TR-IT-0159

意味的類似度と多義解消に基づく文書検索手法

Document Retrieval System Based on
Semantic Similarity and Word Sense Disambiguation

大井 耕三　　隅田 英一郎　　飯田 仁
Kozo OI　　Eiichiro SUMITA　　Hitoshi IIDA

1996.3

概要

本報告では、(1) 質問中の単語と文書中の単語との間の階層的シソーラスに基づく意味的類似度、(2) 質問中の複合語の各単語に類似している文書中の単語間の物理的近さ、(3) 文書内の出現頻度と全文書中の出現文書数に基づく単語の重要度、の3つの尺度に基づいた質問－文書間の関連度計算に加え、コーパスに基づく単語の多義解消手法を導入した文書検索手法について述べる。英語の標準的テストセットを使って実験を行なった結果、単語の重要度で拡張したブーリアンモデルに基づく従来手法に比べて精度の向上を確認した。

ATR 音声翻訳通信研究所
ATR Interpreting Telecommunications Research Laboratories

目次

表目次

図目次

第 1 章

はじめに

近年、膨大な電子化された情報の中から必要な情報を適切に抽出する技術が強く求められている。インターネットの爆発的な普及に伴って、ユーザが求める情報を持つ www サーバを検索するシステムが数多く出現してきている。これらの検索システムのほとんどは、ユーザが入力した検索キーワードを含むテキスト (に対応した www サーバ) を検索するタイプのシステムである。入力されたキーワード「計算機」に対して「コンピュータ」を含むテキストを検索するといった、意味的に類似しているキーワードまで考慮した検索は行なわれず、必ずしも精度が良いシステムであるとは言い難い。

このような問題に対処すべく、さまざまな検索手法が提案されており [住田 96]、近年では、類似性に基づく検索技術の研究 [隅田 91, Sat92, 美馬 96] なども行なわれてきている。

本報告では、(1) 質問中の単語と文書中の単語との間の階層的シソーラスに基づく意味的類似度、(2) 質問中の複合語の各単語に類似している文書中の単語間の物理的近さ、(3) 文書内の出現頻度と全文書中の出現文書数に基づく単語の重要度、の 3 つの尺度に基づいた質問 – 文書間の関連度計算に加え、コーパスに基づく単語の多義解消手法を導入した文書検索手法を提案し、英語の標準的テストセットを使った実験結果について報告する。

第 2 章

検索手法

提案する手法は、各文書に対して質問との関連度を求め、関連度の大きい順に文書をランキングする文書検索に関するものである。

本手法では、階層的シソーラスに基づいた単語間の意味的類似度、および、質問中の複合語の各単語に類似している文書中の単語間の物理的近さ、の 2 つの尺度を定義し、これらに従来から提案されている単語の文書毎の重要度を加えた 3 つの尺度に基づいた質問 - 文書間の関連度に従って文書をランキングする。

さらに精度を向上させるために、これまでに提案されているコーパスに基づく単語の多義解消手法を導入している。

2.1 意味的類似度に基づく関連度

2.1.1 単語間の意味的類似度

階層的シソーラス

単語間の意味的類似度は、単語に付与されているシソーラス上の概念の間の類似度により求め、質問中の単語に類似している単語の検索、および、質問 - 文書間の関連度の計算に用いる。

広く入手可能な電子化されているシソーラスには、角川類語新辞典 [大野 81]、分類語彙表 [国語研 64]、EDR 電子化辞書 [EDR93] の概念辞書 (以下、EDR シソーラスと称する)、WordNet[Mil90]、Roget シソーラス [Cha77] などがあり、いずれも概念を上位 - 下位の関係で階層的に構成している。また、一般に各単語はその意味に対応して 1 つ以上の概念が付与されている。これらのシソーラスの特徴を表 2.1[1] に示す。

角川類語新辞典, 分類語彙表, Roget シソーラスは、ルート概念から末端概念までの階層の深さがほぼ一定である。一方、EDR シソーラス, WordNet は階層の深さは一定でない。EDR シソーラスは概念数が最も多く、また、日本語・英語共通となっており、両言語の文書検索に利用できる。WordNet は品詞毎にシソーラスが異なる

[1] 角川類語新辞典は 1981 年版、分類語彙表は 1964 年版、EDR 概念辞書は評価版 2.1 版、WordNet は Version 1.5、Roget シソーラスは Version 1.02 に関するデータである。

表 2.1: 各種シソーラスの特徴

シソーラス	階層の深さ	概念数	単語数	言語	その他
角川類語新辞典	3(or 4)	1,111	約 60,000	日本語	
分類語彙表	5	1,008	約 32,600	日本語	
EDR シソーラス	18	457,080	202,847	日本語	
			212,827	英語	
WordNet	不明	91,591	不明	英語	品詞毎のシソーラス
Roget シソーラス	4	1,170	不明	英語	

ため、品詞の異なる単語同士の類似度を定義するためには、単語の派生関係を考慮する必要がある。

概念間の類似度

[Sum92] は、角川類語新辞典の上で概念間の類似度を定義し、用例に基づく翻訳に応用している。類似度は、階層の深さが一定であることを利用して、概念間の最も近い共通上位概念の階層上での位置に基づいている。

本稿では、より一般性を持たせるために、EDR シソーラスのように階層の深さが一定でないシソーラスにおいても利用できるように、概念間の類似度を以下のように定義する (図 2.1参照)。

- 各概念に対して、その概念の具体度を表す**レベル**を割り当てる (詳細は後述)。レベルは、下位に位置する概念ほど値が大きくなるように割り当てる。図ではレベルの数 (NL) は 9 となっている。
- 概念 A, B の間の類似度 Sim は、A と B の相対的な位置関係 (3 種類) に応じて、次のように定義する (A, B の最も近い共通上位概念を C とする)。

(1) A と B が同じ場合 ($C = A = B$)、$Sim = 1.0$。

(2) C が A でも B でもない場合、$Sim = LC/NL$。($LC : C$ のレベル).

(3) $A(B)$ が $B(A)$ の上位概念の場合 ($C = A(C = B)$)、$Sim = (LC + 1)/NL$。

Sim は $0, 1/NL, 2/NL, \cdots, 1$ の離散値をとる。この定義は「下位概念 (末端までの概念すべて) の総数が少ない概念ほど、概念の具体度 (レベル) が高い」という仮説に基づくもので、類似度は概念間の最も近い共通上位概念の具体度に比例した値としている。

レベルの割り当て方法に関しては、あらかじめレベルの数 NL(類似度の離散値の数) を決めておき、階層の深さが $NL - 1$(ルート概念の深さを 0 とする) であるバラン

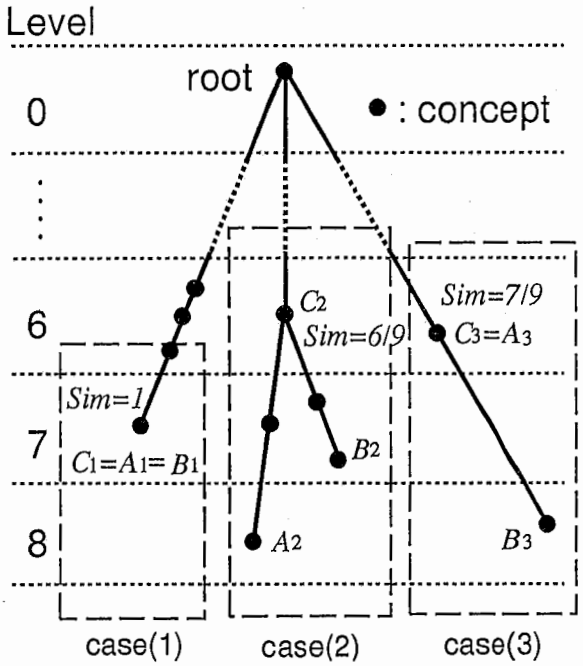

図 2.1: シソーラス上の概念間の意味的類似度の例 (Sim: 概念 A, B の間の類似度, C: A と B の最も近い上位の共通概念)

スのとれたシソーラス (階層の深さが一定で、各概念からの下位概念への分岐数が一定) を想定した時の、深さ d に位置する概念の下位概念の総数を、実際のシソーラスのレベル d の概念の下位概念の総数に対応させる。各概念のレベルの詳細な割り当て手順を次に示す。

1. シソーラスの延べ概念数 TC を求める。直上位の概念 (親概念) を複数もつ概念は、その数だけ別個の概念が存在していると考えて、延べ概念数を求める。
2. レベルの数 NL を任意に決める。
3. バランスのとれたシソーラスとして、階層の深さが $NL-1$(ルート概念の深さを 0 とする) で、総概念数が TC であるシソーラスを想定する。そのときの各概念からの下位概念への分岐数 NB を求める。NB が求まれば、深さが d の概念の下位概念の総数 $TLD(d)$ は次式となる。

$$TLD(d) = \begin{cases} \sum_{k=1}^{NL-1-d} NB^k & (d < NL-1) \\ 0 & (d = NL-1) \end{cases}$$

4. シソーラスの各概念に対して、
   (a) その概念の下位概念の総数 TLC を求める。

(b) $TLC \leq TLD(d)$ である最小の d をその概念に**レベル**[2]として割り当てる。

図 2.2に、延べ概念数が 40 万のシソーラスに対して、レベルの数 $NL = 9$ に設定したときの、下位概念の総数とレベルの関係を示す。

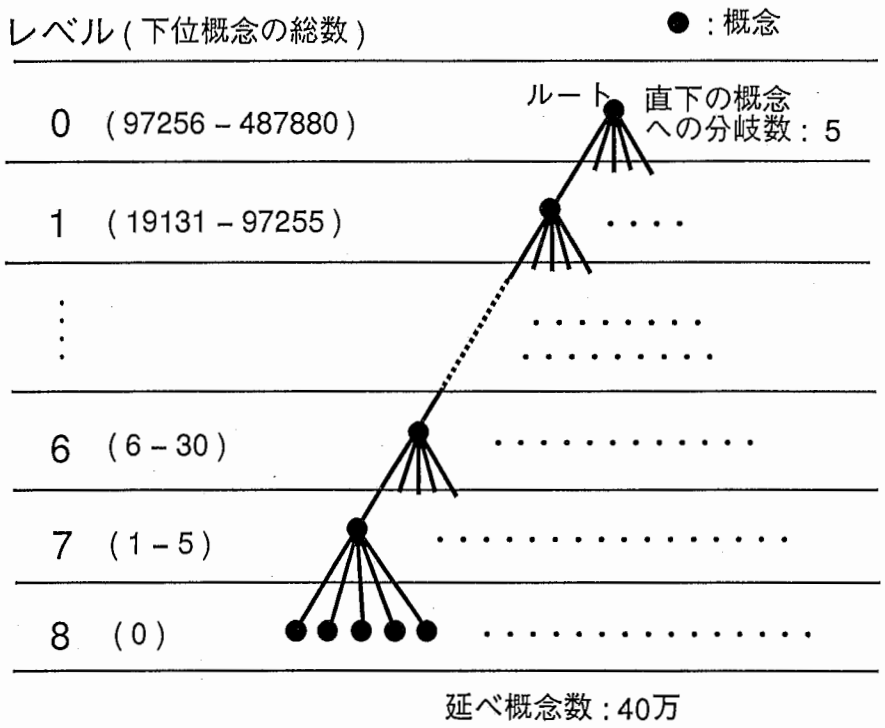

図 2.2: バランスのとれたシソーラスと概念のレベル (具体度)

本報告の実験では、階層の深さが一定でないシソーラスの代表として EDR 電子化辞書の概念辞書と単語辞書[3]を使用した。英語単語辞書は約 213,000 語から成り、各単語には 1 つ以上の概念が付与されている。

単語間の類似度

単語間の類似度は次のように求める。

(1) 単語一致 (単語の見出しと品詞が同じ) の場合
類似度は単語の概念が一致している場合の類似度よりも大きい値に設定する。実験時の設定値は 3.2節の表 3.2を参照。

(2) 単語一致以外の場合
類似度は、それぞれの概念のすべての組合せにおける 2 概念間の類似度の最大値とする。2 概念間の類似度は前項で述べた定義により求める。

[2]実際のシソーラスでは、このレベルはルート概念からの深さを表す値ではない。
[3]単語辞書に単語と概念の対応が、概念辞書に概念間の関係が記述されている。

2.1.2 単語間の物理的近さ

本手法では、質問中には、単語のほかに複合語 (2 単語以上からなるもの) の指定も可能にしている。質問中に複合語が指定された場合は、その複合語中の各単語に類似している文書中の単語間の物理的近さを、質問 - 文書間の関連度に反映させる。

物理的近さは、例えば、複合語が “parallel algorithm” の場合、“parallel” と “algorithm” それぞれに類似した単語が文書中に出現しているとき、その類似した単語間の物理的距離が近いほど大きい値になるよう定義する。この尺度は、質問中の複合語と文書間の類似度の乗数として用いる (詳細は 2.1.4節参照)。

物理的近さ PN の定義式を次に示す。

$$PN = c_1 \times \frac{1}{\frac{c_1 - 1}{c_2} \times (Dis + 1 - N) + 1} \tag{2.1}$$

ここで、

c_1, c_2: 定数
Dis: 類似単語間の物理的距離 (単語間に存在する単語数 +1) の最小値。
N: 複合語の単語数。

式中の定数 c_1 は、類似単語が隣接している場合の PN に相当する。例えば $c_1 = 2$ に設定したときに、類似単語が隣接している場合は PN=2 となり、質問中の複合語と文書間の類似度は 2 倍される。定数 c_2 は、例えば 2 単語からなる複合語の場合、類似単語間の物理的距離が $c_2 + 1$ のときに PN=1 となる値である。実験時の設定値は 3.2節の表 3.2を参照のこと。

2.1.3 単語の文書毎の重要度

単語の重要度の定義に関しては様々な手法が提案されている [Sal88]。本手法では、単語の文書内の出現頻度と全文書中の出現文書数に基づいた重要度として [Tur91] で定義された重要度を使う。文書 D 内の単語 dt の重要度 w は、次のように定義される。

$$w = \frac{tf}{max\_tf} \times \frac{\log \frac{ND}{f}}{\log ND} \tag{2.2}$$

ここで、

tf: 文書 D 内の単語 dt の出現頻度。
max_tf: 文書 D 内の各単語の出現頻度のうち、最大の出現頻度。
f: 単語 dt が出現している文書の数。
ND: 文書数。

2.1.4 質問－文書間の関連度

質問

本手法では、質問 Q は次に示すようなブーリアンの形式で表現可能なものを前提とする。

$$Q = q_1 \mid q_2 \mid \cdots \mid q_K \tag{2.3}$$

$$q = \underbrace{qt_{11}, \cdots, qt_{1N_1}}_{\text{質問語 } qc_1} \& \cdots \& \underbrace{qt_{M1}, \cdots, qt_{MN_M}}_{\text{質問語 } qc_M} \tag{2.4}$$

ここで、'|' と '&' はそれぞれ 'OR' と 'AND' のオペレーションを表す。この形式は、K 個の q が 'OR' オペレーションで結合され、各 q は M 個の質問語 qc(単語または複合語) が 'AND' オペレーションで結合され、各質問語 qc_i は N_i 個の単語 qt から成っている。

関連度

質問 Q と文書 D との間の関連度 $Sim(Q,D)$ は次式のように定義する。

$$Sim(Q,D) = max\{Sim(q_1,D), Sim(q_2,D), \cdots, Sim(q_K,D)\} \tag{2.5}$$

$$Sim(q,D) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N_i}\{Sim(qt_{ij},D) \times PN(qc_i,D)\}^2}{\sum_{i=1}^{M}\sum_{j=1}^{N_i}\{Sim(qt_{ij},D) \times PN(qc_i,D)\}} \tag{2.6}$$

$$Sim(qt,D) = max(w_1, w_2, \cdots, w_L) \times max\{Sim(qt,dt)\}$$

ここで、

- L: qt に最も類似している単語 ($Sim(qt,dt)$ が最大の単語) の数。
- w: qt に最も類似している単語の重要度。
- $PN(qc,D)$: 文書 D 内での、qc 中の各 qt に最も類似している単語間の物理的近さ (2.1.2節の式 (2.1))。
- $Sim(qt,dt)$: 質問 Q 中の 1 つの単語 qt と文書中の 1 つの単語 dt との間の類似度。詳細は 2.1.1節を参照。

$Sim(Q,D)$ の計算は、$q_1, q_2, \cdots, q_K$ のうち、少なくとも 1 つの q に含まれる単語それぞれに対して類似している単語が出現している文書に対してのみ行なう。「類似

している単語」とは、$Sim(qt, dt) \geq$ [あらかじめ指定された**しきい値** T] である単語 dt を意味する。

$Sim(q, D)$ は、単に q 中の単語 qt_{ij} すべてに対する $Sim(qt_{ij}, D) \times PN(qc_i, D)$ の値の平均とするよりも、$Sim(qt_{ij}, D) \times PN(qc_i, D)$ が大きい値ほどその値の重要度が大きくなる定義式になっている。

2.2 多義解消手法の導入

多くの単語は、いくつかの意味 (多義) を持っている。単語間の類似度は、2.1.1節で述べたように、概念のすべての組合せにおける 2 概念間の類似度の最大値としている。この場合、実際にその文書で用いられている意味と異なった意味での類似度が高くなって、誤って関連文書として検索されてしまう場合が生じる。

例えば、'performance evaluation' という質問に対して、次のような文書が検索される。

[検索された文書]

```
 ..........................................................
It is important to choose a class of signals which is, at
present, undergoing a good deal of visual inspection by trained
people for the purpose of pattern recognition.  In this way
comparisons between machine and human performance may
 ..........................................................
```

質問中の単語 "evaluation" の概念は、"to assign a value to; to appraise; to value" である[4]。

文書中の単語 "recognition" に付与されている概念 (概念数:10) の中には、文書中で使われている概念である "to perceive something" のほかに、"to assign a value to; to appraise; to value" があるため、"evaluation" の概念と一致してこの文書が検索されてしまう。

従って、実際に文書中で使われている意味を同定することができれば、検索の精度はかなり向上することが期待できる。そこで、単語の多義をコーパスに基づいて解消する手法を導入することにした。

---

[4]質問は、単語・複合語の AND/OR 結合で簡潔に表現される。このような場合にも有効な機械的多義解消は容易ではないと考えられる。しかし、ユーザとの会話的処理によって、この質問中の単語の多義解消することは実現可能である。本手法ではこれを前提としている。実験に際しては、あらかじめ各単語に適切な概念を人手で付与した。

2.2.1 Voorhees の手法の適用

Voorhees は、同一文書内では関連する単語が出現しやすいという仮説に基づいた手法により多義解消を行ない、文書検索の実験を行なっている[5][Voo93]。

文書 D 中の単語 dt の Voorhees の多義解消手法の概要は次のとおりである。単語 dt の各意味に対して、***hood***(その意味を含み他の意味を含まない最も上位の意味) を求めた後、式 (2.7) の差異を求め、差異が最大である意味を単語 dt の意味として選択する。式中の「内容語」とは、ストップリスト (冠詞や前置詞など重要と考えられない単語のリスト) に属さない単語を意味する。

$$\text{差異} = \frac{hood\text{下の意味を持つ内容語の文書}D\text{内出現数}}{\text{内容語全体の文書}D\text{内出現数}} - \frac{hood\text{下の意味を持つ内容語の全文書内出現数}}{\text{内容語全体の全文書内出現数}} \tag{2.7}$$

この手法を我々の検索手法に適用するに際し、次の変更を加えた。

(1) 1つの単語の複数の概念の中で似ている概念をまとめるために、各概念をあるレベル (2.1.1節参照) の上位概念 (以下、**置換概念**と称する) に置き換える。

(2) *hood* は最も上位の概念でなくレベルによって制限する (抽象的すぎる概念まで含むのを避けるため)。

(3) 差異が正である概念 (置換概念) を単語 dt の概念として選択する。

(1) および (2) のレベルの実験時の設定は、3.2節の表 3.2を参照のこと。

2.2.2 Yarowsky の手法の適用

Yarowsky は、Roget シソーラスを用いて多義解消の実験を行なっている [Yar92]。

文書中の単語 dt の多義解消手法の概要は次のとおりである。コーパス中の各単語に対して、前後 50 語ずつ合計 100 語からなる文脈を抽出し、単語 dt が属するシソーラス中のカテゴリ RCat 毎に、単語 dt の文脈中の単語のうち式 (2.9) があるしきい値 $\boldsymbol{Y}$ 以上 (実験時の設定値は 3.2節の表 3.2を参照) の単語 ct に対して、式 (2.8) の $Score$ を求め、最大の $Score$ となるカテゴリを単語 dt の意味としている。

[5]Voorhees はシソーラスとして WordNet を使い、ベクトル空間モデルに基づいた手法により文書検索を行なっている。多義解消された単語の意味はベクトルとして表現している。一方、我々は EDR シソーラスを使い、ブーリアンモデルを意味的類似度と物理的近さと重要度で拡張した手法により文書検索を行なっている。

$$Score = \sum_{ct} \log \frac{Pr(ct|RCat)}{Pr(ct)} \tag{2.8}$$

$$\frac{Pr(ct|RCat)}{Pr(ct)} = \frac{RCatに属する単語の文脈中の単語ctの出現確率}{単語ctの全文脈中の出現確率} \tag{2.9}$$

この手法を我々の検索手法に適用するに際し、次の変更を加えた。

(1) 1 つの単語の複数の概念の中で似ている概念をまとめるために、単語の各概念を置換概念 (2.2.1節参照) に置き換える。

(2) Roget シソーラスのカテゴリを Voorhees の手法における *hood* 下の概念に置き換える[6]。

(3) スコアが正である概念 (置換概念) を選択する (Voorhees の手法の適用と同様)。

(4) 前後 4 語ずつ合計 8 語を文脈とする[7]。

2.3 検索処理

実際の検索処理に先だって、(a) 質問中の単語への概念の付与 (手作業)、(b) 全文書に対するインデックスファイル (各文書の各見出し語・品詞のペアに対して、文書番号、重要度、出現位置が付与されたファイル) の作成、を行なった。

インデックスファイルの作成手順は次のようになる。

(i) 文書の中から、ストップリスト (冠詞や前置詞など重要と考えられない単語のリスト) に属する単語を除き、すべてのアルファベットを小文字化する [Fox92]。

(ii) 形態素解析[8]により、各単語に対して、可能な見出し語・品詞のペア、出現位置の情報を付与する。

(iii) 2.1.3節の式 (2.2) に従って、各文書の各見出し語・品詞のペア対して、重要度を計算し付与する。

質問に対する検索処理手順は次のとおり。

(1) 質問を 2.1.4節の式 (2.3),(2.4) の形に変換する。

[6]Roget シソーラスのカテゴリを EDR の概念にそのまま置き換えたとすると、EDR シソーラスの概念は、特に最下位概念はそのほとんどが非常に具体的な概念であるため、各概念に属する単語の文脈は非常にスパースとなる。

[7]文脈幅を変えて実験を行なったが、8 語以上にしても、結果に大きな差は生じなかった。

[8]本報告の実験では、[Kar92] の形態素解析プログラムを使用した。

(2) 式 (2.3) の少なくとも 1 つの q 中の単語 qt それぞれに対して、
$Sim(qt, dt) \geq T$(しきい値) を満たす単語 dt が出現している文書を検索する。

(3) 検索された文書毎に、質問 - 文書間の関連度 $Sim(Q, D)$ を 2.1.4節の式 (2.5) から求める。

(4) 検索された文書を関連度順にランキングする。

第 3 章

実験

3.1 標準的テストセット

評価用に作成され公開されている英語のテストセットの 1 つに Fox[Fox90] が作成したものがある。実験では、その中の CACM と呼ばれるセットを使った。CACM には、コンピュータサイエンスに関する 3,204 の文書 (タイトル, アブストラクト)、3 種類の質問セット (自然言語文からなる NLQ, ブーリアン形式の BLQ1, BLQ2)、質問ごとの関連文書の文書番号が含まれている。各質問セットには 64 個の質問が含まれている。NLQ はオリジナルの質問セットで、BLQ1, BLQ2 は NLQ を基にして作られている。

実験では、NLQ の質問文中の単語が比較的多く使われている BLQ2 の中で、NOT オペレーションを含むものと正解の関連文書がないものを除く 47 個の質問に変更を加えたもの (以下では、変更版 BLQ2 と呼ぶ) を用いた。変更版 BLQ2 は、複合語と考えられる部分を複合語として指定したものである。NLQ, BLQ2, 変更版 BLQ2 の質問番号 35 の質問内容を下に示す。変更版 BLQ2 において、シングルクウォート (') で囲まれた部分が 1 つの質問語を表している。

[NLQ]

```
Probabilistic algorithms especially those dealing with algebraic
and symbolic manipulation.
Some examples: Rabiin, "Probabilistic algorithm on finite field", SIAM.
Waztch, "Probabilistic testing of polynomial identities", SIAM.
```

[BLQ2]

```
#q35= #and( 'probabilistic', 'algorithm',
            #or( 'algebraic', 'symbolic'), 'manipulation');
```

[変更版 BLQ2]

```
#q35= #and( 'probabilistic algorithm',
            #or( 'algebraic manipulation', 'symbolic manipulation'));
```

3.2 評価方法および実験条件

評価は、情報検索の分野で一般的によく使われている再現率 (recall) と適合率 (precision) を用いた。再現率および適合率は次式で定義される。

$$\text{再現率} = \frac{\text{ランク}N\text{位までの検索文書中の関連文書数}}{\text{関連文書数}} \tag{3.1}$$

$$\text{適合率} = \frac{\text{ランク}N\text{位までの検索文書中の関連文書数}}{N} \tag{3.2}$$

再現率 – 適合率のグラフは次に示す手順で作成した。

[再現率 – 適合率グラフの作成手順]

(1) 式 (3.1),(3.2) の値 N を任意に複数個決める。

(2) 質問ごとに、式 (3.1),(3.2) により N における再現率と適合率を求める。

(3) N における再現率と適合率の全質問に対する平均を求め、プロットする。

次節の結果では、N を $10, 20, 30, \cdots, 200$ に設定し、表 3.1に示す 4 種類の手法の実験結果を比較した。このうち WM(単語の重要度で拡張したブーリアンモデルに基づく手法) を比較の基準とした。

表 3.1: 実験を行なった 4 種類の手法

手法名	使用する尺度または多義解消手法				
	意味的類似度 (2.1.1節)	物理的近さ (2.1.2節)	単語の重要度 (2.1.3節)	Voorhees の手法 (2.2.1節)	Yarowsky の手法 (2.2.2節)
WM			○		
AM	○	○	○		
DM(Voo)	○	○	○	○	
DM(Yar)	○	○	○		○

実験条件を表3.2に示す。

表 3.2: 実験条件

項目 (関連する章節)	設定
シソーラスの概念のレベル数 NL (2.1.1節)	9
単語一致の時の $Sim(qt, dt)$ (2.1.4節)	$10/9(=(NL+1)/NL)$
物理的近さの定義式 (2.1) の定数 c_1 (2.1.2節)	2
物理的近さの定義式 (2.1) の定数 c_2 (2.1.2節)	10
Voorhees および Yarowsky の手法における置換概念のレベル (2.2.1節,2.2.2節)	6[9]
Voorhees の手法における *hood* 概念のレベル (2.2.1節)	3 より具体的
Yarowsky の手法において単語 ct を選択するための式 (2.9) のしきい値 Y (2.2.2節)	2.0 以上
Yarowsky の手法における使用コーパス	TIPSTER の ZIFF 文書

表中、Yarowsky の手法において使用したコーパス「TIPSTER の ZIFF 文書」は、米国 ARPA(Advanced Research Projects Agency) が先進技術の研究促進を目的として推進している情報抽出・情報検索のプロジェクト TIPSTER で使われている文書の1つである。ZIFF 文書の総単語数は約 9,600 万語であるが、実験では多義解消の処理時間を短縮するためにそのうちの約 800 万語を使用した[10]。

3.3 結果

3.3.1 文書検索の結果

図 3.1–3.4 は、しきい値 T(2.1.4節参照) がそれぞれ 9/9, 8/9, 7/9, 6/9 の場合の再現率 – 適合率グラフである。

[9] レベル 6 の概念が存在しない場合は、より上位の概念に置換する。

[10] Yarowsky の手法は使用するコーパスのサイズと文脈幅が大きくなると処理時間も増えていく。800 万語の ZIFF 文書をコーパスとして多義解消に要した時間は DEC Alpha 上で約 330 時間である。参考までに検索対象文書 (約 17 万語) をコーパスとして使用した場合の多義解消の処理時間は約 11 時間である。また、Voorhees の手法の場合の多義解消時間は約 3 時間である。

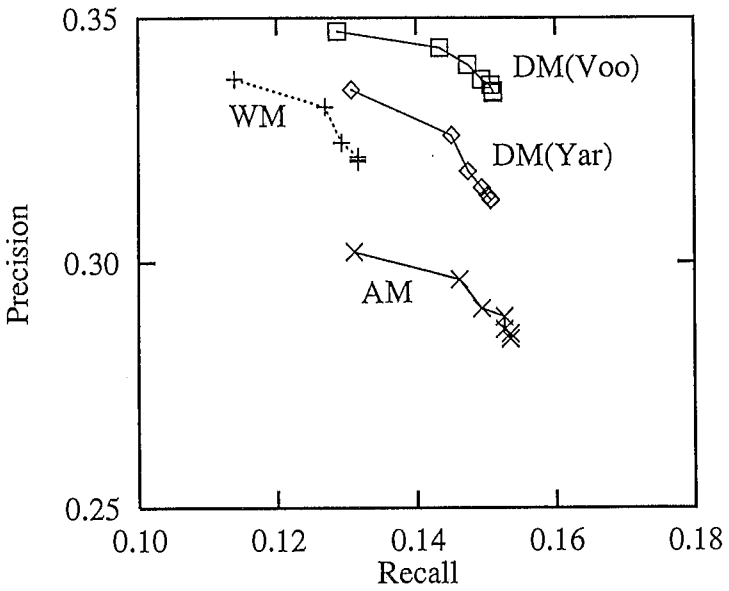

図 3.1: 再現率 – 適合率 (recall-precision)[しきい値 T=9/9]

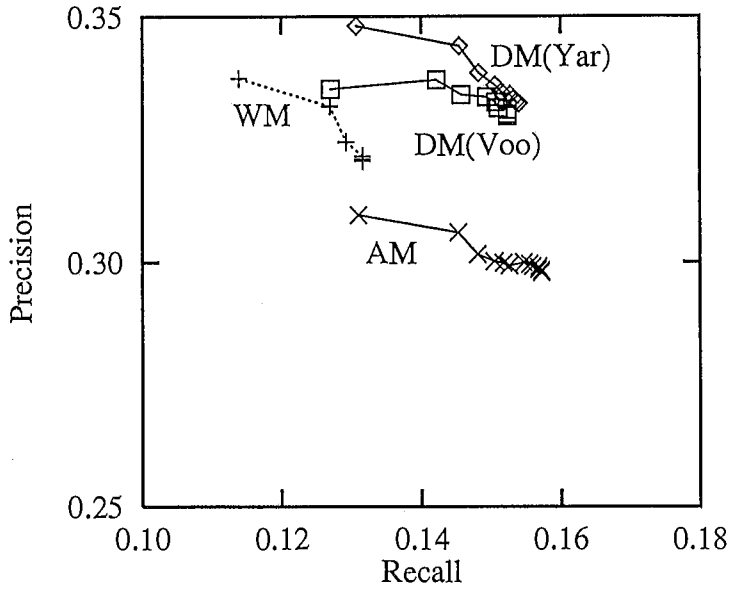

図 3.2: 再現率 – 適合率 (recall-precision)[しきい値 T=8/9]

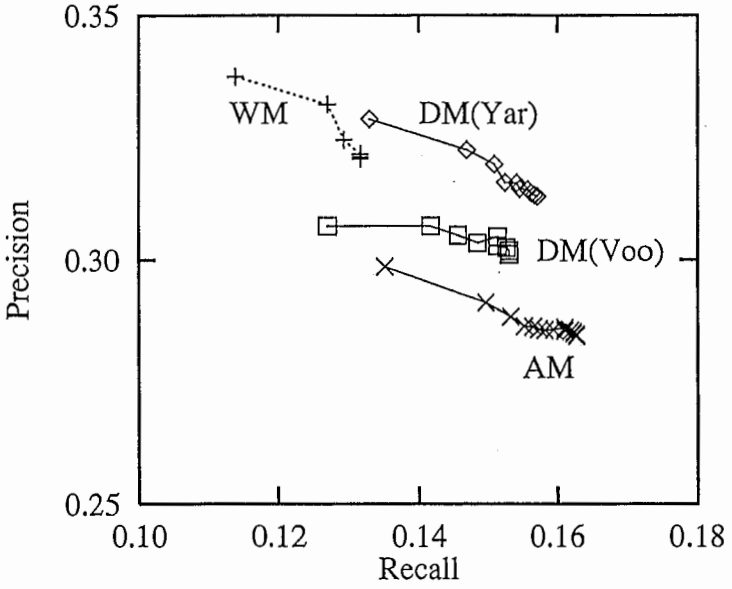

図 3.3: 再現率 – 適合率 (recall-precision)[しきい値 T=7/9]

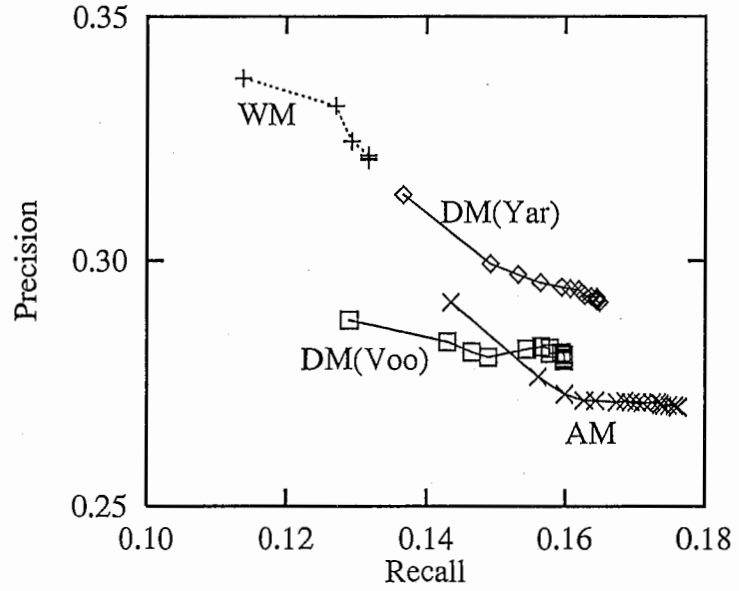

図 3.4: 再現率 – 適合率 (recall-precision)[しきい値 T=6/9]

いずれのしきい値 T においても、AM は、WM と比べて再現率は向上する一方、適合率は減少している。この原因の 1 つに単語の多義性がある。単語間の類似度を求める際にすべての概念を考慮しているので、実際に文書中で使われている概念とは異なる概念に起因して多くの非関連文書が検索されてしまうのである。

多義解消手法を導入したDM(Voo)とDM(Yar)をAMと比べた場合、T=9/9(図3.1), T=8/9(図3.2)では、再現率はわずかに減少しているものの、適合率は明らかに向上している。これは、AMで検索されていた関連文書がDMでは多義解消の失敗により検索されなくなった場合がわずかに生じているものの、それ以上にAMで検索されていた非関連文書がDMでは多義解消の効果により検索されなくなったためである。T=7/9(図3.1), T=6/9(図3.4)では、DM(Yar)はあまり再現率が減少することなく適合率が向上しているが、DM(Voo)は多義解消の失敗の影響で再現率が減少しているのが目立つ。

DMをWMと比べると、いずれのしきい値Tにおいても再現率は向上している。DM(Yar)の適合率は、T=8/9(図3.2)の場合、明らかに向上している。T=9/9(図3.1),T=7/9(図3.3)ではわずかに減少しているが、ある再現率(例えば0.13あたり)で比較すると適合率が向上しているのが分かる。T=6/9(図3.4)では、適合率は減少している。DM(Voo)の適合率は、T=9/9(図3.1), T=8/9(図3.2)の場合、向上している。T=7/9(図3.1), T=6/9(図3.4)では、再現率は向上しているが、適合率は減少している。従って、本手法は、ある程度しきい値Tが高い値で有効であると言える。DM(Voo)とDM(Yar)では全体的にDM(Yar)の方が優れている結果が得られた。

3.3.2 多義解消の結果

本節では、Voorheesの多義解消手法(以下、DA(Voo)と称する)およびYarowskyの多義解消手法(以下、DA(Yar)と称する)の精度について述べる。

関連文書内の単語の中で、T=6/9において質問中の単語と類似している単語(異なり数: 45, 延べ数105)のうち、多義解消により1つ以上の置換概念[11]を選択できた単語(延べ数[DA(Voo):104, DA(Yar):93]について多義解消の結果を調べた。表3.3に結果を示す。

表3.3: 多義解消結果

手法	A: 延べ単語数	B: 成功単語数	B/A: 成功率	C: 延べ置換概念数	D: 選択置換概念数	B/D / A/C
DA(Voo)	104	73	0.70	938	540	1.219
DA(Yar)	93	59	0.63	913	292	1.984

2.2.1および2.2.2節で述べたように、多義解消ではスコアが正である置換概念を選択している。成功単語数は、選択した置換概念の中に正解概念(文書中で使われている概念)が含まれている単語の数を表す。選択置換概念数は、多義解消により選択された置換概念の数を表す。

[11]「置換概念」に関しては2.2.1節を参照。

多義解消の成功率 (B/A) は、DA(Voo) の方が DA(Yar) よりも高い。しかしながら、$T = 6/9$ のときの検索結果 (図 3.4) では DA(Yar) の方が精度が高い。スコアが正である置換概念を選択する我々の手法においては、延べ置換概念数に対する選択置換概念数の割合が高ければ成功率も高くなるので、成功率が直接検索結果に反映されるわけではない。多義解消の精度を比較するためには、延べ置換概念数に対する選択置換概念数の割合を考慮する必要がある。

そこで、多義解消なしの場合の延べ置換概念中に含まれている正解置換概念[12]の割合 (A/C) と、多義解消した場合の選択置換概念中の正解置換概念の割合 (B/D) を比較することにする。その場合、正解置換概念の割合の向上率 (B/D)/(A/C) は、DA(Voo) では約 1.219 倍、DA(Yar) では約 1.984 倍であり、どちらも向上しているが、DA(Yar) の方がより精度が優れており、このことが図 3.4 の検索結果において DA(Yar) が DA(Voo) より優れていることに結び付いている。

多義解消例

DA(Voo) の多義解消例として、文書番号 1719 の文書 (図 3.5) 中の単語 "evaluating" の結果を表 3.4に、DA(Yar) の多義解消例として、文書番号 1411 の文書 (図 3.6) 中の単語 "formulas" の結果を表 3.5に示す。表中、「概念」とは置換する前の概念、すなわち、単語辞書中で付与されている概念を表す。「置換概念」および「概念」の欄に示している文字列は、EDR シソーラスに記述されている英語概念説明または日本語概念説明である[13]。「スコアの計算に貢献した文書 (または文脈) 内の単語」は、DA(Voo) では *hood* 下の概念をもつ文書内の単語を、DA(Yar) では文脈中の単語のうち 2.2.2節の式 (2.9) のしきい値 Y が 2.0 以上 (表 3.2の実験条件を参照) の単語を表す。

[12]正解置換概念数は延べ単語数に等しい。

[13]EDR シソーラスの各概念には、概念識別子と呼ばれる 16 進数、日本語・英語概念見出し、日本語・英語概念説明が付与されている。日本語・英語概念見出し、日本語・英語概念説明の中には付与されていない項目もあるため、表には基本的に英語概念説明を記し、英語概念説明が付与されていないものについては日本語概念説明を記している。

表 3.4: DA(Voo) による単語 “evaluating” の多義解消例

置換概念	概念	スコア	スコアの計算に貢献した文書内の単語
“評価する”	“to determine a numerical representation for”, “to assign a value to; to appraise; to value”	0.123	calculation, calculating, cost, outlined, types, optimizing, optimization
“the act of calculating”	“the act of calculating”	0.079	calculation, calculating
“いろいろな思考的活動”	“to take a theoretical guess”	-0.063	typified, illustrated, reservation

```
 ..........................................................
Real-time data processing systems as typified by the automated
airline reservation system are discussed in this paper.
Criteria for evaluating performance are described; a methodology
for calculating and optimizing is outlined; and the method is
 ..........................................................
```

図 3.5: 単語 “evaluating” が出現する文書 (文書番号 1719)

表 3.5: DA(Yar) による単語 “formulas” の多義解消例

置換概念	概念	スコア	スコアの計算に貢献した文脈内の単語
“数や記号の組合せからなる式”	“a formula that leads to an answer”	3.033	value, evaluate
“抽象的生産物”	“a solution employing several methods, plans, and ···”	0.000	
“a group of words that express a complete thought”	“a predetermined set of words or actions for an occasion”	0.000	
“伝達内容”	“a hackneyed expression”	0.000	
“規則”	“rules or ways of doing that are observed without thought and performed mechanically”	0.000	
“情報媒体”	“an expression using numeric and alphabetic symbols to represent a chemical compound”	0.000	
“nourishment for a person’s body”	“a nutrient given to a baby when the baby is not using mother’s milk”	0.000	
“書物”	“a list of the ingredients of a mixture such as a medicinal preparation, a food or a drink, etc.”	0.000	

```
.....................................................
For each statistic, the algorithm included the usual computing
formulas, correction due to an accumulated error term, and a
recursive computation of the current value of the statistic.
The usual computing formulas were also evaluated in double
precision.  Large errors were noted for some calculation using
.....................................................
```

図 3.6: 単語 “formulas” が出現する文書 (文書番号 1411)

第 4 章

関連研究

シソーラスを使った検索手法に関して、 Rada[Rad85], Paice[Pai91], 松尾 [松尾 91] などの先行研究がある。

Rada はシソーラス上の概念間の類似度を使っている。しかしながら、 Rada はシソーラス上の概念を個々の質問と文書に対して手作業で付与し、文書や質問中の単語に付与していない。よって我々の手法のように質問中の単語に類似した単語を含む文書を検索することはできない。

Paice は意味ネットワークからなるシソーラスを使っている。質問単語に対応するシソーラス中のノードから一定の距離内にあるノード (単語) に重み付けを行ない、それを拡張単語として検索を行なうことを提案している。我々の手法では、概念間の類似度を階層的シソーラス上の共通上位概念の位置に基づいて求め、質問中の単語に対してあるしきい値以上の類似度である単語をもつ文書を検索している。また、 Paice では実験結果は示されていない。

松尾は、文の検索を対象に、各単語に付与されたシソーラス上の意味属性の集合として表現された文の間の類似度を、意味属性および表記の一致度をベースに求めており、我々の意味概念の間の類似度をシソーラスに基づいて求める手法と異なる。

第 5 章

おわりに

階層的シソーラスに基づく単語間の意味的類似度、単語間の物理的近さ、単語の重要度、の 3 つの尺度に基づいた質問 - 文書間の関連度計算に加え、コーパスに基づく単語の多義解消手法を導入した文書検索手法について報告した。英語の標準的テストセットを使った実験の結果、本手法は単語の重要度で拡張したブーリアンモデルに基づく検索手法に比べ再現率・適合率ともに向上することを確認した。

謝辞

本研究では EDR 電子化辞書の英語単語辞書・概念辞書 (評価版第 2.1 版) を使用している。関係各位に深謝する。

参考文献

[Cha77] Robert Chapman, editor. **Roget's International Thesaurus (Fourth Edition)**. Harper and Row, New York, 1977.

[EDR93] 日本電子化辞書研究所. **EDR 電子化辞書仕様説明書**, March 1993.

[Fox90] Edward A. Fox, editor. **Virginia Disk One**. Virginia Polytechnic Institute and State University Press, Blacksburg, 1990.

[Fox92] Christopher Fox. **Lexical Analysis and Stoplists**. In William B. Frakes and Ricardo Baeza-Yates, editors, *Information Retrieval: Data Structures and Algorithms*, chapter 7, pp. 102–130. Prentice Hall, 1992.

[Kar92] D. Karp, Y. Schabes, M. Zaidel, and D. Egedi. **A Freely Available Wide Coverage Morphological Analyzer for English**. In *Proc. of the 14th International Conference on Computational Linguistics (COLING'92)*, pp. 950–954, August 1992.

[Mil90] George Miller. **WordNet: An On-line Lexical Database**. *International Journal of Lexicography*, Vol. 3, No. 4, 1990. (Special Issue).

[Pai91] Chris D. Paice. **A Thesaural Model of Information Retrieval**. *Information Processing & Management*, Vol. 27, No. 5, pp. 433–447, May 1991.

[Rad85] Roy Rada, Susanne Humphrey, and Craig Coccia. **A Knowledge-Base for Retrieval Evaluation**. In *Annual Proc. of the ACM (Association of Computing Machinery)*, pp. 360–367, 1985.

[Sal88] Gerard Salton and Christopher Buckley. **Term-Weighting Approaches in Automatic Text Retrieval**. *Information Processing & Management*, Vol. 24, No. 5, pp. 513–523, May 1988.

[Sat92] Satoshi Sato. **CTM: An Example-Based Translation Aid System**. In *Proc. of the 14th International Conference on Computational Linguistics (COLING'92)*, pp. 1259–1263, August 1992.

[Sum92] Eiichiro Sumita and Hitoshi Iida. **Example-Based Transfer of Japanese Adnominal Particles into English**. *IEICE TRANS. INF. & SYST.*, Vol. E75-D, No. 4, pp. 585–594, April 1992.

[Tur91] Howard Turtle and W. Bruce Croft. **Evaluation of an Inference Network-Based Retrieval Model**. *ACM Transactions on Information Systems*, Vol. 9, No. 3, pp. 187–222, July 1991.

[Voo93] Ellen M. Voorhees. **Using WordNet to Disambiguate Word Senses for Text Retrieval**. In *Proc. of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93)*, pp. 171–180, June 1993.

[Yar92] David Yarowsky. **Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora**. In *Proc. of the 14th International Conference on Computational Linguistics (COLING'92)*, pp. 454–460, August 1992.

[大野 81] 大野晋, 浜西正人. **角川類語新辞典**. 角川書店, 1981.

[国語研 64] 国立国語研究所. **分類語彙表**. 秀英出版, 1964.

[隅田 91] 隅田英一郎, 堤豊. **翻訳支援のための類似用例の実用的検索法**. 電子情報通信学会論文誌, Vol. J74-D-II, No. 10, pp. 1437–1447, October 1991.

[住田 96] 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, January 1996.

[松尾 91] 松尾比呂志, 内野一. **意味属性に基づくテキストベース検索方式**. 情報処理学会論文誌, Vol. 32, No. 9, pp. 1172–1179, September 1991.

[美馬 96] 美馬秀樹, 隅田英一郎, 飯田仁. **類似検索を用いた情報検索システム**. 言語処理学会第 2 回年次大会, A5-3, March 1996.