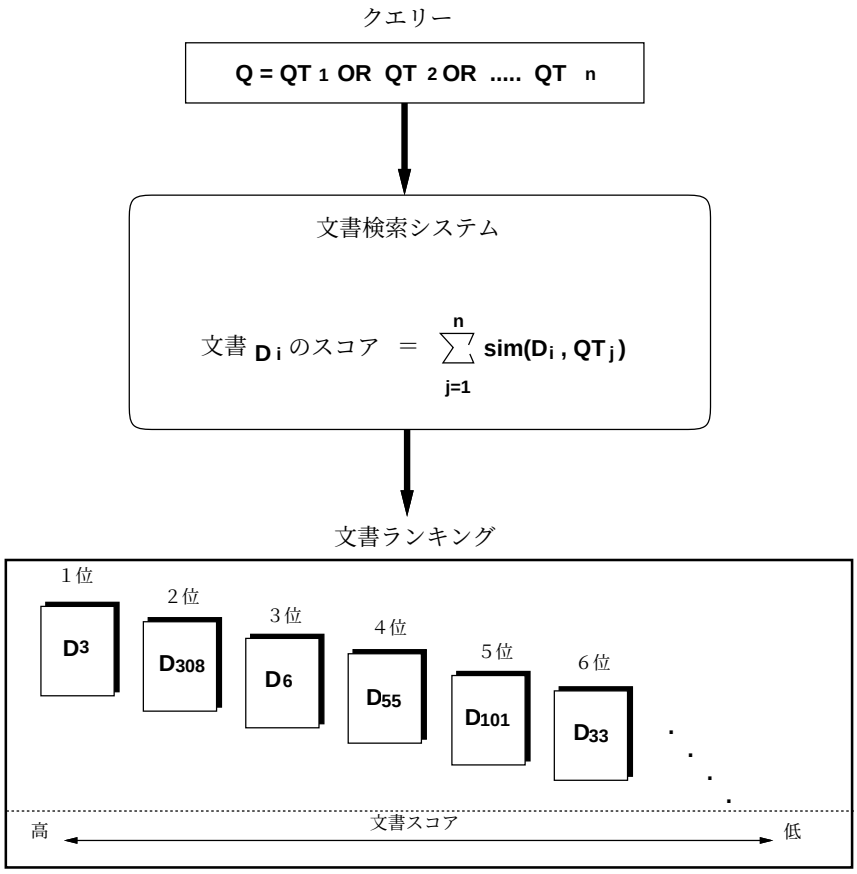



図 2.2: 文書検索システムの概観

1. 入力されたクエリーに対しても, 2.2.1 の 2 を計算する.

2. 検索用データベースを検索し, 文書のベクトルとクエリーのベクトルの類似度を その内積によって求める (詳しい式については 6 章で述べる).

3. 2 で得た類似度をその文書のスコアとして, 検索された全文書よりスコアが高いものほど上位に順位付けされるような正解文書ランキング・リストを生成し, これを検索システムの出力とする.

以降の章では, 上記のようなクエリー拡張システム および文書検索システムを用いることを前提に, 図 2.1に示した 本研究で提案するクエリー拡張手法および 語義の曖昧性解消手法について 個々の説明を行う.

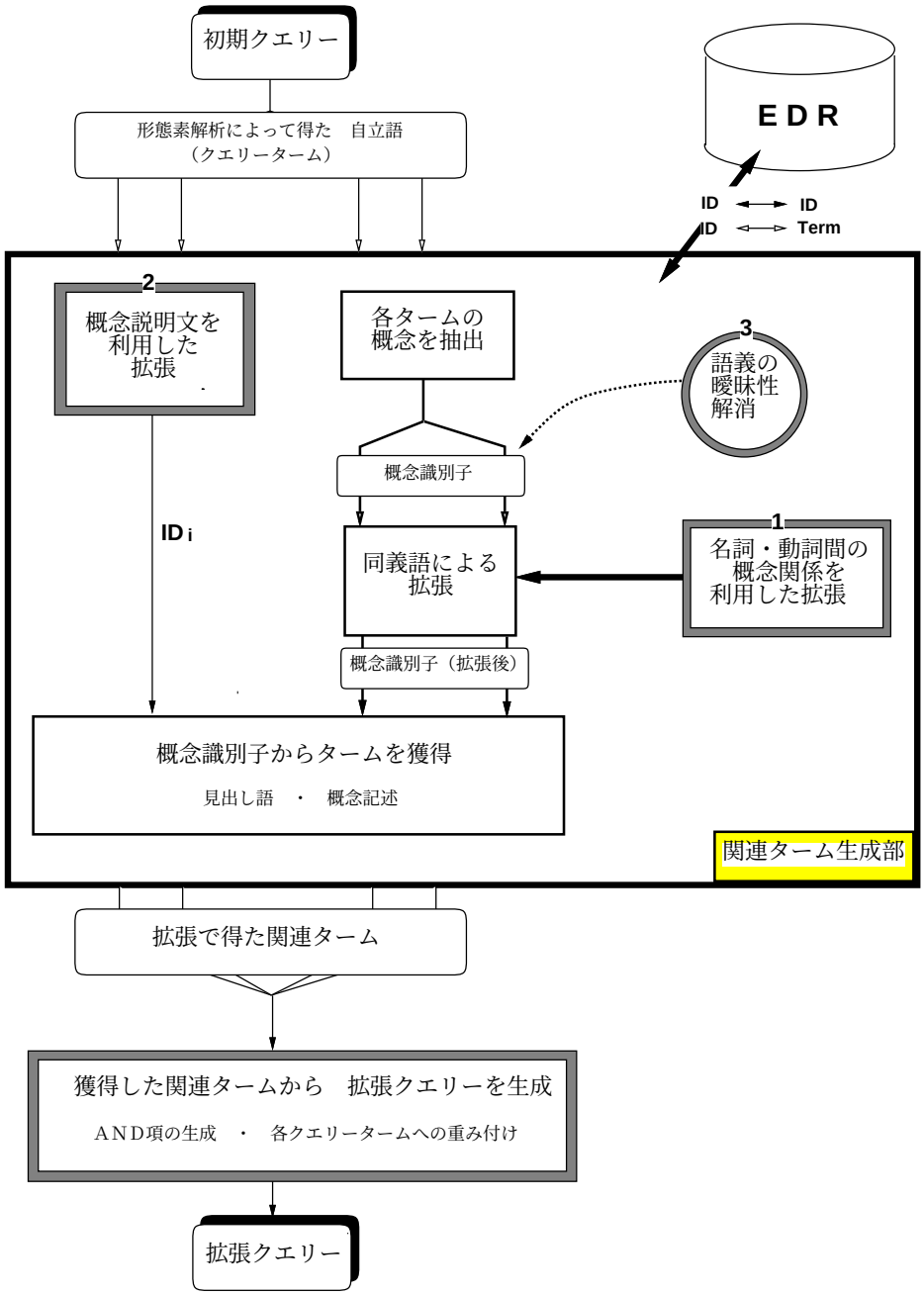



図 2.1: クエリー拡張システムの概要

# 第 3 章

# 名詞・動詞間の概念関係を利用したクエリーの拡張

クエリー拡張を行うことによって, 初期クエリーに何らかの情報が付加されることになる. そのため, クエリー拡張の効果として, 拡張前よりも 多くの正解文書を獲得できる可能性が高まることになるが, それと同時に, システムが不正解文書も正解として獲得してしまう可能性も高くなる. つまり, クエリー拡張においては, 情報の「拡張」技術と同時に, 必要な情報だけをうまく抽出するという情報の「フィルタリング」技術も必要とされる訳である.

そこで, 本研究では, 「フィルタリング」に着目したクエリー拡張の一手法として, 「名詞的な概念と動詞的な概念」という 2 つの概念間関係を利用して関連語を獲得し, それをフィルタリングに用いる手法を提案する. 本章では, 初めに本手法により獲得される関連語と検索におけるその効果について説明し, その後, 具体的な関連語獲得の手順およびその関連語を用いた拡張クエリーの生成方法について述べる. またさらに, 上記の方法を使って, より効率的に関連語を獲得するために行った “利用する概念識別子のフィルタリング” についても述べる.

## 3.1 名詞・動詞間の概念関係を利用した関連語の獲得と検索における効果

**名詞・動詞間の概念関係の利用について**

2 つのターム間の関係には, シソーラスに表されるような上位・下位関係の他に, *object*(動作・変化の影響を受ける対象) や *agent*(主体) といったような関係がある. これらの関係は格関係と呼ばれ, 特に上記のように, “表記’ に依らない意味的な格関係を「深層格」と呼ぶ.

このような深層格で支配されるという条件下では, 2 つのタームの語義, すなわち 本研究における「概念」は一意に決まると考えられる.

例えば, 文「電話をかける」を考える. EDR 日本語単語辞書において, 文中の名詞「電話」は 2 種, 動詞「掛ける」は 24 種の語義を持つ. これらの中で, この文の意味「電話を使って, 誰か人と話す (話そうとする)」に一致するような語義のペアは, 「電話」では「0ffdaa(電話機を使った通話)」, 「掛ける」では「3c1208(電話する)」のみであり, この 2 つの概念識別子は *object* という深層格によって支配される. このように, 2 つのタームは, ある深層格を伴って実際には 2 つの概念間関係を表す.

そこで本手法では, 人手で作成された深層格を伴う 2 つの概念間関係を記述したデータベースを利用することによって, クエリータームの持つ概念から意味的に関連性のある概念を獲得し, それに基づき拡張タームを生成する.

ここでは, データベースとして EDR の概念記述辞書を利用する. 概念記述辞書は, 2 つの概念間関係のうち, 「動詞的概念」と, それがある深層格によって支配する「名詞的概念」のペアを記述した辞書である. つまり, このデータベースの使用により, 意味的に最もらしい体言と用言の共起による拡張を行うことができる.

なお, この辞書では深層格として, {*object, agent, goal, implement, a-object, place, scene, cause*}[1] の 8 種のうちいずれかを持つ概念間関係のみが記述されている.

概念記述辞書のレコード例を図 3.1に, 関連概念の獲得例を図 3.2に示す.

[1] EDR で定義された深層格「概念関係子」は他にも存在するが, ここで使われる 8 種についてのみ その説明を付録 Aに付す.

| 名詞的概念 | 概念関係子 | 動詞的概念 |
|---|---|---|
| **0ffdaa** | **object** | **3c1208** |

図 3.1: 概念記述辞書のレコード例 および 概要

**関連概念として獲得**

⇓

| 名詞的概念 | 概念関係子 | 動詞的概念 |
|---|---|---|
| **0ffdaa (通話)** | object | 3c1208 (電話する) |
| | object | 0e5308 ((情報) を入れる) |
| | scene | 0e6ee1 (応対する) |
| | implement | 0f9c38 (説得する) |
| | implement | 0fbf9e (確かめる) |
| | implement | 0fe90a (伝える) |
| | object | 3bba73 (殺到する) |

※ () 内は 概念見出し

図 3.2: クエリーターム 「電話」の語義 **“0ffdaa”** の場合の関連概念の獲得例

**本手法で得られた関連語の, 検索における効果について**

上記のように, 本手法によって あるクエリータームから, 意味的に最もらしい共起関係を持つような体言 もしくは 用言が獲得できる. しかし, この手法で獲得されたタームをそのまま拡張クエリーとした場合, あまり意味をなさないのではないかと考える.

例えば, 体言「電話」から 用言「掛ける」が獲得された場合を考える.
本来 本手法は体言と用言の共起によって得られる意味的な最もらしさをクエリー拡張に利用している. しかし, もし今「掛ける」のみを拡張タームとしてしまえば,「掛ける」が出現する文書であれば「電話」が出現していなくてもシステムは正解としてその文書を獲得する. つまり, 「掛ける」のみでは, 拡張の源となっている概念間の意味的な関連性を利用できず, 結果的に不要な正解文書 (不正解文書) を導いてしまうと考えられる.

よってここでは, 本手法により生成された関連語を単独で使用せず, 拡張のもととなっているクエリータームおよびその同義語との AND 項を生成し, それを拡張タームとして利用する (上記の例では 「電話 AND 掛ける」が拡張タームとなる).

そして, このようにして生成された AND 項を同義語のみによる拡張で得られたクエリーに加えることによって, 従来の同義語のみによる検索精度と比較し, より高い適合率を得ることを目的とする.

## 3.2 拡張クエリーの生成方法

### 3.2.1 名詞・動詞間の概念関係を利用した関連語の獲得方法

本手法では以下の手順によって, クエリータームの関連概念を獲得し, さらに その関連概念から関連語を獲得する.

1. 初期クエリーを形態素解析器にかけ, 2 章での説明と同様に クエリー拡張に使用するタームのみを抽出する.

   なお, その際には 本手法においては, 品詞が動詞 または 名詞のもののみを抽出することとし, 抽出の際には 各タームの品詞も記憶しておく.

2. 1 で得たターム毎に, 日本語単語辞書を用いてそのタームの持つ全ての概念識別子を獲得する.

3. 2 で得た概念識別子 "$ID_a$" 毎に, 概念記述辞書を検索する.

   なお, その際には, 1 で記憶しておいた品詞記述を使用し, クエリータームの品詞が

   - "名詞" だった場合には, 「"名詞的概念" が "$ID_a$" である」レコード
   - "動詞" だった場合には, 「"動詞的概念" が "$ID_a$" である」レコード

   を検索する. また得られた全レコードから, 前者の場合は 動詞的概念識別子を, 後者の場合は 名詞的概念識別子を関連概念として獲得する.

4. 3 で得られた概念識別子を基に, 2 章と同様の方法で, 単語見出し・概念見出し・概念説明文より関連語を獲得する.

### 3.2.2 クエリータームの生成方法

3.1節で述べたように, 3.2.1で得た関連語をそのままクエリーとして検索に用いることは意味をなさないと考えられる.

そこで, ここでは 3.2.1で得た関連語を利用してどのような拡張クエリータームを生成するかについて, その手順を説明する.

なお, 本研究では 図 3.3に示した通り, この手法で得られた関連語を同義語による拡張クエリーに加えたものを拡張クエリーとして検索に使うことによって, 同義語のみによる拡張クエリーよりも高い適合率を得ることを目的としている.

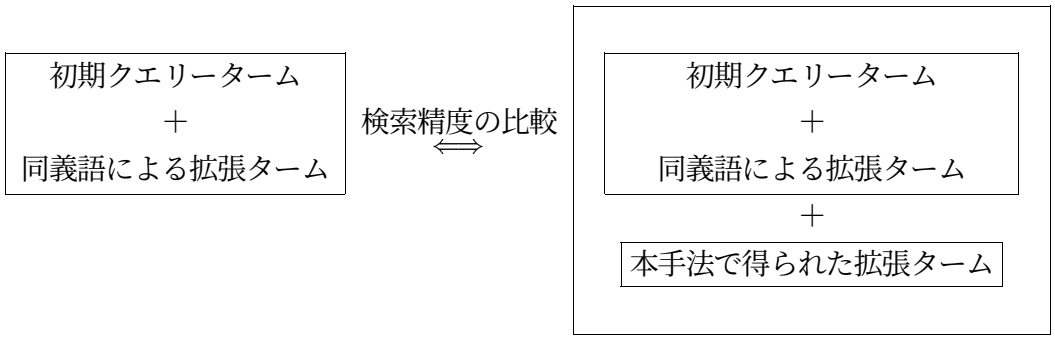



図 3.3: 本手法により生成される拡張クエリーと比較対象の関係

1. 今, 初期クエリーは $n$ 個のクエリータームからなるとする.
   3.2.1で得た関連語を “$t_{nv}$” とし, その集合を “$T_{NV}$” と表す.
   また 2 章で説明した方法で, 初期クエリータームの同義語 “$t_s$” および 初期クエリータームを合わせた集合 “$T_{SAME}$” を作成しておく ($T_{SAME}$は, 図 3.3の左側のセットにあたる. ).

2. 1 で準備した拡張タームを使って, 以下の (a)–(c) の 3 通りの拡張クエリータームを生成し, 実験により各検索精度の違いを調査する.

   なお, $t_{s_m}$は 初期クエリーターム $t_{f_m}$から得られた同義語を表す.

   (a) $\underline{t_{s_m}}$ AND $t_{nv}$
   ～ 同義語 $t_s$ のうちいずれか 1 つ (下線部分) と $t_{nv}$ との AND 項

(b) $\underline{t_{s_l} \text{ AND } t_{s_m}} \text{ AND } t_{nv}$

$\sim$ 異なる 初期クエリーターム $t_{f_l}$ と $t_{f_m}$ から得られた同義語 $t_{s_l}, t_{s_m}$ (下線部分) と $t_{nv}$ との AND 項

(c) $\underline{t_{s_1} \text{ AND } \cdots \text{ AND } t_{s_m} \text{ AND } \cdots \text{ AND } t_{s_n}} \text{ AND } t_{nv}$

$\sim$ 初期クエリー中の 全 $n$ 個のタームから得られた同義語 (下線部分) と $t_{nv}$ の AND 項

(b),(c) のタイプのクエリータームは, 上記の (a) に比べ “より多くの同義語の共起” によって制約を強めたものと言える.

以上のような (a)–(c) の 3 種類のタイプ別に, $T_{NV}$ 中の全ての $t_{nv}$ を使ってクエリータームを生成し, できたクエリータームを全て “OR” で結んだものを拡張クエリーとして, 検索システムの入力に使用する.

## 3.3 拡張に利用する概念識別子のフィルタリング

これまでに, 本手法における拡張方法 および クエリータームの生成方法について述べた.

3.2.1の手順 **3** で述べたように, 全節までの方法では, 概念記述辞書のレコード 全てを関連概念獲得のために利用している. しかし, 概念記述辞書のレコードを調査したところ, 多くの “動詞的概念” と共起する “名詞的概念” が存在し, またその逆のケースも見られた. つまり, 全節までの方法では, 多くの “動詞的概念” と共起しやすい 「一般的な」“名詞的概念” と特殊な “動詞的概念” としか共起しないような「特徴的な」“名詞的概念” とを区別なく拡張に使用してしまっていることになる.

そこで本節では, ある基準にのっとって「特徴的な」概念ほど大きな重みを付け, 閾値以下の重みを持つ概念は 「一般的」であるため拡張に使用しないこととする「概念識別子のフィルタリング」方法を提案する.

本手法では, 拡張に用いる概念識別子を決定する閾値を調節し, 最適な閾値の設定による検索精度の向上を導くことを目的とする.

具体的なフィルタリング方法について以下に述べる.

本手法では, tf.idf 法での考え方を応用し, 以下のような手順でフィルタリングを行う.

process 1

tf.idf の世界における “文書” を ここでは 概念記述辞書のレコード に置き換え, ある

概念識別子 $ID_a$を「名詞的概念」として含むレコードの数の逆数を $ID_a$ の重みとして計算する[2].

なお, 概念記述辞書では 同一の概念識別子が, レコードによって名詞的概念識別子として扱われている場合と動詞的概念識別子として扱われている場合の両者の可能性がある. よって $ID_a$に対する重みの計算は, 「名詞的概念 としての $ID_a$」の場合と「動詞的概念 としての $ID_a$」の場合の両方を計算することとする.

process 2

概念記述辞書中の全ての概念識別子に対して 1 の計算を行い, 全てが終了したら, “キー” を 名詞的概念識別子, “値” を共起する動詞的概念識別子の異なり数とするような検索用データベースを作成する (動詞的概念識別子については, “キー” が 動詞的概念識別子, “値” が 共起する名詞的概念識別子の異なり数となる).

process 3

初期クエリータームの持つ概念識別子の拡張過程における手順は 3.2.1で述べたものとほぼ同様だが, 手順 **3** においては, 以下の下線部分のように変更して実行する.

1. ～ 2. 3.2.1と同様.

**3.** 2 で得た概念識別子 “$ID_a$” 毎に, 概念記述辞書を検索する.

なお, その際には, 1 で記憶しておいた品詞記述を使用し, クエリータームの品詞が

- “名詞” だった場合には, 「“名詞的概念” が “$ID_a$” である」レコード
- “動詞” だった場合には, 「“動詞的概念” が “$ID_a$” である」レコード

を検索する.

また得られた全レコードから, 前者の場合は 動詞的概念識別子を, 後者の場合は名詞的概念識別子を抽出し, 抽出した概念識別子をもとに, 上記の “process 2” において作成したデーターベースを検索する.

そして, データベースから得られた “値” が 閾値以上だった場合のみ, その概念識別子を関連概念として獲得する.

[2]tf にあたる「あるレコードにおける概念識別子 $ID_a$の頻度」は, いずれのレコードにおいても 1 か 0 であるため, この場合は Record Frequency の逆数のみを重み付けに用いる

**4.** 3.2.1と同様.

以上のような **process 1** − **process 3** を複数の閾値において行い, 各閾値の場合の検索精度の変化を実験によって調査する.

# 第 4 章

# 概念説明文を利用したクエリーの拡張

本章では, EDR の概念説明文 (語釈文) を利用して, 初期クエリーに関連する概念を獲得するという方法について述べる.

ここでは, まず本手法の特徴を述べた後, その具体的な拡張手法について述べる.

## 4.1 概念説明文を利用したクエリー拡張

前章までで述べて来た拡張方法では, 初期クエリーの持つ概念をシソーラスや共起データを用いて拡張し, その結果得られた概念から, その概念説明文中に出現するターム他を関連語として獲得した.

それに対し, 本手法では, 「概念説明文中にクエリータームが出現するならば, その概念説明文の持つ概念とクエリータームとの間には何らかの関連がある」という仮説に基づき, ある概念説明文中にクエリータームもしくはその同義語が 2 つ以上共起していた場合, その概念説明文はクエリーに関連する概念についての説明文であると判定し, その概念を関連概念識別子として獲得する. つまり, 前章までの手法で拡張の結果として獲得していた情報を, 本手法では 拡張の源情報として用いて関連概念を得る訳である.

これまで, 語釈文中のタームの情報は主に語義の曖昧性解消に利用されてきているが [13] [14] [17], それは 語釈文中に出現する複数のタームの情報が意味的な制約になると考えられるためである. そこで本研究では, この特性をクエリー拡張に用いることにより, これまでのシソーラスを用いたクエリー拡張の研究で行われて来たような同一・上位・下位概念を利用した拡張とは異なるタイプの関連語を獲得しようとするものである.

では, 次の章では 本手法による関連概念の獲得方法について具体的に述べる.

## 4.2 関連概念識別子の獲得手法

本手法では, 以下のアルゴリズムにしたがって関連概念を獲得し, そして最終的にその概念から関連タームを獲得する.

なお, ここではクエリー「国内・航空・大手」中のタームのうち, 「国内・航空」というタームセットを扱う場合の例を付す.

1. 初期クエリー中のターム別に, 2 章で説明した手順で同義語集合を作成する.

   (例) 「国内」の同義語集合 ( この場合 「国内」自身も含む) を 同$_{携帯}$ と表す.
   「航空」についても同様.

2. 異なる初期クエリータームの同義語からなるタームのペアを, 全ての組み合わせ分作成する.

   (例) ( [**同$_{国内}$のうち 1 ターム**], [**同$_{航空}$のうち 1 ターム**])
   を満たすようなタームのペアを作成する.
   例：(国内 , 航空), (国内 , フライト),...etc.

3. 2 で作成したタームのペアを含む概念説明文[1]を検索する.

   なお今回は, 2 つのタームは各々概念説明文中のどこに出現していても良いとし, 構文情報等は用いず検索を行っている.

4. 3 で得られたレコードから 概念識別子を抽出し, これを関連概念として獲得する.

   (例) 概念説明文中に (国内 , 航空) を含むレコードの例

| 概念識別子 | 概念説明文 |
|---|---|
| 0f11cf | 国内の空港を結ぶ航空路線 |

⇒ 0f11cf を獲得

[1] ここでは, 概念見出し辞書のレコードを利用

5. 4 で獲得した全ての概念識別子より, 日本語単語辞書を用いてこの概念識別子を持つ「単語見出し」または「概念見出し」[2] を抽出し, 関連ターム (拡張ターム) とする.

(例) 概念識別子 “0f11cf” を持つ日本語単語辞書のレコード例

| 単語見出し | 概念識別子 | 概念見出し | 概念説明文 |
|---|---|---|---|
| 国内線 | 0f11cf | 国内線 | 国内の空港を結ぶ航空路線 |

抽↓出　　抽↓出

「国内線」が関連タームとして獲得される

[2]「概念見出し」については, 概念見出し辞書から獲得してもよい.

# 第 5 章

# 概念を利用した, クエリータームの語義の曖昧性の解消

クエリー拡張において, 情報のフィルタリング技術が必要とされることは, 1 章をはじめ各所で述べた通りである. そこで本研究では, フィルタリングの一手法として, 初期クエリータームの語義の曖昧性解消に注目し, この曖昧性解消により, 「検索精度の向上に有効な拡張タームを効率的に獲得できる」という仮定をたて, その実証を試みる.

本章では, まず クエリー拡張における クエリータームの語義の曖昧性解消の有効性について述べ, その後 本研究で提案する概念を利用した 3 種の語義の曖昧性解消手法について説明を行う.

## 5.1 クエリー拡張におけるクエリータームの語義の曖昧性解消

今 システムに

**航空**　AND　**大手**

という 2 つのタームからなるクエリーが 入力されたとする. そして, 「大手」は 図 5.1に示すような 6 つの語義を持つ多義語であるとする.

この場合, 「『航空』が共起する」という条件を使用して, 大多数の人は「大手」の語義が “f” であると答えることができるだろう. しかし, もしクエリーが「大手」だけであっ

| | |
|---|---|
| a. | 肩から手の指先まで |
| b. | あたりをはばからない態度をとること |
| c. | 喜んで人を迎えること |
| d. | 敵の正面から攻め込む部隊 |
| e. | 多額の商取引をする人 |
| f. | 大量の取引をする会社 |

図 5.1: 「大手」の語義の例 (EDR 日本語単語辞書による)

たなら, その「大手」がどの語義を指すのかは このクエリーを入力した本人しかわからないはずである. このように, 人は「『航空』が共起する」という条件のような 2 つ以上のタームの共起を以って生じる制約をもとに, 各タームの語義を推測・決定していると考えられる. そこで, 本研究では この「共起によって生じる制約」を電子化辞書中の概念に関する記述から見付け出し, それを用いて初期クエリー中の各タームの語義を決定することを試みる.

本クエリー拡張システムでは, 2 章で述べたように, まず 「初期クエリー中の各タームの持つ概念 (概念識別子) を抽出」し, その後 「抽出された “概念” に対して 様々な拡張を行う」という基本的なプロセスを経て, 検索に有効な拡張クエリーを生成する. したがって本システムでは, 拡張の前段階で 初期クエリー中のタームの語義の曖昧性を解消することによって, 複数タームの共起によって最もらしいと判断された概念のみを利用した拡張を行うことが可能となる. そしてこのように, “概念” の段階で不適切な拡張源を除去することによって, これまで不正解を導いていた不要な拡張タームの削減と, 初期クエリーに意味的に関連した有効な拡張タームの獲得を効率良く実現することができると考えられる.

## 5.2 概念を利用した語義の曖昧性解消

本研究では, 概念を利用した語義の曖昧性解消の手法として,

1. 語釈文とシソーラスを利用した手法
2. 日本語共起辞書を利用した手法

3. 拡張で得られた共通概念を利用した手法

という 3 つの方法を提案する.

語義の曖昧性の解消に関しては, これまで 非常に多くの研究が行われて来ているが [12], ここでは, 各手法についての関連研究に触れながら, 本研究で提案する手法の特徴について述べる.

1. 語釈文とシソーラスを利用した手法

   機械可読辞書中の語釈文を用いた曖昧性の解消に関する研究では, 同一文中に共起する 2 つの単語 (多義語) の語釈文を利用し, 「各々の持つ語釈文中に共通タームを含むものがあれば, その共通タームを含む語釈文の語義を 各単語の語義に決定する」という手法 [13][14] や, 語釈文における単語間の共起頻度をもとに各単語の特徴を表すベクトルを作成し, そのベクトルの類似度を計算することによって語義を決定する手法 [15] の他, 単語の属性として語釈文中に出現するタームを用い, ユーザの観点に適した類義語の獲得を行った研究 [17] 等がある.

   [13][14] の手法は, 「同一文中に出現する 2 つの単語は意味的に似ている」という前提に基づくものであり, 本研究でもこの手法の実装・実験を行ったが, 上記の条件を満たす初期クエリーは今回の実験セット中には存在しなかった. これについては, 実験結果の分析の結果, 「共通タームを含む」条件による制約が強すぎたため, 該当する概念説明文を持つような単語のペアが獲得できなかったことが主な原因であることがわかった. よって, 単語を意味的にまとまったいくかの単語の集合 (カテゴリ) に分類し, 条件を「語釈文中に同じカテゴリに属すのタームを含む」というように緩和することによって, この前提により近い実装が行えるのではないかと考える.

   また, [17] では, 語釈文中に出現するタームをその見出し語の概念の特徴を表す属性として概念データベースを作成し, それに基づきユーザから与えられた観点 (属性値) からユーザの意図する概念に対応した類義語の獲得を行うもので, 語釈文中のタームを概念の区別に用いることによって, 語義の曖昧性の解消を行っている点が興味深い.

   このように, これまでに様々な方法で語釈文を多義性解消に用いた研究が行われて来ているが, これらに対し, 本研究では, 辞書中の語釈文の説明文としての特性に注

目して,「語釈文中に出現する単語の持つ語義と その語釈文が表す語義は意味的に似ている」という前提に基き,「語釈文中に 2 つの初期クエリータームが出現している場合, その一方の語義のいずれかと その語釈文の語義が同義 もしくは 上位・下位関係を持つならば, その時の語義をそのクエリータームの語義として決定する」という方法で, 語義の曖昧性解消を試みる. ここで用いる 2 つの概念の上位・下位関係については, 概念体系辞書という概念シソーラスを利用して獲得する.

語義決定の条件を「同義 もしくは 上位・下位関係をもつ場合」と設定したのは,「説明文が表す語義は, 説明文中に出現する単語の語義よりも下位の概念ではないはずである」という仮定に基づいたものである. また, 単に クエリー中のターム 1 つだけが出現するような語釈文を用いると, クエリー全体が表す概念と全く関連のない概念が多く獲得されてしまう恐れがあるため, ここでは「もう一方のクエリータームが共起する」という条件を付すことにより意味的な制約を作り, それを防ぐ.

上記の関連研究の他, 既存のシソーラスを用いた多義性解消の研究として, Voorhees による WordNet を用いた研究がある [16][9].

[16] では, WordNet の synset [1] および シソーラスを利用して, 各単語の持つ語義を表現するようなカテゴリ (synset よりも大きな単語集合) をヒューリスティクスによって作成し, そのカテゴリを利用することによって語義の曖昧性の解消を試みている. この研究の曖昧性解消の手法自体は本提案手法とは全く異なるタイプのものだが, シソーラスを利用した上位・下位概念語によるカテゴリの作成は, 本提案手法にも応用できるものとして興味深い.

2. 日本語共起辞書を利用した手法

一方, コーパスを用いた過去の多義性解消の研究では, 見出し語への語義の付与 (tagging) に対訳コーパス中の対訳文を利用し, その上で単語間の共起確率を利用した語義決定を行った研究 [18] など, 統計的手法を用いた方法が多く行われている. しかし, このような統計的な手法では, 統計処理に用いるデータを取る際, 同綴異義語の違いを考慮できないという問題がある. そこで本手法では, 単に共起句の表層的情報だけでなく, 人手によって付与された語義ラベル (概念識別子) 付きの共起データである日本語共起辞書を用いて, 初期クエリータームの語義の曖昧性解消を試みる.

[1] WordNet で定義されている, 同じ語義を持つ単語の集合

本手法では, 初期クエリータームの持つ語義から概念識別子のペアを作成し, 共起辞書中にそのペアを含むレコードがある場合のみ語義決定を行う. また今回は, 一般的に行われているような統計的手法は採らず, 確実な語義のみを決定するという立場をとって非常に簡単な方法のみで曖昧性の解消を試みているが, 本手法は, 語義ラベル付き共起データによるクエリータームの語義の曖昧性解消による効果を知る上で意義のあるものであると考える.

3. 拡張で得られた共通概念を利用した手法

   本手法は, 「初期クエリー中の 2 つのタームにおいて, ある拡張手法によって共通のタームが得られた場合, その時の拡張もととなった各タームの語義を各々のタームの正しい語義とする」というものである. これは「2 つのタームの語義から共通タームが得られるならば, その語義間になんらかの意味的な関連性がある」という仮説に基づくもので, 今回は, 拡張手法として 3 章でも使用した「動詞的概念と名詞的概念の関係」を用いた拡張手法を使い, 拡張によって得られる共通ターム数を利用したヒューリスティクスを用いて, 語義の曖昧性解消を試みる.

   ここでは, 概念記述辞書という共起データを使用することになるが, この共起データは 2 つの “概念” に関する共起データであり, ここでも上記の 2 で述べたような同綴意義語に関する問題は回避できると考える.

## 5.3 語義の曖昧性解消手法について

ここでは, 5.2節で提案した各手法について, 具体的な語義曖昧性解消の方法を詳しく説明する.

### 5.3.1 語釈文とシソーラスを利用した手法

本手法では, 図 5.2に示すように, 「語釈文中に 2 つの初期クエリータームが出現していて, なおかつ 一方のタームが持つ語義のいずれかと その語釈文の語義が同義 もしくは 上位・下位関係を持つ場合, その関係成立時の語義をそのクエリータームの語義とする」という方法によって語義の曖昧性の解消を試みる.

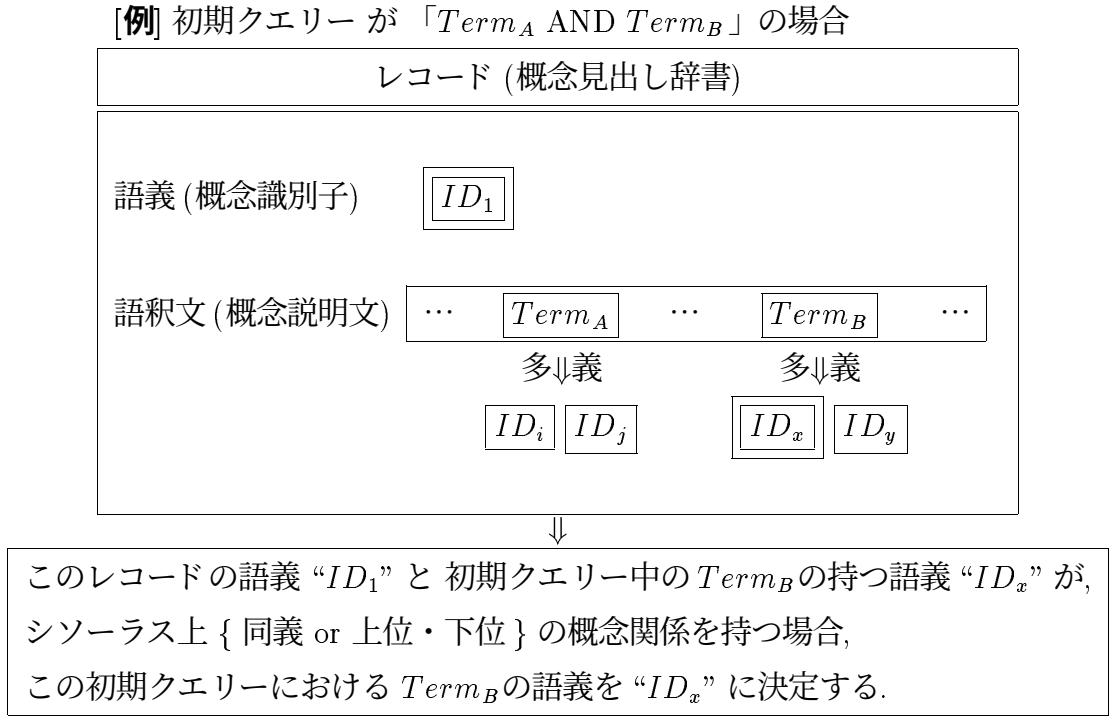



図 5.2: 語釈文とシソーラスを利用した語義の曖昧性の解消方法

本手法におけるクエリータームの語義の曖昧性解消は, 以下のアルゴリズムに従って行われる.

なおここでは, 「携帯・電話」という 2 タームからなるクエリーを例として付す.

1. 〜 3. は 4.2 と 同様.

4. シソーラスとして概念体系辞書を利用し, 3 で獲得したレコード中の概念識別子の上位概念を検索する.
   得られた上位の概念識別子と 概念説明文中に出現するクエリータームの持つ概念識別子のうち, 一致するものがあればそれを正解とする.

   ただし, この方法によって, クエリータームの語義として, 複数の候補があがった場合は, 語義の決定は行わない.

   (例) クエリーターム「電話」の持つ語義のうち,「電話機」を表す概念識別子 "0ea3d0

” と レコードの持つ概念識別子の上位概念 “3becc6” が一致した例.
(図 5.3は この例を含むシソーラスの一部)

● ← 0ea3d0(「電話機」)

◎ ← 3becc6(「データホン」)

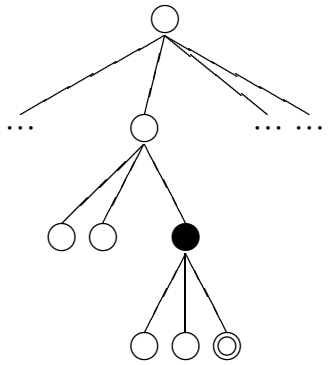


図 5.3: 概念識別子 0ea3d0(「電話機」) を含むシソーラスの一部

### 5.3.2 日本語共起辞書を利用した手法

本手法は, 以下のような手順によりクエリータームの語義決定を行う. なおここでは,「携帯・電話」という 2 タームからなるクエリーを例として付す.

1. 初期クエリー中の各タームが持つ概念識別子から, 異なる 2 つのタームの概念識別子からなる概念のペアを全ての組み合わせを網羅するように作成する.

   (例) クエリーに, 2 つの語義を持つ「電話」が含まれるため, 以下のような 2 種類の概念識別子のペアが生成される.

| クエリーターム | タームが持つ概念識別子 (概念見出し) |
|---|---|
| 「携帯」 | 3d007b |
| 「電話」 | 3bdeeb(通話), 0ea3d0(電話機) |

⇓

生成される概念のペア

| (3d007b,3bdeeb) , (3d007b , 0ea3d0) |
|---|

2. 1 で生成された概念のペアが, 以下のいずれかの条件を満たすような日本語共起辞書のレコード[2]を検索する.

   - (「係り側概念識別子」,「受け側概念識別子」)
   - (「受け側概念識別子」,「係り側概念識別子」)

   (例) (「係り側概念識別子」,「受け側概念識別子」) =(3d007b,3bdeeb)
   であるレコードの抜粋

| | | | |
|---|---|---|---|
| < 共起句構成要素情報 > | | | |
| < 要素番号 > | 1 | 2 | 3 |
| < 形態素表記 > | 電話 | を | 持 |
| < 概念識別子 > | 3dbeed | - | 3d007b |

3. 2 の検索で得られた全てのレコードをもとに, 各ターム毎に以下のルールを適用する.

   (a) 全くレコードが得られなかったタームについては, 語義は決定できないと判定する.

   (b) あるタームに対して, 該当レコードは得られたが, 語義の候補が複数存在した場合, 語義は決定できないと判定する.

   (c) あるタームの概念識別子の候補が 1 つだけであり, かつ その他のタームの語義決定において, 矛盾が起きない場合のみ, その概念識別子をそのタームの語義に決定する.

   本手法では,「共起データによって語義の曖昧性が全くないという確実な場合に限って, 語義を決定する」という立場をとり, あるタームについて, ルール 3 を満たす条件がそろっていた場合は語義の決定を行うが, それ他の場合については語義の曖昧性を持たせたままにしておく.

   (例 1) 語義のペア (3d007b,3bdeeb) を含む 3 レコードのみが獲得された場合, クエリーターム「電話」の語義を “3bdeeb” に決定する.

   (例 2) 語義のペア (3d007b,3bdeeb) を含む 1 レコード および (3d007b,0ea3d0) を含む 2 レコード が獲得された場合, 語義は決定できないと判定する.

[2]付録 B参照

### 5.3.3 拡張で得られた共通概念を利用した手法

本手法では, 概念記述辞書を用いた拡張手法をもとに, 共通の概念識別子を生成する語義のペアを見付け, 共通概念数を利用したヒューリスティクスによって, 語義の曖昧性の解消を試みる.

本手法では, 以下の手順に従い, 語義の曖昧性解消を行う.

なおここでは, 「携帯・電話」という 2 タームからなるクエリーを例として用いる.

1. 初期クエリー中タームが持つ概念識別子毎に, 3 章で述べた方法を用いて拡張概念を獲得し, その数もカウントしておく.

2. 初期クエリー中の各タームが持つ概念識別子から, 異なる 2 つのタームの概念識別子からなる概念のペアを全ての組み合わせを網羅するように作成する.

3. 2 で得られた概念識別子のペア毎に, 1 で個々の概念識別子から得た拡張概念を比較し, 共通する概念識別子の数をカウントする.

4. 2,3 で得た概念識別子の数を利用し, 以下のようなヒューリスティクスによって語義を決定する.

   以下では, 「携帯 (3d007b)」「電話 (3bdeeb,0ffdaa)」の 2 タームからなるクエリーを直接例にとって説明する (括弧の中は 各タームの持つ概念識別子)

   (a) まず, 表 5.1のように, 片方の概念識別子が共通な概念のペアを比較対象として用意する (ここでは, この 2 組の概念のペアを「比較セット」と呼び, この比較セットにおいて, 共通な概念識別子を「基準 ID」, それ以外の概念識別子を「比較 ID」と呼ぶ).

   (b) 上記の 2,3 より得られたデータから, 比較 ID に対して各概念のペアに対するスコアを計算する.
   スコアは以下の式 5.1による.

$$\text{スコア} = \frac{\text{基準 ID と比較 ID の拡張によって獲得した共通の概念識別子の数}}{\text{比較 ID の拡張によって獲得した概念識別子の数}} * 100 \qquad (5.1)$$

このスコアは, 「基準 ID にとって, 比較 ID のうちのどれが最もふさわしい語義か」を表すものとし, 比較 ID 間で公平なスコア比較を行うために, ここでは「共通の概念識別子の数」を「比較 ID の拡張によって得た概念識別子の数」によって正規化したものを使用する.

(c) b で得たスコアを比較し, スコアの高い比較 ID を, ターム「電話」の持つ語義のうち, 基準 ID から見て最もふさわしい語義であると決定する ( この語義を基準 ID に対する最適 ID と呼ぶ).

なお, スコアが同点であった場合は, 最高スコアを持つ比較 ID 全てを正しい語義とし, 全ての比較 ID のスコアが 0 だった場合には, この基準 ID と比較 ID における語義の曖昧性解消は不可能とする. [3]. また, 比較 ID が 1 つしかない (「携帯」のような多義でない語の語義が比較 ID である) 場合は, その基準 ID にとってその比較 ID がふさわしいという判定しておく.

このように, ここまでで, 全ての比較ID のスコアが 0 でない限り, ある比較セットにおける基準 ID と 最適 ID のペアが生成されることになる.

表 5.1: 共通の概念識別子数を利用した多義性解消の例

| クエリーターム | 区別 | 概念識別子 | 獲得した概念識別子数 | 獲得した**共通の**概念識別子数 | スコア | |
|---|---|---|---|---|---|---|
| 携帯 | 基準 ID | 3d007b | 353 | - | - | |
| 電話 | 比較 ID1 | 0ffdaa | 50 | 4 | 8.000 | |
| | 比較 ID2 | 3bdeeb | 71 | 15 | 21.127 | ←決定 |

(d) 2 で生成された全ての概念のペアから作り得る全比較セットに対して b,c の処理を行い, その結果をもとに, 以下のルールに従って語義の曖昧性解消を行う.

[3] 比較 ID のスコアが全て 0 である場合は, 全ての比較 ID において基準 ID との共通する拡張概念識別子が得られないという状態である. よって, これの状態は「基準 ID にとって全てがふさわしくない」と考えるより, 「本手法がうまく適用できない」ためであると考える方が妥当であり, ここでは「この比較セットの基準 ID についての最適 ID は判定できない」と判定するようにした.

**ルール** 2 つのクエリータームの持つ概念識別子が 2 つの異なる比較セットにおいて基準 ID と最適 ID である場合, 各クエリータームの語義を各比較セットにおける最適 ID (もしくは基準 ID) に決定する.

よって, クエリー「携帯・電話」の場合「携帯」の持つ概念識別子 “3d007b” と「電話」が持つ “3bdeeb” が共に, 図 5.2のように, 各セットの基準 ID および最適 ID である場合にのみ, 「携帯」の語義 を “3d007b” に「電話」の語義 を “3bdeeb” に 決定する.

このヒューリスティクスでは, 共起する 2 つのタームにおいて, 各タームの持つ概念識別子が互いに最適であると判定されている場合にのみ各クエリータームの語義を決定するものである. なお, ここでは語義の仮決定を行い, 最終的には次のステップで全ての語義が本決定する.

表 5.2: 語義決定の条件 1

| 比較セット名 \ID の区別 | 基準 ID | 最適 ID |
|---|---|---|
| 比較セット$_A$ | 3d007b (携帯) | 3bdeeb (電話) |
| 比較セット$_B$ | 3bdeeb (電話) | 3d007b (携帯) |

※ ( ) 内は 上記の概念識別子を語義としてもつクエリーターム

(e) d で全ての可能な語義の仮決定を行った後, その結果中に矛盾が生じるものがあれば, それらのタームに関するは語義決定は本手法では行えないと判定する. このような状況は, クエリータームが 3 つ以上ある場合に生じる可能性がある. それ以外の場合は, d の仮決定を本決定とする.

(例) クエリー「多角・事業・低迷」の場合

以下の (1) と (2) は d で仮決定されたものだが, 両者による「事業」の語義が異なる ($ID_A$と $ID_B$) ため, (1)(2) の仮決定は共に無効とし, これらの語義の曖昧性解消は不可能として, 曖昧性を残す.

| | | | |
|---|---|---|---|
| (1) | $ID_A$ (事業) | 最適ID $\Longleftrightarrow$ | $ID_I$ (多角) |
| (2) | $ID_B$ (事業) | 最適ID $\Longleftrightarrow$ | $ID_J$ (低迷) |

## 5.4 手法の組み合わせについて

本研究で提案する クエリータームの語義の曖昧性を解消するための 3 つ手法は, 前節で説明した通り, 各々異なる特徴を持つ方法である.

これらの手法では共に「語義を決定できるだけの確実な情報があるとみなされた場合のみ語義を決定する」という立場をとっており, 一つの手法でほぼ全ての語義を決定できる可能性は低いと予想される.

しかし一方で, 各手法の特徴の違いより, 各々の手法により多義性を解消できるクエリータームは異なるという状況も考え得る.

そこで, 本研究ではこれらの 3 手法から得られる正解を分析し, その結果から各手法の適用順序を検討する. そして, 単一の手法よりも高い正解率を得ることを目的として, 最も適切と思われる順序によって 3 手法を組み合わせた曖昧性解消方法の効果についても実験・評価を行う.

# 第 6 章

# 実験と評価

これまで, 2 章で基本的な拡張について述べ, 3 章と 4 章で概念を利用したクエリー拡張手法の提案を行い, 5 章では概念の拡張を行う前に初期クエリータームの語義の曖昧性の解消を行うことが検索精度の向上に有効であるという主張を行ってきた. 本章では, これまで提案してきた手法の有用性を調査するために評価実験について述べる.

以下ではまず, 1 節で評価実験に使用した実験セットや評価方法などについて説明し, その後の 2 節では, 2 章で説明した従来の方法を用いた基本的な拡張結果を用いた実験結果を示す. そして 3 節以降は, 2 節の結果を基準として 3〜5 章で提案した方法による実験の結果と評価を行う.

## 6.1 実験の方法

まず, 本研究では 2 章で述べたように, 各クエリー拡張の手法を実装し, 図 2.1のような構造を持つクエリー拡張システムを試作した.

本システムの入力には, 情報検索システム評価用ベンチマーク “BMIR-J1” で用意されたクエリー[1]を用いる. BMIR-J1 では, 検索対象文書 600 件と, クエリー 60 個およびその正解が与えられ, クエリーは検索の機能に基づき 6 グループに分類され, 各研究課題に応じて使用するクエリーを選択できるようになっている [19][20]. 今回は, そのうちから 以下のような 2 種類のクエリーセットを用意した.

[1] BMIR-J1 の解説書では「検索要求文」と記述されている

**クエリーセット “A”**:
「基本機能」[2]に分類された 10 個中, 8 個のクエリー[3]

**クエリーセット “B”**:
本手法では明らかに正解を導き出すのが困難だと思われる「数値・レンジ機能」[4]に分類された 5 個を除いた 55 個中, 40 個のクエリー[5]

このクエリー拡張システムの基本的な枠組みと拡張に際して行われる処理に関する詳細は 2 章で既に述べた通りである. なお, 各手法により生成されたクエリータームの評価方法は手法毎に各々異なるため, 次節以降でその都度述べる.

一方, 上記のシステムによって生成された拡張クエリーの検索における評価を定量的に行う必要があるため, 本研究では小規模な文書検索システムを試作した.

この文書検索システムでは, 検索対象文書集合として BMIR-J1 で用意された全 600 文書を用いる. また本システムでは, 上記の 2 種類のクエリーセットを入力とし, tfidf 法 [11] によりスコアリングされた検索文書のリストを出力とする. 検索文書に対しては, まず 初期クエリーと同様に形態素解析し, その結果から必要な自立語以外を除去[6]した後, 残ったターム毎に 式 6.1にしたがって, 重要度 $w_{dt}$を計算する.

$$w_dt = \frac{1 + log(tf_{dt})}{log(length(d))} * log\frac{N}{n_t} \tag{6.1}$$

ただし, $w_{dt}$は ターム $t$ に対する文書 $d$ の重要度, $tf_{dt}$は $d$ における $t$ の出現頻度, $length(d)$ は $d$ における文書長 (ターム数), $N$は 総テキスト数, $n_t$は $t$ を含む文書の数である.

以上の処理によって作られたインデックスより, 文書ベクトル $D_i$は, $N$文書からなるデータベースに出現する全 $m$ 個の必要な自立語の重要度を要素とする $m$ 次元のベクトルとして表せる. クエリーベクトル $Q_j$についても同様である. 検索システムが出力する文書のスコアは, クエリーベクトルと文書ベクトルの類似度であり, 式 6.2のような内積計算によって求める.

[2] キーワードの存在確認, あるいは, キーワードのシソーラスによる展開語の存在確認, および, それらの語の存在に関する論理式 (AND や OR など) の充足判定などが可能なクエリー

[3] 解説書において, 正解文書数が適切な範囲内にあり, 使用を推奨されているもの

[4] 数の数え上げや, 数値などの範囲の正しい解釈, 数値の大小比較や単位の理解・変換が可能なクエリー

[5] 脚注 3と同様の理由による

[6] { 普通・固有・時相・サ変 } 名詞, 動詞, 形容詞以外を除去した後, さらに ストップワードを除去する

$$sim(D_i, Q_j) = \sum_{k=1}^{t} w_{ik} w_{jk} \tag{6.2}$$

また, 検索システムの出力に対する評価は 式6.3, 式6.4に定義した recall と precision を用いて行う.

$$recall(\%) = \frac{\text{システムの出力における正解文書数}}{\text{データベース中の全正解文書数}} * 100 \tag{6.3}$$

$$precision(\%) = \frac{\text{システムの出力における正解文書数}}{\text{システムが出力した文書数}} * 100 \tag{6.4}$$

## 6.2 従来の拡張方法についての実験〜同義語による拡張〜

本節では, 次節より各提案手法の実験と評価を始める前段階として, 従来法の 1 つである同義語による拡張クエリーを用いた検索実験を行い, 各手法の比較対象である同手法の精度を示す.

なお, その際には, 以下の閾値実験を行い, その結果より各々の最適閾値を決定し, 以降の節ではその閾値設定を default 設定として 実験を行っていくものとする.

1. 初期クエリーターム と 同義語クエリータームに対する重み付け実験
2. 対象行数を考慮した AND 項と文書とのマッチングの実験

これらの実験では, 2 章で述べた手順で 各初期クエリータームの同義語は獲得できているものとし, その同義語クエリーをそのまま検索システムの入力とした場合の検索精度, そしてさらに 上記の 1, 2 で手を加えた同義語クエリータームによる検索精度を実験により調査・考察していく.

1. **初期クエリーターム と 同義語クエリータームに対する重み付け実験**

   ここでは, 初期クエリーターム と 同義語クエリータームの重みを $n : 1$ $(1 \leq n \leq 10)$[7]とした場合の検索精度の変化について調査した.

[7]なお, 予備実験により 初期クエリータームより同義語クエリータームに高い重みをつけることによって検索精度の向上は見られなかったため, $n$ の値をこのように設定した.

その結果を表6.1に示す. この表より, 同義度のみの場合の最適重み比率は 3:1 の場合であることがわかる. また, 全く重みを付けなかった場合が最も精度が悪く, クエリータームの重み付けによって検索精度の向上が得られることがわかった.

表 6.1: クエリーの重み付けにおける検索精度の比較

| $n$ | BreakEvenPoint |
|---|---|
| 1:1 | 28.42 |
| 2:1 | 34.22 |
| 3:1 | 38.58 |
| 4:1 | 38.23 |
| 5:1 | 37.33 |
| 6:1 | 35.24 |
| 7:1 | 34.01 |
| 8:1 | 34.03 |
| 9:1 | 32.40 |
| 10:1 | 31.35 |

2. **ターム間距離を考慮した AND 項と文書とのマッチングの実験**

本研究で実装した文書検索システムでは, クエリーに AND 演算子や OR 演算子が使用出来る. しかし, AND 項であるクエリー中に出現する各タームは, 通常設定では, 同一文書の中に存在していれば どんなに離れた部分に存在していても正解文書として獲得される.

そこでここでは, 「AND 項中のタームとタームの間隔が $n$ 文であるような文書のみ, それを正解として獲得する」という条件を付けて実験を行った. その結果が 表6.2である.

この結果から, まず $n = 2$ の時, つまり ターム間距離が 2 文以内である場合に最もよい精度が得られていることがわかる. ただし, 個々のクエリーの結果を分析してみると, 通常設定の全文書にした場合が最も精度がよかったクエリーもあった.

表 6.2: ターム間距離による検索精度の比較

| $n$ | BreakEvenPoint |
|---|---|
| 1 | 33.49 |
| 2 | 35.38 |
| 3 | 31.22 |
| 4 | 29.98 |
| 5 | 30.01 |
| 6 | 28.33 |
| 7 | 26.24 |
| 8 | 26.10 |
| 9 | 25.50 |
| 10 | 27.74 |
| max | 30.58 |

よって, 以降の章では, 初期クエリー と 同義語クエリーの重み比率は 3:1, AND 項については, その文書における全タームの出現間が 2 文以内である場合のみを正解とした結果だけを示す.

## 6.3 名詞・動詞間の概念関係を利用したクエリーの拡張

3 章では, 従来の拡張方法における検索精度を改善するため, 名詞・動詞間の概念関係を利用して得た関連語を加えた拡張クエリーを生成する方法を提案した.

本節では, 本手法による検索精度への影響の調査を目的として, まず 以下の実験 (1) と実験 (2) という 2 通りの評価実験を行った.

**実験 (1)** 概念記述辞書のレコードを全て用いた場合

**実験 (2)** 概念記述辞書のレコードをフィルタリングして用いた場合

以下では この 2 つの実験について, 各結果とそれに対する考察を順に述べていく. また, これらの考察を受けて実験 (3) も行っているが, これについては実験 (2) の考察においてその結果を示す.

### 6.3.1 実験 (1)

#### 実験の目的と結果

本手法では, 概念記述辞書のレコードを利用して, クエリータームと最もらしい共起を為す動詞的もしくは名詞的な概念識別子を関連概念として獲得し, そこから得られる関連語と従来の拡張手法で得られる拡張タームを “AND” で繋いだものを拡張タームとして生成する.

このように, 従来の方法による拡張タームに本手法による関連語を加えることによって, 正解文書のスコアを増加させ, 検索システムの出力するランキングにおける順位を上げることが本手法の目的である.

そこで, 実験 (1) ではまず本手法の検索精度への影響について調査する第一段階として, 概念記述辞書のレコードを全て利用して関連語を生成した場合における検索精度の変化を調査するための評価実験を行った.

以下, クエリーセット “A” を基にした拡張を行った場合の結果を図 6.1に, クエリーセット “B” の場合の結果を図 6.2に示す. なお, これらのデータは全クエリーの平均値を plot したものである.

#### 考察

図 6.1, 図 6.2より, 結果として 両クエリーセット共に, 本手法によって precision 値が減少してしまうことがわかった. クエリーセット “A” で最大 9%, クエリーセット “B” で 最大 4.5%減少している. そこで, 個々のクエリーについての検索結果を分析し, 原因を調べた. その考察を以下に述べる.

なお, 3 章で説明した通り, ここで検索システムの入力とした拡張クエリーセットには, 同義語による拡張タームも含まれているため, recall の最大値は変化しない.

#### 原因 (1) 検索文書集合における頻出タームによる影響

precision 値が下がった最も大きな理由として, 「メーカー」, 「経営」, 「会社」, 「日本」といった検索文書の大部分に出現するタームによる影響が挙げられる. これらのタームは, 検索文書全 600 件のうち, 約 80–90%の文書中に出現している.

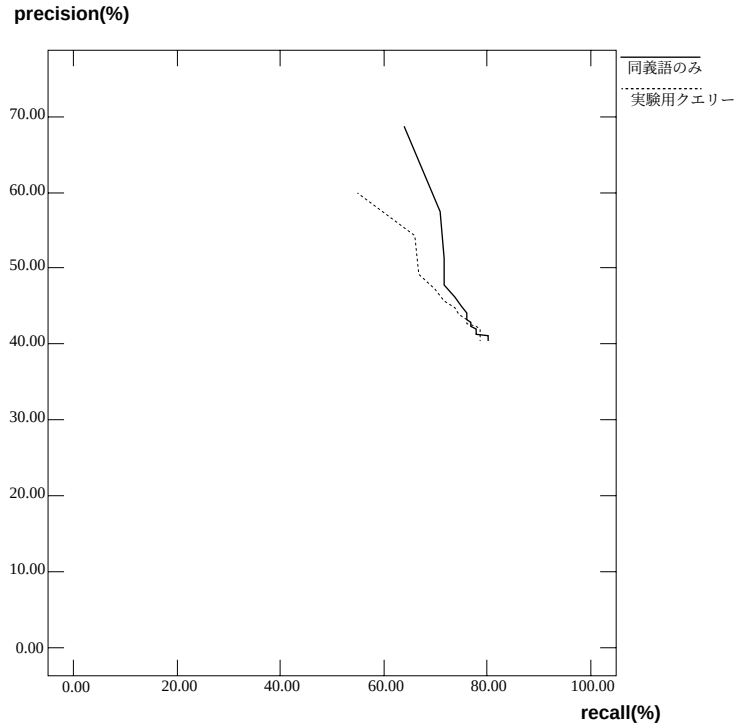



図 6.1: 実験 (1) の結果 ( クエリーセット “A”)

本手法では, 単にこれらのタームそのものをクエリータームとするのではなく,「メーカー AND 販売する」のような AND 項を生成しており, それによってマッチングにおける制約も強まり precision 値も上がると考えられた. が, 上記のような頻出ターム (つまり, Idf 値の低いターム) においては「メーカー AND 販売する」のような 2 タームの共起も比較的多くの文書において出現するため, システムは 不正解文書を導いてしまったものと考えられる. precision 値の減少が著しかったほぼ全てのクエリーにおいて, このような頻出タームが見られた.

このような頻出タームの特徴は, 検索文書の特徴に依存する.
今回の場合 使用した BMIR-J1 が日本経済新聞の記事を検索文書としているため, 上記に挙げたようなタームが結果的に不正解文書を導いていることがわかった. そこで, 検索文書集合における情報は, あらかじめ検索材料として利用できるものであるため, この問題については, 検索文書集合における出現タームのふるまいからその特徴を抽出し, クエリー拡張に活かすという解決策が効果的であると考えられる. そ

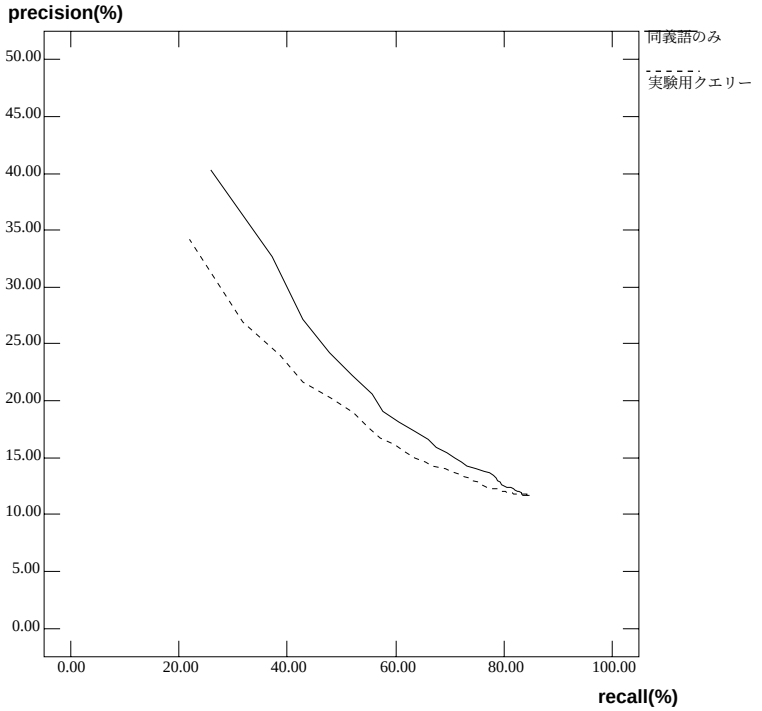



図 6.2: 実験 (1) の結果 ( クエリーセット “B”)

の方法の 1 つとしては, 各タームの Idf 値を計算し, Idf 値が低いタームについては拡張してもあまり効果が得られないとして低い重み付けを, 逆に Idf 値の高いタームについては高い重み付けを行う方法が挙げられる.

**原因 2** クエリー中のタームの係り受け関係を考慮していない

今回利用したクエリーの中で, クエリーセット “B” には, 動詞的概念を持つと判定され, 名詞的概念を関連概念として獲得したタームを含むものが 16 クエリーあり, いずれのクエリーにおいても, 係り受け関係を考えた場合, 動詞的概念を持つターム (係り側) に対応する受け側の名詞がクエリー中に存在するものがほとんどだった.

例えば, クエリー「コンピュータメーカーの人員削減」を形態素解析すると, 「コンピュータ」,「メーカー」,「人員」,「削減」の 4 タームが得られるが, そのうち, 動詞的概念を持つタームが「削減」, その受け側にあたる名詞が「人員」ということになる[8].

[8] その他, 「日本の製造業における生産性向上またはコストダウンの事例」,「業績悪化を原因とする企業

しかし, 本研究では単にクエリーを形態素解析し, 必要な自立語のみを獲得するのみで, 単語間の係り受け関係の解析は行っていない. よって, 上記の例で「削減 (する)」という動詞的概念について, 「人員 (を)」という受け側要素が存在するにも関わらず, 全くそれとは関連のない「経費」「エネルギー」を表す名詞的概念を関連概念として獲得してしまっている. つまり, 獲得される概念の属すると思われる分類 (この場合では, 「人間」の数) が分散してしまい, 非常に多くの関係のない分類 (「静物」の数) について記述した文書を獲得してしまうのである. precision 値が下がったのは, “原因 1” に並び, この理由によるところも非常に大きいと考えられる.

また, 名詞的概念から 動詞的概念を得る場合にも同様の現象は起きているが, 動詞的概念に基づく拡張による場合の方が非常に多くの不正解文書を導いていることがわかっている.

この問題については, 単純に動詞的概念からは拡張を行う際には, その他のクエリーターム全てについて共起しやすい名詞かどうかのチェックを加えるという解決法もあるが, 今後は 係り受け関係を利用していく必要があるだろうと考える.

ただ, 実際には 検索システムに入力されるクエリーとしての自然言語文は 通常の文とは異なり, 動詞的概念を持つ単語が名詞化されて入力される傾向があると考えられるため, 特に 名詞化された動詞等にも対応できるような係り受け解析が必要となるだろう. またそれと同時に, 検索システムに入力される自然言語文のクエリーの特徴の分析を行っていくことが必須であると考える.

**原因 3 クエリー全体の意図を反映できていない**

本手法では, クエリーに含まれる複数のタームのうち, 1 タームのみに注目し, そのタームと最もらしい共起をする名詞的もしくは動詞的概念を関連概念として獲得している. しかし, ユーザの意図はクエリー全体に込められているものであり, 単にそのうちの 1 つのタームを拡張しても, クエリー全体の意図には一致しない文書 (つまり 不正解文書) を獲得してしまう恐れがある.

このことは, 単に本手法だけでなく 同義語のみの拡張にも言えることではあるが, 本手法においては「同義語のみによって獲得している正解文書のうち, 真の正解文書のスコアを上げる」ことが目的であり, 現状では “ 同義語のみによるクエリー” の

の合併の事例」など

良い部分も悪い部分もそのまま拡張していることになるとも考えられる.

今回の分析結果でも,「国内」,「航空」,「大手」の 3 タームからなるクエリーにおいて, 本手法による拡張クエリータームが, 単に 大手会社 (大手メーカー) についての記事を獲得してしまっているケース等が多く見られた. この問題は, 単に クエリー中の 1 タームとその関連語だけでは, クエリーの意図を表現しきれないため発生すると考えられ, そのための対策としては, 現在生成している「クエリーターム A とクエリーターム A の関連語」に, さらにその他のクエリータームも共起しているような文書を獲得する方法がある.

今回この方法については,

- **b タイプ** [ クエリーターム A [9] ] AND [ クエリーターム A の関連語 ] AND [ クエリーターム B ] (ただし A $\neq$ B)
- **c タイプ** [ 全てのクエリーターム ] AND [ クエリーターム A の関連語 ]

という 2 つのタイプのクエリータームを本手法により生成し, 実験を試みている. なお, この実験 (3) の結果と考察については, 6.3.3で述べる.

### 6.3.2 実験 (2)

#### 実験の目的と結果

実験 (2) では, 3.3.3 で説明した概念識別子のフィルタリングについての実験の目的と結果を述べる.

この実験には, ある初期クエリータームが持つ概念識別子から関連概念を得る場合に,「どのような名詞的 ( もしくは 動詞的) 概念とも共起するような概念識別子は 関連概念として拡張しない」ようにした場合, 検索精度がどう変化するのかを調査する目的がある.

今回 共起する概念識別子数の逆数を重みとし, その重みに 9 通りの閾値を与えて, 実験を行った. その結果 ( クエリーセット “B” の場合) 獲得された概念識別子数 (1 クエリータームあたり) と BreakEven Point の値を 表 6.3に示す. なお, 概念記述辞書における各概念識別子 (名詞的概念識別子, 動詞的概念識別子) の異なり数と, それら各概念識別子と同辞書において共起する概念識別子の最大数を 表 6.4に示す.

[9] ここで,「クエリーターム」とは,「初期クエリーターム もしくは その同義語」 を指すこととする.

表6.3では, 各閾値における検索精度から, 横軸に評価対象とする文書数[10]として recall と precision のグラフを描き, その2本のグラフの交点となる **BreakEven Point** の座標を比較している. なお, 表6.3の "$n$" とは, 初期クエリータームの持つ概念識別子 に対し, 概念記述辞書において共起する概念識別子数(異なり数) を指し, それに対応する BreakEven Point は, $n$ が左記の値域内にある概念識別子のみを検索に用いた場合の BreakEven Point(recall, precision) を示す.

表 6.3: 概念識別子のフィルタリングによる検索精度の比較

| $n$ | 獲得される概念数 (1 タームあたり) | BreakEvenPoint (%) |
|---|---|---|
| 1∼ 10 | 19.72 | 16.48 |
| 1∼ 30 | 30.19 | 20.29 |
| 1∼ 50 | 50.65 | 30.22 |
| 1∼ 100 | 79.71 | 32.78 |
| 1∼ 200 | 102.24 | 31.61 |
| 1∼ 400 | 121.46 | 31.20 |
| 1∼ 600 | 131.01 | 30.45 |
| 1∼ 800 | 134.43 | 28.50 |
| 1∼1000 | 137.04 | 29.74 |
| 1∼ max | 143.12 | 29.95 |

表 6.4: 概念記述辞書における { 名詞的・動詞的 } 概念識別子に関するデータ

| 概念識別子の区別 | 異なり数 | 共起する概念数の最大値 |
|---|---|---|
| 名詞的概念識別子 | 51965 | 2203 |
| 動詞的概念識別子 | 26577 | 9581 |

[10] 2 章参照

**考察**

1. **共起概念数が 1000 以上の概念識別子における効果**
   概念記述辞書において共起する概念識別子数が 1000 以上の概念識別子は, 動詞的概念では「なる」「行う」「いる」等, 名詞的概念では「もの」「こと」「日本」「米国」「人」「人々」「会社」「個」等を表す概念であった. この例でわかるように, これらは一般にはストップワードとして扱われるものがほとんどであった. しかし, これらの概念を持つタームについては, 本研究では 独自に作成したストップワード・リストに既に登録されているものが多く, これらは事前にクエリーとして採用しないよう処理されているため, 実際に $1 \leq n \leq max$ つまり 全て概念識別子を使った場合と, $1 \leq n \leq 1000$ の場合の検索精度の差はあまり生じていない.

2. **最適閾値について**
   表 6.3から, 共起する概念識別子数が $1 \sim 100$ である概念識別子を用いた場合, 最も検索精度の平均値が高いことがわかった.

   この時の検索精度と 全概念識別子を拡張に利用した場合の検索精度の比較した結果 (クエリーセット “B”) を 図 6.3に示す.

   図 6.3と表 6.3より, 概念識別子のフィルタリングにおいて, 共起概念数が 100 以下の概念識別子を使うことによって 1 タームあたり 平均 79.71 個の関連概念を獲得し, それによって, 全概念識別子を利用した拡張の場合に比べ precision 値が 最高で 4.53% 向上させる効果を持つことがわかった.

   このような全概念識別子を利用した場合との検索精度の比較によって, このフィルタリングの実験の前提となっている「どんな概念とも共起しやすい概念は検索においてノイズとなるであろう」という仮説は, 本実験セットにおいては正しかったと言えよう.

   ただし, 先に考察を行った 共起頻度数が 1000 以上の概念識別子に対し, 1000 未満の共起頻度を持つ概念識別子については, 人間の目から見たところその差異を明示することは難しく, また 個々のクエリーの結果を見てみると, フィルタリングをしない場合の方が precision 値が高かったケースも見られた. よって, 今回は 大規模な日本語文書の実験セットがなかったために, closed test しか行えていないが, 本フィルタリング手法の有効性を正しく評価するためには, 大規模な実験セットにおける

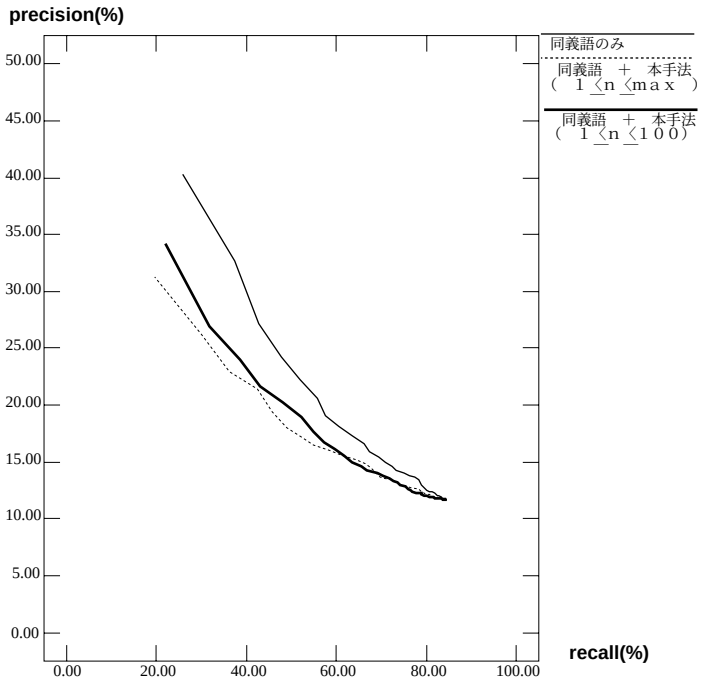



図 6.3: 概念識別子のフィルタリングによる検索精度の比較

open test を使った実験を行う必要があると考えている.

また, フィルタリングにおける効果は上がったものの, 依然同義語のみの場合を上回る精度は得られていない. この原因は, フィルタリングによって効率的に関連概念を得ることができるようになったものの, 本手法によって精度が下がる原因は, 根本的に 6.3.1 で述べた 3 つの原因であり, これらがフィルタリングによっては, 解決されていないためであると考えられる. そこで, 次の実験 (3) では 6.3.1 であげた “原因 3” について考慮した実験を試みる.

なお, 以下に, 共起頻度数がこの値域にある時獲得できたタームとクエリーの例を挙げておく.

| 初期クエリーターム or その同義語 | 獲得した関連語 |
|---|---|
| 冷夏 | 厳しい, 緩和する, 規制 |
| 被害 | 重大だ, 広がる, 賠償する, 発生する |
| 損傷 | 激しい |
| 計画 | あきらめる, 完了する, 引き継ぐ, 難航する |

### 6.3.3 実験 (3)

#### 実験の目的と結果

ここでは, 6.3.1 で述べた “原因 3” について考慮したクエリーを生成した上で, 実験 (2) で得られた最適閾値に基づくフィルタリングを行った場合の検索精度を調査する.

実験 (3) では, まず 実験 (2) より 共起概念数が $1 \sim 100$ であるような概念のみを拡張に用いて, 関連語を生成する. そして, 得られた関連語をもとに, 6.3.1 で述べた b タイプと c タイプという 2 種類の クエリータームを生成する. これら “b タイプ” と “c タイプ” を検索システムの入力とした場合の結果のうち, “b タイプ” の結果を 図 6.4に示す. なお, 図中で (a タイプ) と書かれたクエリーのグラフは, 実験 (2) で示した $n$ の値域が $1 \sim 100$ である場合の結果と同じものである.

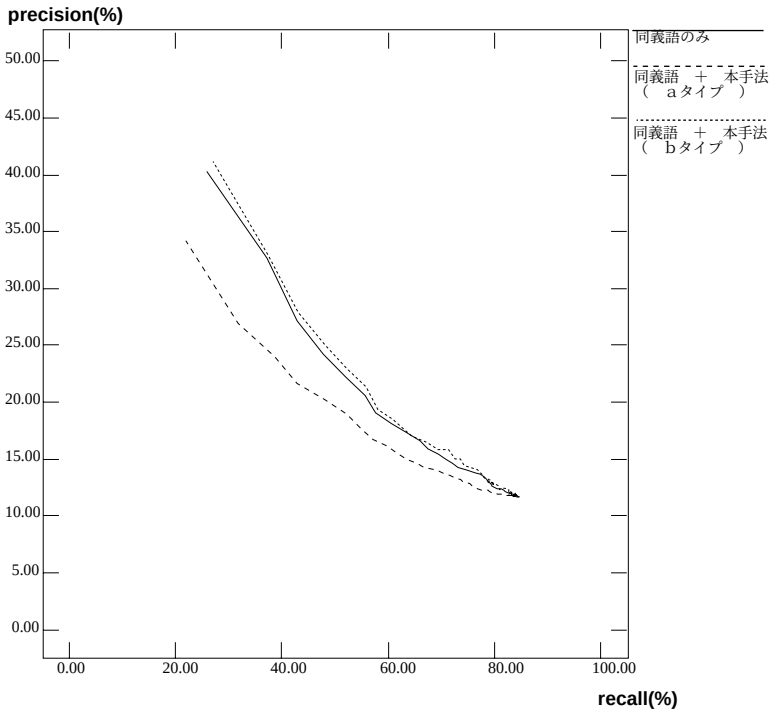



図 6.4: 実験 (3) の結果 ( クエリーセット “B”)

### precision 値に関する考察

図 6.4より, b タイプのクエリータームを生成した場合, 同義語のみを用いた拡張に比べ, precision 値の向上が得られた. ただし, その差は わずか 1.3%程である.

まず, “a タイプ” のクエリー および 同義語のみからなるクエリーと比較し precision 値が向上した理由は, 単に 初期クエリータームのうちの 1 つに対する拡張しか行わないのではなく, その他の初期クエリータームを加えてやることで, クエリーターム全体が表す意味に近いものだけを獲得できているからであると考える.

例えば, クエリー「冷夏による被害」において, 単に「被害 AND 広がる」(“a タイプ”) だけでは様々な「被害」に関する文書が正解として獲得されてしまうが, 「冷夏 AND 被害 AND 広がる」(“b” タイプ) というクエリータームを生成することによって, その中の「冷夏」に関連する被害について書いた記事のみが獲得される訳である.

このようにして, “b タイプ” のクエリータームは, 同義語のみによって獲得した文書のうちから, 一部の正解文書のスコアを上げる効果を出し, その結果わずかではあったが, precision 値が向上したと考えられる.

なお, “c タイプ” のクエリータームについては, “a タイプ” の場合よりは精度が向上したものの, 同義語のみによる場合の結果と同様の結果であった. これは検索結果を分析した結果, 「初期クエリーに含まれる全てのターム (もしくは その同義語)」という制約が強すぎて, 本拡張によるクエリータームにはどの文書もヒットしなかったためであるとわかった.

### 6.3.4 まとめ

実験 (1) と実験 (3) より, 名詞と動詞の意味的に最もらしい共起による拡張は, 検索の上で わずかながら精度の向上に役立つ情報であることがわかった.

また実験 (2) の結果より, 共起しやすい名詞または動詞は拡張に用いてもあまり効果は得られず, ある程度特徴的な共起語のみに拡張の範囲をしぼることによって, 検索の精度は向上することがわかった.

今回 本研究では EDR 辞書における概念共起データを用いたが, 本手法は 大量の名詞と動詞の共起データによっても同様に実現が可能であると考えられる.

なお, 本手法は これまで特に注目されてこなかった動詞情報の利用による検索精度向上の一手法を提案したが, その手法自体は非常に simple なものであった. しかし, 今回注目した名詞と動詞の共起関係は, 文の構造に大きく関わるものであり, 今後は クエリーそのものや, 生成された拡張クエリーと検索文書とのマッチング等においても文の構造に考慮していくことが必要であり, それによってこの情報を使った精度の向上が期待できるのではないかと考える.

## 6.4 概念説明文を利用したクエリーの拡張

この節では, 4 章で提案した手法について, (1) 本手法においてどの程度の関連語獲得が行えるのか, (2) 獲得できた関連語を使った場合の検索精度の向上にどれくらい貢献するのかについて調査するための実験とその考察を行う.

### 6.4.1 実験

実験 (1) では, 4 章で説明した方法で拡張を行うよう設定したクエリー拡張システムにおいて, クエリーセット “A” および クエリーセット “B” を入力とした場合, どのような

関連語が獲得できるかを調べた.

また, 実験 (2) では, 実験 (1) によって獲得された関連語を 初期クエリーセットに加えた場合の検索システムの出力結果を, recall・precision を用いて評価した.

すなわち, 実験 (1) では本手法の適用可能率 (coverage) を, 実験 (2) では 本手法を適用した場合の検索における正解率 (accuracy) を調査する.

以下, 実験 (1) の結果を表 6.5と表 6.6に, に, 実験 (2) の結果を図 6.5に示す.
なお表 6.5で評価基準に用いている coverage は式 6.5による.

$$coverage(\%) = \frac{\text{関連語を獲得できたクエリー数}}{\text{全クエリー数}} * 100 \tag{6.5}$$

表 6.5: 実験 (1) の結果

| | クエリーセット “A” | クエリーセット “B” |
|---|---|---|
| 関連語を獲得できたクエリー数 | 3 | 22 |
| 全クエリー数 | 8 | 40 |
| coverage(%) | 37.50 | 55.00 |

表 6.6: 実験 (1) の結果

| | クエリーセット “A” | クエリーセット “B” |
|---|---|---|
| 平均獲得概念数 (異なり数) | 1.66 | 9.77 |
| 平均獲得ターム数 (異なり数) | 3.75 | 11.91 |

### 6.4.2 実験 (1) の評価・考察

#### 獲得された関連語について

本手法における coverage は, クエリーセット “A” で 37.50%, クエリーセット “B”

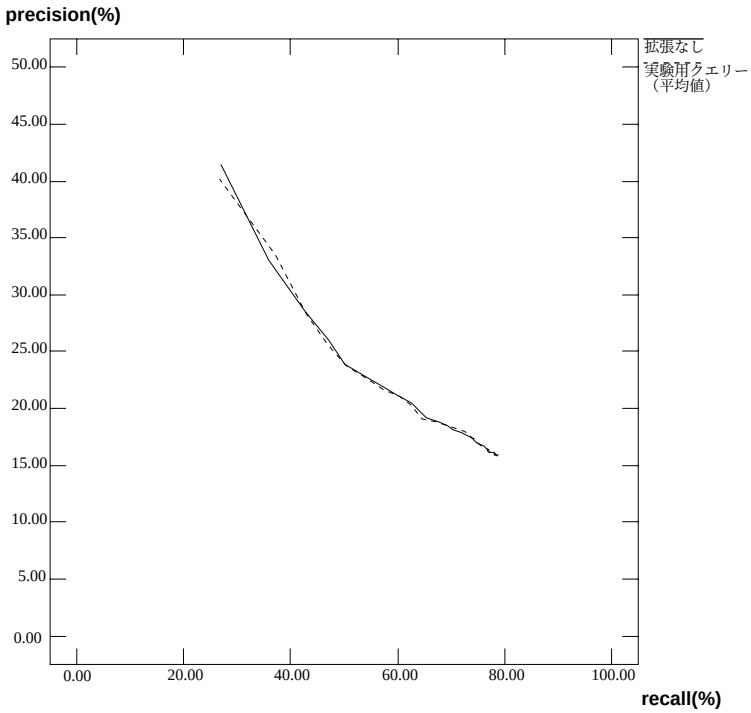



図 6.5: 実験 (2) の結果 ( クエリーセット “B”)

で 55.00% と比較的低かったものの (表 6.5), 関連語を獲得できた場合には, クエリーセット “B” において 平均 9 種の関連概念, 関連語にして約 12 種のタームが獲得できることがわかった (表 6.6).

まず, 今回本手法により獲得できた関連語の例を表 6.7 上段に示す. その下段に添付したのは, 例にあげた概念説明文中に出現する各クエリータームの同義語, 上位概念語, 下位概念語の抜粋である[11].

表 6.7の上段より, 例 1 のように クエリータームと 概念の意味するものの主体は同様である ( この場合は「企業」のことを指す) が, クエリータームの持つ概念をさらに細かく分類した概念や, また例 2,3 のように, クエリータームと概念の意味するものの主体は全く異なるが, 機械的に獲得することが難しいと思われるようなクエリータームに関連の深い概念が本手法によって獲得できていることがわかる. 例 4

[11] 表中に記した単語は 同義, 上位, 下位の概念識別子が持つ単語見出しである (括弧付きの場合は, 概念見出し).
また, “-” は該当する概念識別子が存在しなかったことを示す

は, 「人員」という制約が伴った場合のターム「削減」が表す動詞的な意味あいを本手法によりうまく捉えて関連概念を獲得した例である.

また, 表6.7下段にあげたような従来のシソーラスを用いた拡張方法で得られる関連語と比較することによって, 本手法が これまで得られなかったような関連語の獲得を実現していることは明らかだろう.

表 6.7: 獲得できた関連語の例

| 例 | 概念説明文 | 獲得できたターム |
|---|---|---|
| 1 | 価格水準を決定したり, 価格を操作したりする力のある企業 | プライスリーダー |
| 2 | 国内の空港を結ぶ航空路線 | 国内線 |
| 3 | 価格に関する企業間協定 | 価格カルテル |
| 4 | (企業が機械化や不況の影響などで)人員を削減する | 合理化 |

| ターム | 同義語 | 上位概念語 | 下位概念語 |
|---|---|---|---|
| 価格 | コスト, 値, 値段 | (金額) | 買い値, 課税価格, 魚価, 減価, 現価, 米相場, 糸価 |
| 企業 | コーポレーション | (経済組織) | - |
| 国内 | - | (領地) | - |
| 航空 | 飛行する, 翔ける | (飛ぶ) | - |
| 人員 | 頭数 | (人間) | - |
| 削減 | カットする, 削る | (減らす) | 軽減する, 節略する, 減員する |

**coverage の改善について**

最初に述べたように, 本手法は coverage が低いという欠点を持つ. 今回, このように適用率が低かった一因として, サンプル数が少なかったという点があげられる.

しかし, 本手法では現在, クエリータームと その同義語のみを 概念説明文の検索に使用しており, 根本的な問題として, このような単純な拡張ではカバーできない概念説明文がまだ多く存在するという点があげられ, 本手法をこうした問題に対応できるよう拡張していく必要があると考えられる.

そこで以下では, 現手法で関連する概念説明文の獲得に失敗している例をあげなが

ら, その対策について述べる.

対策 1

「冷夏・被害」というクエリー[12]に対し, 本手法では 「冷夏・被害」以外に「被害」の同義語を用いて「冷夏・損傷」「冷夏・ダメージ」といったタームのペアを生成し, 概念説明文の検索を行っている. しかし, 実際に本手法で獲得したい概念説明文に非常に近いと思われる「夏に気温が低かったり日照不足などで農作物が不作になる被害」という概念説明文は, 上記のようなタームのペアからは検索できない.

この他にもこのようなケースで概念説明文の獲得に失敗した例が 2-3 あり, この問題については今後 クエリータームそのものの概念説明文中のタームを利用する[13]などの工夫をすることによって, ある程度補完できるのではないかと考える.

対策 2

本手法では, 「菓子・メーカー」というクエリーから「菓子・メーカ」,「菓子・会社」といったタームのペアが生成されるため, 実験前の時点では, 関連語として「製菓」「製菓業」また「菓子店」などといった語が獲得できるのではないかと予想された.

しかし, 実験結果および関連レコードの分析の結果, 実際は「製菓」の概念説明文は「販売するために菓子を作ること」であること, また「製菓業」という見出し語は辞書中に存在せず, 「菓子製造業」という概念説明文を持つ「コンフェクショナリ」という見出し語ならば存在することがわかった.

つまり, 本手法のように単に同義語を用いるだけでなく, この場合であれば「メーカー」と「菓子」が取り得る共通の用言「製造 (する)」や「販売する」というタームを用いることによって, さらに有用な関連語が獲得できる可能性があると考えられる. これには 現在 EDR の概念記述辞書が利用できると考えている.

対策 3

上記の例以外の場合, 実際に生成したタームのペアが意味するような概念は細

[12] もとは 「冷夏による被害」というクエリー.

[13]「冷夏」の概念説明文 ：「気温が低い夏」

かすぎてEDR中に存在しない[14] 場合, もしくは 生成されるタームのペアから関連概念を得ることが辞書のレベルでは不可能であるかのどちらかである場合が多かった.

例えば, 「不況におけるディスカウンターの台頭」というクエリーからは, 「不況・ディスカウンター」といったタームのペアが生成されるが, 「不況」と「ディスカウンター」に関連性を見出せるのは近年の経済動向に関する知識が必要であり, いわゆる辞書のレベルではこのような流動的な知識に対応することはほぼ不可能である.

この点は, 既存のデーターベースを利用した手法における最も不利な部分であり, 複数のターム間の情報を用いてこのようなクエリーの関連語を得る場合には, おそらく検索文書中の情報を利用するのが有効であると考える.

### 6.4.3 実験 (2) の評価・考察

図 6.5では, 本手法の有効性のみを評価するために, クエリーを拡張しない場合 (初期クエリーのみを使用) と初期クエリーに本手法によって得られた関連タームを加えた「実験用クエリー」を用いた場合の比較を行った. なお, 図 6.5は 全クエリーの平均値を plot したものである.

まず, 本手法により システムによって得られる最大の recall 値がわずかに 0.35%だけ伸びている. precision 値については, recall 値が 30–40%の範囲以外は全般的に 初期クエリーとほぼ同等かそれ以上の値が得られている. この結果からは, 本手法によって検索精度を特に大きく左右するような効果は生じない傾向にあることがわかる.

しかし, 本手法を適用できたクエリー数は全体の約半数であったことから, さらに 適用できたクエリーに焦点をあてた本手法による検索精度への影響を調査する必要がある. そこで以下では, 本手法を適用できた 22 個のクエリーのみに対し, 本手法で得た関連語によって検索された文書についての分析とそれに対する定性的な評価を行う.

1. 精度に変化が生じなかったクエリーに関する考察

   拡張なしの場合と比較し 検索精度に変化がなかったクエリーが, 全 22 クエリー中 10

[14] なお, EDR には, 部分的ではあるが, 概念が非常に細かく分類さている部分とそうでない部分の差が存在するという問題があると考えている.

クエリーあった. このようなクエリーでは, 関連語は生成できたものの, それらの語を含むような文書がデータベース中になかったため, 精度の向上が見られなかった.

ただし, このタイプのクエリーから得られた関連語は通常使用される語も多く含まれ, よって獲得された関連語の特殊性からこのような結果が得られた訳ではないと考えられる.

[**例**] クエリー「携帯 AND 電話」 → 関連語 「留守番電話」,「データホン」など.

なお, クエリーセット "A" のクエリーは全てこのタイプのものであったため, 検索精度が拡張前と全く変わらず, よって図 6.5のような結果は示していない[15].

2. 本手法による関連語 と recall 値の向上に関する考察

検索結果の分析により, 全体の傾向として, 初期クエリーでは獲得不可能な正解文書を導くような関連語 (recall 値を上昇させる効果がある関連語) は, 本手法によってはあまり得られないということがわかった (recall 値の向上が見られたのは 2 クエリーのみだった).

今回の実験はサンプル数が少ないため, その結果だけでこれが本手法によって確実に生じる傾向であるとは断言できないが, 今回の結果を分析した範囲では, 獲得された関連語が複合語であり, かつ その関連語中の単語の 1 つが初期クエリータームと同一である場合が多かったことが原因となって, この傾向が生じていると考えられる.

本システムでは, クエリーは 一度形態素解析器にかけられる. もし 2 つ以上の形態素が得られた場合, システムは得られた全ての形態素から "AND" 項を生成し, それをクエリーの 1 つとして使用する.

**例** : 「留守番電話」 ⇒ 「留守番 AND 電話」
( ↑ 「持つ・電話」から得られた関連語 )

よって, 関連語「留守番 AND 電話」が獲得できても, このクエリーによって得られる正解文書は, 既に初期クエリーターム「電話」でも検索できており, これによって新たな正解文書を獲得することはできない. つまり, このような関連語では, recall

[15] なお, このうちクエリー「国内・航空・大手」については, 正解文書 a を導く関連語「国内線」が獲得できたが, この正解文書 a は, 初期クエリー中のタームによっても検索可能な文書であり, この場合「国内線」からは 正解文書 a の 文書スコアをあげるという効果が得られただけで, 特にそれによるランキングへの影響がなかったため, 精度に変化が見られなかった.

値は高くならないのである.

しかし, このようなケースとは異なる場合, 本手法によって得られた関連語は, recall 値の向上に有効に働く.

例えば, 今回の実験では, クエリー「コンピューターメーカーの人員削減」において「人員」と「削減」の 2 タームより 関連語「合理化」が獲得できた.

この関連語「合理化」を含む実験用クエリー (NO.19) における拡張なしの場合との比較を行った結果, 拡張なしの場合の recall 値を最高で 10.00%上回る精度が得られた (図 6.6). この recall 値の向上は, 本手法によってより多くの正解文書が検索できたことを意味する.

なお, その他の部分で拡張なしを下回っているのは, 「合理化」により検索された一部の不正解文書による影響である.

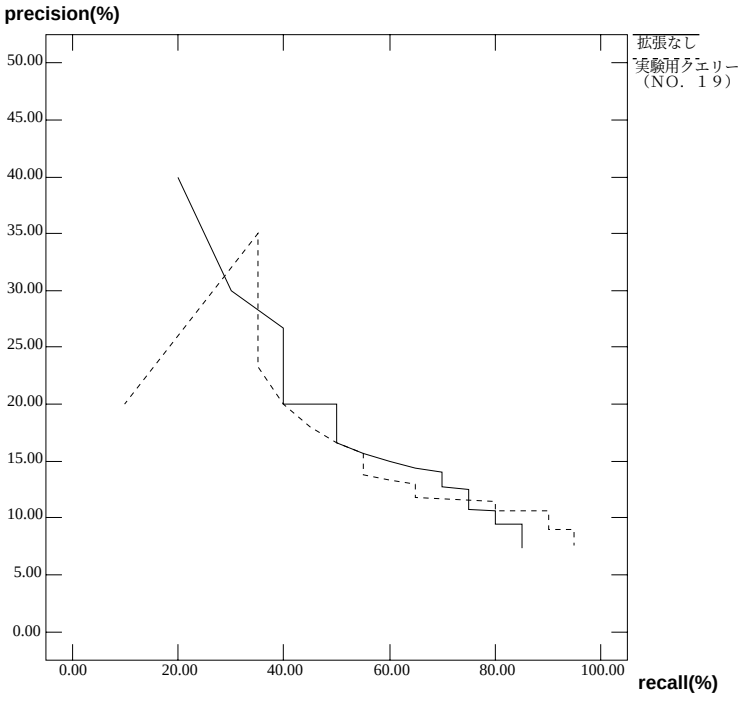



図 6.6: クエリー NO.19 「コンピューターメーカーの人員削減」の結果

3. 関連語による precision 値の変化に対する考察

本手法により あらゆる recall 値において, precision 値を大きく上昇させるような効果が得られたクエリーはほとんど存在しなかった. また, 全22 クエリーから上記の 1 のタイプのクエリーを除いた 12 のクエリーのうち, precision 値がほぼ全ての recall 値において拡張なしを下回っていたのが 3 クエリーで, それ以外の 9 クエリーでは, いくつかの recall 値において拡張なしを上回る precision 値を得ていた.

成功例

本手法によって precision 値の向上に成功した例として, クエリー NO.23 「円高による物価の低下」の結果を図 6.7に示す.

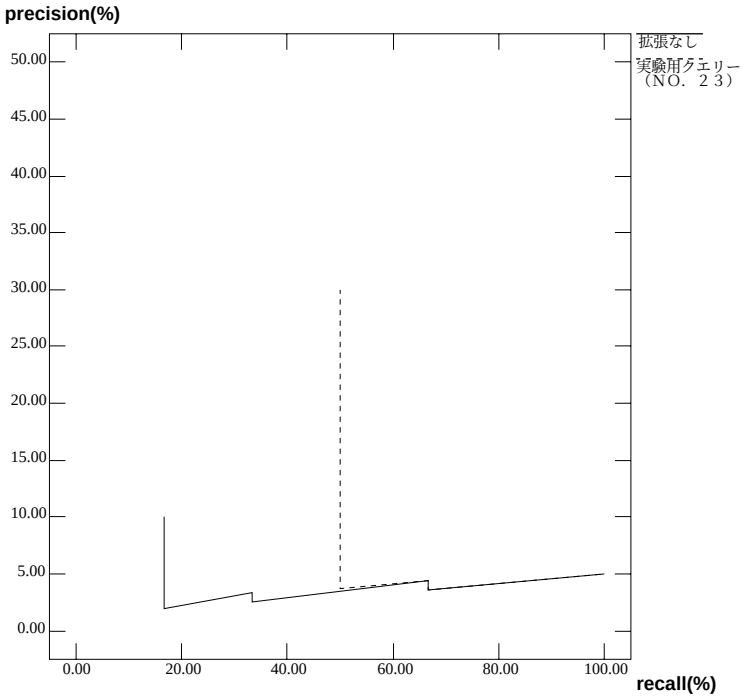



図 6.7: クエリー NO.23 「円高による物価の低下」の結果

形態素解析によって得られた「円高・物価・低下」のうち「円高・低下」によって 「差益還元」他 9 種の関連語が獲得できた.

これらの関連語を加えた 実験用クエリー (NO.23) を用いた検索では, 拡張源である「円高」が出現し, かつ本クエリーに対する正解である文書 ( ここではこれらをまとめて “正解文書集合$_{円高}$” と記述する) において関連語「差益還元」が頻出しており, それによって “正解文書集合$_{円高}$” のスコアが増し, その結果

として システムが出力する文書ランキングの上位にこれら “正解文書集合$_{円高}$” が押し上げられたため, precision 値が向上した.

失敗例

precision 値の改善に失敗した例として クエリー (NO.40) 「管理部門の統廃合と営業部門の強化を行う会社」の結果を図 6.8に示す.

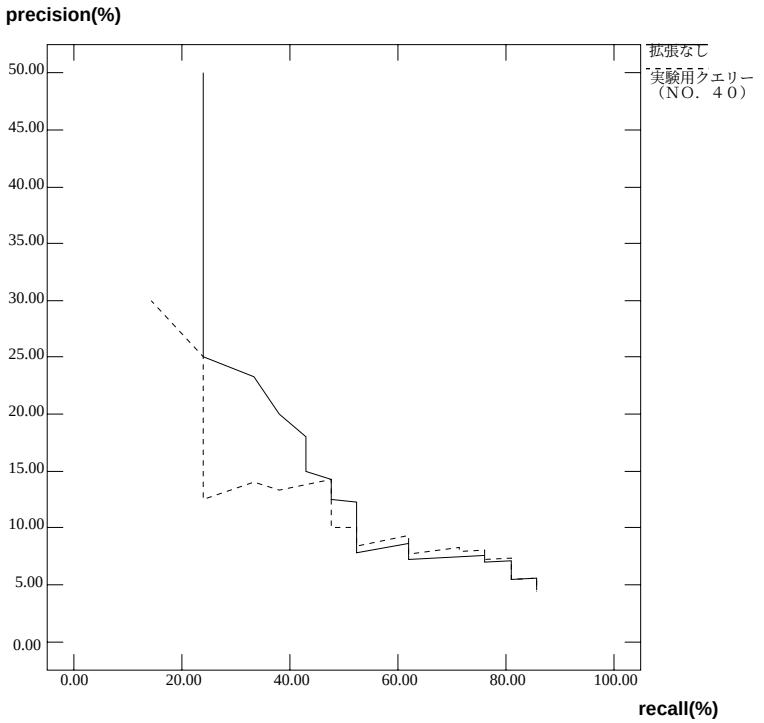



図 6.8: クエリー NO.40 「管理部門の統廃合と営業部門の強化を行う会社」の結果

このクエリーでは, 形態素解析によって「管理・部門・統・廃合・営業・強化・会社」というタームが得られ, そのうち「部[16]・会社」から 「部内」や「総務部」, 「営業・会社」からは「金融会社」や「代理店」といった関連語を獲得しているが, これらの関連語のうち「部内」「総務部」以外は 全て多くの不正解文書を導いてしまっている.

ここでは, この例のように不正解文書を導く関連語と precision 値の向上に貢献する関連語の差を考えてみる.

[16]「部門」の同義語

まず, 2 で例にあげた関連語「合理化」では, 概念説明文における 2 つのターム「人員・削減」の係り受け関係が, もとのクエリー「コンピューターメーカーの人員削減」での係り受け関係とほぼ同じであると考えられる.

| 合理化 | ← | (企業が機械化や不況の影響などで) 人員 を 削減する |
|---|---|---|
| 関連語 | | 概念説明文 |

これに対し, 「金融会社」,「代理店」では, クエリーにおける 2 つのタームの係り受け関係と各概念説明文のそれは異なり, それが 1 つの原因となって直接クエリーとは関連のない語が関連語として獲得されていると考えられる.

| 金融会社 | ← | 金融業を 営業 する 会社 |
|---|---|---|
| 代理店 | ← | 特定の 会社 から委託されて 営業 の代理を行う店 |

また一方で, 本手法では現在, クエリーから概念説明文を検索するための 2 つのタームのペアを作る際, クエリーにおけるそれらのターム間の意味的, 構文的なつながりを何も考慮していないため, それが上記のような不要な関連語を生成する根本的な原因となっていると考えられる. この点が原因となり, 不要な関連語を生成していると思われるケースは他にも多く見られた.

以上から, 本手法では, 2 つのタームと概念説明文間のマッチングにおける信頼性を強化するための対策として, 今後 **(1) クエリーからタームのペアを生成する際のルールなどによる制約の検討**, および **(2) タームと概念説明文とのマッチングの際の構文情報の利用の検討**等を行って行く必要があるだろう.

### 6.4.4 まとめ

2 つの実験によって, 本手法は, 従来のクエリー拡張手法では獲得できないような関連語を生成することができ, またそれらの関連語は, 検索においても部分的ではあるが 有効に働くことがわかった.

しかし, 6.4.2と 6.4.3で考察したように, 本手法の核となるクエリータームと概念説明文とのマッチングにおいて改良すべき点は多い. 今後はこの点を強化することによって, さらに信頼性のある関連語の獲得が行え, それが検索精度の向上にもつながるのではないかと考える.

## 6.5　クエリータームの語義の曖昧性解消

この節では, 5 章で述べた クエリータームの語義の曖昧性解消の手法 (1)–(3) について, まず 初めに 個々の多義性解消の精度について考察し, その後 6.5.4 において 全手法による検索精度における効果についてまとめて実験結果の提示とその考察を行う.

### 6.5.1　手法 (1) 語釈文とシソーラスを利用した手法

本手法では, 5.3.1 で説明した手順 1–3 に従って概念説明文を検索し, その中から 手順 4 に該当する概念説明文を抽出した. その結果, 全 40 クエリー中 4 クエリーのみで該当する概念説明文を獲得した. しかし, そのうち 2 クエリーについては, 多義語ではないタームによって概念説明文が獲得されていたため, 事実上 本手法による効果が得られるクエリーは 2 クエリーのみであった.

そこで, 以下では この 2 クエリーにおいて得られた概念説明文および それによって決定された語義の正否についての考察を行い, またこのように該当する概念説明文が少なかったことによる考察を行う.

なお, 本手法では該当数が少なかったため, 特に本手法による検索精度の効果を調査するための実験は行っていない.

#### 該当する概念説明文が得られたクエリーの考察

まず, 獲得できた概念説明文とクエリーの関係を以下に示す.

なお, □は初期クエリーターム (もしくは その同義語) を指す. また, 『□』 の持つ概念識別子のうち 1 つは, シソーラス上で, 獲得されたレコードの概念識別子との間に “『□』との関係” に記された概念関係を持つ.

また, 1 における初期クエリーは「携帯電話」, 2 は「銀行の経営計画」である.

| NO | | 概念説明文 | 概念識別子 | 『□』との関係 | 正否 |
|---|---|---|---|---|---|
| 1 | a | データ通信機能を 持つ 『電話機』 | 3becc6 | 下位 | ○ |
| | b | 『携帯』 用の無線 『電話』 器 | 1f2a7a | 下位 | ○ |
| 2 | | 個人の出資者によって 経営 される 『銀行』 | 3c3d78 | 下位 | ○ |

1. 全ての概念説明文において, 獲得概念 が□にとって下位 または同義であるのは, 概念説明文で説明する主体を表すタームが□であるためであると考えられる.
2. 実際に □ の語義を決定しているのは 共起する□の効果によると考えられる

   (1) 電話 は, 語義 (1) 通話, (2) 電話機を持つが, 「持つ」や「携帯」が共起することによって 上記のうち語義 (2) の下位概念が獲得できている.

   (2) 銀行 は 語義 (1) 銀行という建物 ,(2) 金融機関を持つが, 「経営される」が共起することによって 語義 (2) の下位概念が獲得できている.
3. 係り受け解析等を行っていないため, 1-a のように, 意味的なマッチングはとれておらず, 偶然に関連概念を獲得できたという可能性もある.

**該当概念説明文が少なかったことに対する考察**

- 実際には 該当する概念説明文が得られたが, その場合の初期クエリータームに多義性がないようなケースがあった.
  → これらも加味すれば, 大規模な実験セットを使った test で, 本手法により 今回の実験よりも大きな効果が得られる可能性はあると思われる.

以上より, 本手法のように,
(A) 異なる 2 つの初期クエリーターム ( もしくは その同義語) が概念説明文に共起しているもので, かつ
(B) クエリータームのうちのいずれかとシソーラス上関連がある

という比較的制約の強い条件によって, 結果的に概念説明文の意味する主体とクエリータームのいずれかの語義が一致している概念説明文が獲得でき, クエリータームの曖昧性の解消に役立つ情報を獲得できることがわかった.

しかし, 本手法では 2 つのクエリータームのうちの一方の語義しか決定できないため, 他の方法との併用が必須である. また, より本手法の正確性を増すためには, 1-a のような意味的に適さないマッチングを防ぐために, 概念説明文におけるクエリータームの同士の係り受け関係に考慮することも必要であると考えられる.

### 6.5.2 手法 (2) 日本語共起辞書を利用した手法

#### 語義の曖昧性解消の実験

ここでは, 本手法における語義曖昧性の解消に関する評価を行う.

#### 実験と結果

表 6.8は, 5.3.2 で説明した一連のルールが適用可能であったタームの割合を coverage[17]で, 適用できたクエリーのうち 正解の語義が得られたタームの割合を accuracy によって評価したものである. なお, これらは クエリーセット “B” に含まれる全 40 クエリーの中で, 多義性のある 55 タームのみを対象としている.

表 6.8: 共起辞書を利用した語義の曖昧性解消の精度

| 評価基準 | (1) 概念を利用した場合 | (2) 表記のみ利用した場合 |
|---|---|---|
| coverage | 69.09 $(\frac{38}{55})$ | 41.82 $(\frac{23}{55})$ |
| accuracy | 55.26 $(\frac{21}{38})$ | 43.50 $(\frac{10}{23})$ |

accuracy は 式 6.6による.

$$accuracy(\%) = \frac{\text{正解の語義が得られたターム数}}{\text{本手法のルールを適用できたターム数}} * 100 \qquad (6.6)$$

#### 考察

1. 本手法における語義曖昧性解消の精度は, 表 6.8のうち, “(1) 概念を利用した場合” の数値である.

   本手法は

   (a) 語義ラベル付きの共起データを用いている

   (b) 2 つ異なる初期クエリータームの概念のペアを作成し, その概念のペアを使って, それに一致する共起データを獲得している

[17] coverage は 6.4 において用いた式による

という 2 つの特徴を持つ.

そこで, 一般の方法 (2 つの異なる初期クエリータームの表記のペアの利用) との精度を比較するために, 5.3.2 と全く同様の方法で, 単に 共起データの検索に表記のペアを使った場合の精度を表 6.8の “(2) 表記を利用した場合” に示す.

(1) と (2) の比較より, 概念を利用した場合の方が coverage においては 約 27%, accuracy においては 約 12% 高い精度が得られることがわかった.

coverage の向上は, 表記のみでは得られなかった共起データを 概念を利用することによって獲得できたことによるもので, 今回利用した日本語共起辞書においては, 概念の共起データを用いることによりスパースネスの問題にも対処できたと言えるだろう.

また, accuracy の向上は, 単に表記のみの共起データを用いるのではなく, 語義も考慮した場合の共起データのみを曖昧性解消に用いているため, 曖昧性解消における正解率を高めることができたことによると考えられる.

2. しかし, 本手法は 獲得された共起データの中で矛盾が生じない場合のみそのデータにおける語義を各初期クエリータームの語義とするという 非常に単純な手法をとる. よって, これによって 1 つのタームに対し, 多くの異なる共起データが得られた場合には全く語義の曖昧性解消は行えず, よって accuracy は 55%に留まっており, 多義である全タームのうちの約 38%に対してしか 正しい語義を決定することができなかった.

### 6.5.3 手法 (3) 拡張で得られた共通概念を利用した手法

#### 語義の曖昧性解消の実験とその結果

**実験と結果** まず, 手法 (3) による coverage と accuracy の値を 表 6.9に示す.

表 6.9: 拡張による共通タームを利用した語義曖昧性解消の精度

| | |
|---|---|
| coverage | 49.09 $(\frac{27}{55})$ |
| accurcy | 85.19 $(\frac{23}{27})$ |

**考察**

- coverage が 低かった原因は 2 点考えられる.

1. 手法 (3) では, 2 つの異なる概念から 名詞・動詞間の概念間関係を利用した拡張により得られた関連概念のうち, 共通の概念数を調査し, その数を基に 5.3.3 で述べたヒューリスティクスを用いて語義曖昧性の解消を行っている.

   しかし, クエリー中に名詞的概念 A を持つタームと動詞的概念 B を持つタームが両方含まれる場合, 拡張によって各概念から得られる関連概念を考えると, A からは 動詞的概念が, B からは 名詞的概念が獲得され, よってこの拡張によっても共通の概念はほとんど獲得できないことがわかる. この結果, 手法 (3) において共通概念がほとんど得られず, データ不足によって語義の曖昧性の解消が行えないケースがいくつか見られた.

2. 本手法では, クエリー中のある 2 つのタームが持つ概念に注目し, スコアによって互いに最もらしい概念であると判定された場合にのみ, 語義の曖昧性解消を行っている.

   しかし, 実際には クエリー中に タームが 3 個以上あるクエリーは全体の約 70%を占める.

   よって, 本手法による場合, 2 タームの語義が互いに最もらしいと判定された時 はじめてその 2 タームの語義が決まるので, 3 ターム以上からなるクエリーの場合, 全語義を決定するのには, A と B, B と C, C と A が 互いに最適 ID でなくてはならず, 多くの場合 いずれかのペア間で互いに矛盾する語義決定がなされ, その結果として 両方のペアにおいて語義決定ができない状況に陥りやすい.

   しかし 一方で, この制約の強い語義決定のヒューリスティクスによって, 決定された各タームの語義は, 正解の語義である場合が多く, その効果が accuracy を 高めていると言える.

- 本手法で用いる ヒューリスティクスは, 利用するスコアやその語義決定ルール, また利用する拡張方法についてまだ充分な検討を行っていないため, 今後 これらを検討することによって, より最もらしい語義への決定が行えるのではないかと考えられる.

### 6.5.4 手法 (2) と手法 (3) の統合

ヒューリスティクスを用いている手法 (3) に対し, 手法 (1) と手法 (2) は 比較的信頼性が高いと考えられる. また 手法 (2) と手法 (3) での多義性解消の結果を細かく分析してみると, 手法 (2) で不正解であったも 17 タームのうちの 8 タームについて, 手法 (3) により正しく語義の決定が行えていることがわかった.

そこで, ここでは coverage の非常に低かった手法 (1) を除いた手法 (2), 手法 (3) の 2 手法に加え, 手法 (4) として, 「手法 (2) を適用した後, 語義決定できなかったタームに対して手法 (3) を適用する」という 2 手法の統合による検索精度への効果を調査する.

この手法 (4) によって, coverage と accuracy は 以下の表 6.10ようになる.

表 6.10: 手法 (2) と手法 (3) の統合による多義性解消精度

| | (2) 共起辞書中のレコードの利用 | (3) 拡張による共通タームの利用 | (4)2→3 |
|---|---|---|---|
| coverage | 69.09 | 49.09 | 76.36 |
| accuracy | 55.26 | 85.19 | 69.05 |
| 全体の正解獲得率 | 38.18 | 34.55 | 52.73 |

また, 検索精度に対する語義曖昧性解消の効果を評価するためにここでは 手法 (2)–(4) に加えて, 各タームの語義を人手で決定した場合の検索精度も調査した. それらの結果を, 図 6.9にまとめて示す.

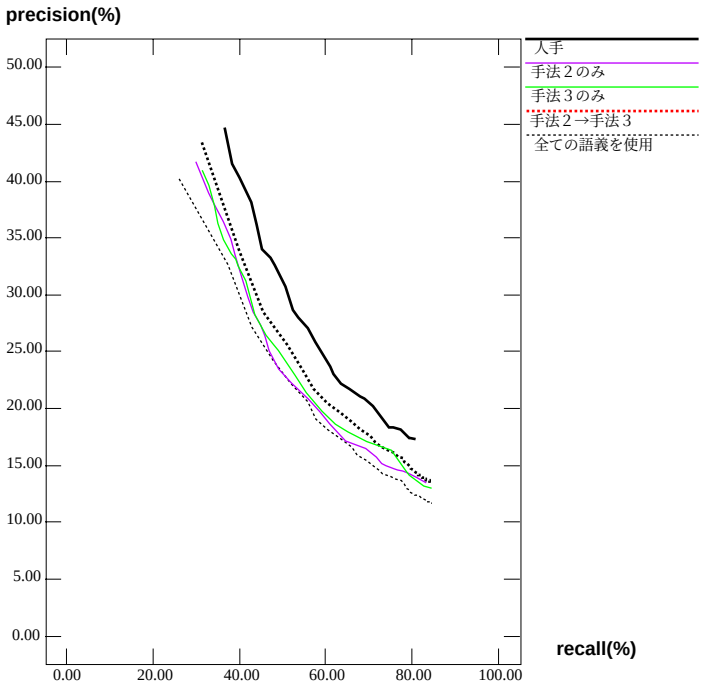



図 6.9: 各多義性解消手法による検索精度 (クエリーセット “B”)

**考察**

- 人手による語義決定による検索精度が最も高く, このことから クエリータームの語義決定が クエリー拡張における精度向上に非常に効果的であることがわかった.
- 手法 (2) のみの場合と手法 (3) のみの場合, 両者共 coverage と accuracy の値もあまり変わらず, 検索精度においても同程度 (2–3%) の precision 値の向上しか得られていない.

  これに対し, 手法 (4) では その他の手法よりさらに 2%程度 precision 値が向上した. これは, 手法 (4) により 全 55 タームの多義語のうち約半数の 52.73%の語義が正しく決定されたことによると考えられる.

### 6.5.5 まとめ

本節では, 概念を用いたクエリータームの語義の曖昧性解消方法についてその曖昧性解消の精度と それによる検索における影響について考察した.

ここで提案した語義曖昧性解消手法は, 2 つの概念の共起によって生じる意味的な制約を用いている. これらの手法は, どれも simple な方法ではあったが, 既存の語義ラベルによって意味的関連性を付されたデータを用いるため, accuracy 値が 約 50%程度において語義の曖昧性の解消が行えることがわかった. また, 確実な語義のみを決定するというスタンスにおいて, これらの手法を統合して行った検索精度調査のための実験では, 各曖昧性手法の特徴を活かし, 統合前に比べより高い precision 値を得ることが可能となった.

今後は, 語義ラベル付きの共起データを用いた統計的な語義曖昧性の手法や, より信頼性のあるヒューリスティクスの構築等を行い, クエリー中に複数のタームが共起する場合には, クエリーターム全体が表す概念を その共起によって自動的にシステムが理解し, さらに クエリーの表す概念により関連の深い概念への拡張に役立てて行くことがクエリー拡張の過程において一つの重要なポイントになると考えられる.

# 第 7 章

# 結論

## 7.1 概念を用いたクエリー拡張と情報検索

本研究では, 概念を様々な方向から利用してクエリー拡張を行い, 各実験より, クエリー拡張に “概念” を使用する場合の情報検索における可能性について検証を行った.

本研究で提案した (1) 名詞・動詞間の概念関係を利用したクエリー拡張手法, (2) 概念説明文を利用したクエリー拡張手法は 共に わずかな精度の向上しか得られなかったが, 個々のクエリーにおける実験結果の考察により, 概念を用いた (1)(2) の手法が有効にはたらくクエリーが存在することがわかった.

また, 本研究では, 初期クエリータームの語義の曖昧性解消について概念を利用した手法を 3 通り提案した. これについての評価実験の結果, 人手による語義決定による精度が最も高く, クエリータームの語義の曖昧性解消は検索精度の向上に有効であることがわかった. また, 提案した 3 手法を統合して曖昧性解消を行うことによって, 人手による語義決定には及ばないものの, 最高で 5%程度の精度の向上が得られた.

なお, 今回全般的に精度の向上がわずかしか得られなかった理由として, 各章で述べた考察以外に, 今回利用した実験セット BMIR-J1 のサンプル数も少ないこと, また 用意されているクエリーと正解文書のセットが クエリー拡張の評価にはあまり適したものではなかったためであるとも考えられる.

## 7.2 今後の課題

- クエリーの拡張手法においては以下のような処理を行うことによって, さらに精度の向上が期待できると考えられる.
  - クエリーおよび概念説明文中の名詞と動詞の係り受け関係の分析と各手法におけるその情報の利用
  - 検索文書集合の特徴を加味した重み付け
- 語義の曖昧性解消手法のうち, 共起辞書のレコードを利用した手法については, 今回の実験結果で 表記を用いるよりも良い結果を得られることがわかったため, 今後は語義ラベル付きの共起データを用いた新しい語義曖昧性の手法に期待できる
- 今回は 各手法共, 拡張クエリーの重み付けについて特に考慮しなかったが, 重み付けによる検索精度の向上の可能性については, 6.2 における単純な実験によりある程度保障されていると言っても良い. そこで, 今後 各手法特徴や検索文書内のタームの分布等によって, 重み付けの手法を考え, その重みが検索精度においてどの程度効果を発揮するのかを調査してみる必要がある.
- 今回は 実験セットのサンプル数が少なかったことから, 各閾値実験においては 全て closed test になってしまっている. よって, 今後は 大規模な実験セットにおける open test の実施を行う必要がある.

  また, 本手法の本来の効果を知るためにも, より クエリー拡張の評価に適した, 大規模な実験セットにおける実験は必要だろう.

# 謝辞

本研究を進めるにあたり終始御指導頂きました奥村学助教授に心から感謝致します. また, 数多くの御助言を頂いた島津明教授, Thanaruk Theeramanunkong 博士に厚く御礼申し上げます.

さらに, 常日頃より議論を重ね, 研究に関して良きアドバイスを下さった自然言語処理学講座の皆様に心より感謝の意を表したいと思います.

最後に, 3 年にわたる JAIST での生活を支えてくれた家族, そして友人に感謝致します.

太田千晶

1998 年 2 月 13 日

# 参考文献


[1] EDR 電子化辞書 仕様説明書EDR,1995

[2] 藤澤浩道 , 絹川 博之, 情報検索における自然言語処理 情報処理, Vol.34, NO.10, pp.1259–1265, 1993

[3] 住田一男, 三池誠司, 知的情報検索の動向, 人工知能学会誌, Vol.11, No.1, pp.10–16, 1995

[4] A.F. Smeaton and C.J. van Rijsbergen, The Retrieval Effects of Query Expansion on Feedback Document Retrieval System, The Computer Journal, VOL.26, NO.3, pp.239–246, 1983.

[5] D. Harman, Relevance Feedboack and Other Query Modification Techniques Information Retrieval - Data Structures & Algorithms pp.241–263, 1992.

[6] 西村英樹, 伊藤耕一郎, 河野浩之, 長谷川利治, 重み付き相関ルール導出アルゴリズムによる WWW データ資源の発見, 電子情報通信学会 第 7 回データ工学ワークショップ (DEWS'96), pp.79–84, 1996

[7] Yonggang Qiu and H.P.Frei, Concept Based Query Expansion Proc. 16th Annual International ACM SIGIR Conference, pp.160–169, 1993.

[8] Ellen, M. Voorhees and Yuan-Wang Hou, Vector expansion in a large collection Proc. First Text REtrieval Conference (TREC-1) pp.343–351, 1993.

[9] Ellen, M. Voorhees, Query Expansion using Lexical-Semantic Relations, Proc. 17th Annual International ACM SIGIR Conference, pp.61–69, 1994.

[10] 松本裕治 , 北内啓 , 山下達雄 , 平野善隆 , 今一修 , 今村友明 日本語形態素解析システム『茶筌』version 1.5 使用説明書1997

[11] G. Salton, Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer Addison-Wesley, 1998

[12] 奥村学, 自然言語の意味的曖昧性の解消法, 人工知能学会誌, Vol.10, No.3, pp.332–339, 1995.

[13] Lesk, M., Automated Sense Disambiguation Using Machine-Readble Dictionaries : How to Tell a Pine Cone from an Ice Cream Cone, Proc.ACM SIGDOC Conference pp.24–26, 1986.

[14] Cowie, J., Guthrie, J. and Gutherie, L., Lexical Disambiguation Using Simulated Annealing, Proc. 14th Int. Conf. on Computational Linguistics, pp.359–365, 1992.

[15] Wilks, Y., Fass, D., Guo, C., McDonald, J., Plate, T. and Slator, B. Providing Machine Tractable Dictionary Tools, Pustejovsky, J.(ed.) *Semantics and the Lexicon* pp.341–401, Kluwer Academic Pub, 1993.

[16] Ellen, M. Voorhees, Using WordNet to Disambiguate Word Senses for Text Retrieval, Proc. 16th Annual International ACM SIGIR Conference, pp.171–180, 1993.

[17] 笠原要, 松澤和光, 概念ベースを用いた常識語の類似検索, 信学技報 TECHNICAL REPORT OF IEICE, AI95-25, 1995.

[18] Brown, P., Pietra, S., Pietra, V. and Mercer, R., Word Sense Disambiguation using Statisical Methods, Proc. 29th Annual Meeting of the Association for Computational Liguistics, pp.264–270, 1991.

[19] 情報検索システム評価用データベース構築ワーキンググループ, 情報検索システム評価用ベンチマーク Ver.1.0 解説書, 1996.

[20] 福島俊一, 小川泰嗣, 石川徹也 他, 日本語情報検索システム評価用テストコレクション BMIR-J1, 自然言語処理シンポジウム「大規模資源と自然言語処理」, pp.1–16, 1996.

[21] 斉藤公一, 森辰則, 中川裕志 概念に基づく検索要求文の拡張, 情処研報 NLP 121-18, pp.127–134, 1997.

# 第 A 章

# 概念記述辞書において使用される「概念関係子」の説明

| 概念関係子 | 説明 |
|---|---|
| *object* | 動作・変化の影響を受ける対象 |
| *agent* | 有意志動作を引き起こす主体 |
| *goal* | 事象の主体または対象の最後の位置 |
| *implement* | 有意志動作における道具・手段 |
| *a-object* | 属性を持つ対象 |
| *place* | 事象の成立する場所 |
| *scene* | 事象の成立する場面 |
| *cause* | 事象の成立する原因 |

# 第 B 章

# 日本語共起辞書のレコード例

<レコード番号> JCC0296769
<句見出し> 携帯/φ/電話
<構文情報>
<共起句構成要素情報>
<要素番号> 1 2
<形態素表記> 携帯 電話
<かな表記> ケイタイ デンワ
<品詞> 名詞 名詞
<構文関係情報>
<受け側要素> 2/電話
<関係要素> ""/φ/""
<係り側要素> 1/携帯
<意味情報>
<始点概念>
<始点要素番号> 2
<概念識別子> 3bdeeb
<概念見出し>
<英語概念見出し> "telephone set"
<日本語概念見出し> 電話機 [デンワキ]
<概念説明>
<英語概念説明> "a telephone set"
<日本語概念説明> 電話機
<概念関係子> modifier
<終点概念>

<終点要素番号> 1
<概念識別子> 3d007b
<概念見出し>
<英語概念見出し> carry
<日本語概念見出し> 携帯する [ケイタイ・スル]
<概念説明>
<英語概念説明> "to carry something with one"
<日本語概念説明> 身につけて持つ
<共起状況情報>
<頻度>
<表層共起頻度> 5
<共起項目頻度> 2
<共起要素頻度>
<受け側共起要素頻度> 190
<係り側共起要素頻度> 18
<例文> 006000019b9a-6-4/"<携帯>(電話)"
<管理情報>
<管理履歴情報> DATE="95/6/16"