データベースシステム 125－34
（２００１．７．１８）

情報検索における単語間の関係の効果

松村 敦† 高須 淳宏‡ 安達 淳‡

† 図書館情報大学図書館情報学部
〒 305-8550 茨城県つくば市春日 1-2
Email: matsumur@ulis.ac.jp

‡ 国立情報学研究所
〒 101-8340 東京都千代田区一ツ橋 2-1-2
Email: {takasu, adachi}@nii.ac.jp

概要

従来から行なわれている文書検索手法は，出現単語とその統計的特徴量のみに頼ったものであるためその精度には限界がある．そこで我々は，単語間の関係として係受けを用いた手法と順序付共起関係を用いた手法の二つを開発し，単語間の関係の情報検索における効果の分析を試みた．評価実験には情報検索システム評価用テストコレクション NTCIR–1 及び NTCIR–2 を用いた．これにより，単語ベースの baseline の性能向上により単語間の関係の効果が小さくなることを示した．さらに，個別の問合せに対する分析を試み，本手法の有効性を議論した．

Effect of Relationships between Words on Information Retrieval

Atsushi MATSUMURA† Atsuhiro TAKASU‡ Jun ADACHI‡

†University of Library and Information Science
1-2, Kasuga, Tsukuba-shi, Ibaraki 305-8550, JAPAN
Email: matsumur@ulis.ac.jp

‡National Institute of Informatics (NII)
2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo 101-8340, JAPAN
Email: {takasu, adachi}@nii.ac.jp

Abstract

Conventional keyword based retrieval methods have obviously limitation on retrieval effectiveness, because they use only keywords and their statistical characteristics. We therefore developed two information retrieval methods using relationships between words. One is a method using dependency relationships between words and another is a method using the ordered co-occurrence information of words in a sentence. Through retrieval experiments with NTCIR–1 and NTCIR–2, which are Japanese test collections for information retrieval, we showed that the effect of using relationships between words on information retrieval is diminishing as the effectiveness of baseline is growing. We also discussed effectiveness of our methods, giving analysis of search results for some search topics.

1 はじめに

インターネットの普及と文書の電子化により，誰もが直接大量の電子文書へアクセスできるようになって来た．こうした状況から，有用な情報だけを効率良く手に入れる高度な情報検索手法への期待が高まっている．しかしながら，従来からの検索手法であるブール演算モデルやベクトル空間モデルではこれらの要求に答えることはできない．これらの手法の問題点の一つとして，例えば Term Frequency Inverted Document Frequency(TF–IDF) 法のように文に含まれる単語の統計的な情報のみしか利用しておらず単語間の関係を適切に扱えないことがあげられる．

情報検索に利用される単語間の関係は，構文的フレーズと統計的フレーズの二種類がある．構文的フレーズは自然言語処理により構文関係を解析した結果を利用するものであり，詳しい単語間の関係を表現できるが処理のコストが大きいことと構文解析の精度に問題が残る．一方，統計的フレーズは共起などの統計情報を利用し構文的フレーズを近似するものである．この手法では計算量的なコストの問題は改善されるが，正しいフレーズを抽出しているかは疑問が残る．このような問題に対して，Mitra 等は TREC-1 コレクション[1] 内の 25 文書以上に出現したすべての語の組を統計的フレーズと定義し，情報検索に利用し分析を行なった [2]．その結果，baseline の向上に伴いフレーズの検索精度への寄与は 1%程度に減少することを報告している．さらに，一単語を同時に利用する場合には構文的フレーズも統計的フレーズも検索精度への寄与は変わらないことを報告している．このような事実は Smeaton と Kelledy による追実験によっても裏付けられている [4]．しかしながら，どのような理由でこれらの現象が起こるのかについては分析が進んでいない．

このような観点から，我々は日本語における単語間の関係を使った情報検索手法の詳細な分析の必要性を感じ研究を行ってきた．これまで文献のタイトルを対象とした係受け関係を用いた情報検索手法の有効性を示し [7]，これを抄録検索における単語間の関係を利用した検索手法に拡張し，その有効性の検証を日本語情報検索用テストコレクション NTCIR–1[2] を用いて行なった [1]．

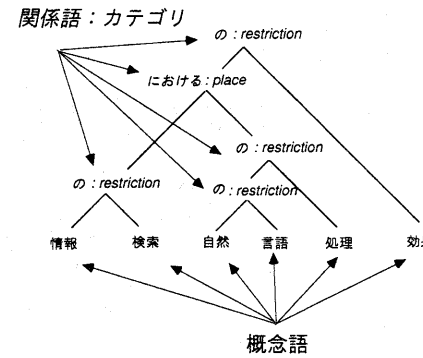

図 1: 構造化インデクスの例

本論文では，baseline となる単語ベースの得点付け関数の改善を行ない，これに伴う情報検索における単語間の関係の効果を見るために，日本語情報検索用テストコレクション NTCIR–1 および NTCIR–2[2] を用いて実験を行なった．以下では，はじめに本手法の概要を述べ，次に検索実験の結果と分析を示し，最後に結論と今後の課題について述べる．

2 単語間の関係を用いた検索手法

本研究で実現した単語間の関係を利用した検索手法は，係受け関係を用いた手法 *ST* と順序付共起関係を用いた手法 *CO* である．それぞれ，単語間の関係は一文内のみに限定している．

ST において単語間の係受け関係を表現する手法が，構造化インデクス手法である．これは，係受け関係を二分木の形で表現したものである．図 1 は‘情報検索における自然言語処理の効果’という文の構造化インデクスの例である．

構造化インデクスの枠組では，単語は概念語と関係語の二つのグループに分けられる．概念語は概念を表す語と定義し，構造化インデクスの葉の部分におかれる．一方，関係語は二つの概念語を関係付ける語と定義し，意味の類似性によってカテゴリに分類され，関係語とそのカテゴリは構造化インデクスの内部ノードにおかれる．このように構造化された文の内部表現を用いることにより単語間の係受け関係，すなわち構文的フレーズを用いた検索が可能と

[1] Text REtrieval Conference
http://trec.nist.gov/

[2] NACSIS Test Collection for IR Systems 1 (NTCIR–1)
NII-NACSIS Test Collection for IR Systems 2 (NTCIR–2)
http://research.nii.ac.jp/ntcir/

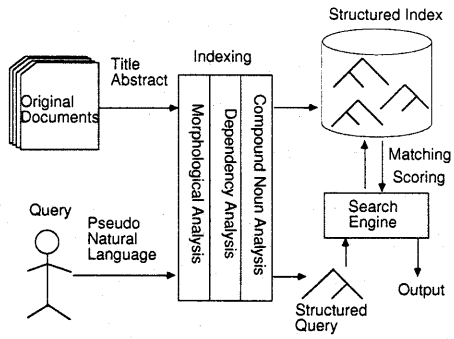

図 2: 処理の流れとシステムの全体像

なる．

一方，係受け解析のコストを下げつつ係受け関係を近似することを目指した手法が，概念語間の順序付共起関係を用いる手法 *CO* である．この手法は，統計的フレーズの一種を用いた検索手法と見ることができる．*CO* は，構造化インデクスから概念語の並びだけを取り出して利用することで実現できる．実際には *CO* に適したインデクス構造を実現する必要があるが，本研究では比較実験を行ないやすくするため構造化インデクス上に *CO* と *ST* の二つの検索手法を実現した．

以下ではインデクス作成については，構造化インデクスの作成方法のみを説明するが，検索と得点付けに関しては *ST* と *CO* に分けて説明する．図 2 に *ST* および *CO* を実現したシステムの全体像と処理の流れを示す．

2.1 インデクス作成

構造化インデクスの作成は形態素解析，係受け解析，複合名詞解析の三段階を経る．

形態素解析　はじめに各文を概念語と関係語に切り分ける．本手法では，概念語は名詞，形容詞，副詞および複合名詞の構成要素，関係語は前置詞，助詞，助動詞，動詞とその組合せと定義する．文内の概念語と関係語を同定するために，形態素解析結果の品詞情報と人手で集めた関係語のデータベースを利用する．形態素解析には茶筅[3]を用いた．これらの関

[3]日本語形態素解析ソフト "茶筅"
http://catus.aist-nara.ac.jp/lab/nlp/chasen.html

表 1: 関係語のカテゴリと主な関係語

カテゴリ名	主な関係語
restriction	の，な，された，される
place	における，での，上の
method	による，を用いた，に基づく
and	と，および，ならびに，も

係語を分類するカテゴリとして 18 種類を定義している．表 1 に，主なカテゴリとそこに分類される主な関係語を示した．関係語の同定で 18 カテゴリに分類できないものについては，other カテゴリを与える．

係受け解析　概念語間の係受け関係の付与には，一文内の関係語の並びをテンプレートとして用いる．例えば '情報検索における自然言語処理の効果' という文は 'A における B の C' というテンプレートに一致する．ここで，A，B，C は概念語またはその並びである．各テンプレートには事例分析によりあらかじめ係受けパタンを対応させてあり，これを用いて係受け判定を行なう．関係語を二つ以上含むテンプレートで係受けパタンが一対一に対応しない場合があるが，このような曖昧性の残る場合は文末に一般語があるかどうかで係受けを付与する．ここでの一般語の定義は，'研究' や '効果' といった重要度が低く文全体を受ける性質を持った概念語である．本手法では 53 の一般語を定義している．

これらのテンプレートは関係語の数が 2 個と 3 個の場合を作成しているが，それ以上の関係語がある場合にはいくつかの経験則を用いて文を分割してから，上述のテンプレートマッチングを行なう．こうすることによって全体では係受けパタンの付与が失敗した場合にも部分的に正しい係受けが付与される可能性が高くなり，係受け付与の失敗による検索精度の劣化が押えられる．

複合名詞解析　複合名詞も，その構成要素間の係受け関係を考慮することによって構造化インデクスの枠組に組み込むことができる．はじめに，複合名詞の構成概念語間に適当な関係語を補って文に変換する [6, 5]．本手法では基本的に概念語間に 'の' を補う．これは，複合名詞中の概念語と関係語の共起関係を調べた結果，関係語として 'の' を補うことが統

計的に最も正しいことが判明したためである [8].

一方，係受けの付与は二つの単語が複合名詞中で隣りあって出現する回数を表す単語 bigram 統計を利用して行なう．単語 bigram は，十分なサンプルをとれば複合名詞中の二つの概念語の間の結び付きの強さを表すものとして利用可能であると考えられる．本システムでは，13615 件の文献タイトルから抽出した 60507 個のうち，10 回以上出現した 814 個の bigram データを利用した．

2.2 検索と得点付け

問合せは文書タイトルのような擬似的自然言語文とし，インデクス作成と同様の処理を行ない構造化する．これにより，*ST* における検索は問合せの二分木と文書の二分木の集合とのマッチングを取ることになる．一方，*CO* はこれらの二分木中の概念語の順序付共起関係のみを用いてマッチングをとる．

2.2.1 文書の総得点

文書 d の総得点 S_d は，単語ベースの TF–IDF による単語得点 SW_d と単語間の関係による得点 SR_d の得点を線形結合で組み合わせたものであり，それぞれの重みを xw，xr として式 (1) によって計算される．

$$S_d = xw \times SW_d + xr \times SR_d \qquad (1)$$

2.2.2 単語に対する得点

文書 d に対する単語得点 SW_d は，問合せに含まれる概念語と一致した概念語に対して，式 (2) による TF–IDF 得点を与えることによって得られる．

$$SW_d = \sum_{j=1}^{n} tfidf(C_j) \qquad (2)$$

ここで，$tfidf(C_j)$ は TF–IDF ベースの重み付け関数である．後に述べる実験では，この関数として以下の二種類を利用した．

$$tfidf(C_j) = \log(tf(C_j)+1)\log(\frac{N_{all}}{df(C_j)}) \qquad (3)$$

$$tfidf(C_j) = \frac{tf(C_j)}{tf(C_j)+1}\log(\frac{N_{all}}{df(C_j)}) \qquad (4)$$

ここで，各変数は以下のとおりである．

C_j ： 問合せ中の j 番目の概念語
n ： 問合せ中の概念語の総数
$tf(C_j)$ ： 文書中の C_j の頻度
$df(C_j)$ ： C_j を含む文書数
N_{all} ： 総文書数

式 (3) はこれまで利用してきた基本的な式であり，式 (4) は TREC などで良い結果を出している Robertson らによる BM11[3] を単純化した式である．

2.2.3 概念語間の関係に対する得点

問合せ及び文書中の各文は，二つの概念語と関係語の三つ組で表される係受け関係の集合とみる．この三つ組は *ST* における得点付けの最小単位である．一方，*CO* における得点付けの最小単位は，この三つ組から関係語を除いた二つの概念語の順序付共起関係である．文書 d の概念語間の関係に対する得点 SR_d は，文書の各要素 b (タイトル (t)，アブストラクト (a) など) に対して与えられる得点を Sr_b，その重みを w_b とし，式 (5) によって定義する．

$$SR_d = \sum_b w_b \times Sr_b \qquad (5)$$

***ST* における三つ組の得点** *ST* では，問合せと文書要素で三つ組を構成する二つの概念語が一致したものに対して得点を与える．この得点は以下に定義する二つの要素を考慮して計算する．

一つは，三つ組間の一致レベルである．これは，同じ二つの概念語を持つ二つの三つ組間の類似度を，関係語の違いによって以下の三つのレベルで評価し得点を与えるものである．

1. *Exact Match*: 二つの関係語が同じ．
2. *Category Match*: 二つの関係語は異なるが，そのカテゴリが等しい．
3. *Wild Match*: 二つの関係語もそのカテゴリも異なる．

もう一つは，三つ組の重要度である．ここでは三つ組の重要度は二つの概念語の重要度で近似し，それぞれの概念語の IDF の積として定義する．さらに，一般語によるノイズを考慮に入れ，三つ組に一般語が含まれている場合には重要度を 0 とする．

以上の一致条件を考慮して，三つ組 TR（左側の概念語を C_l，右側の概念語を C_r とする）の得点は

式 (6) で与える.

$$Sd(TR) = LD(TR)ID(C_l, C_r) \qquad (6)$$

ここで，式中の関数はそれぞれ以下のとおりである.

$$\begin{aligned} LD(TR) &: TR\text{の一致レベルに応じた重み} \\ &= \begin{cases} we & \text{if Exact Match} \\ wc & \text{if Category Match} \\ ww & \text{if Wild Match} \end{cases} \\ ID(C_l, C_r) &: \text{文書集合中での } TR \text{ の重要度} \\ &= idf(C_l)idf(C_r)gw(C_l)gw(C_r) \quad (7) \\ idf(C) &= \log(\frac{N_{all}}{df(C)}) \\ gw(C) &= \begin{cases} 0 & C\text{が一般語の場合} \\ 1 & \text{それ以外} \end{cases} \end{aligned}$$

CO における二つ組の得点　*CO* では，問合せと文書要素で，概念語の順序付共起関係が一致した二つ組に対して得点付けを行なう．二つ組の得点はその重要度のみとし，*ST* と同様に二つの概念語の IDF の積と一般語を考慮に入れる．すなわち，式 (7) が *CO* の二つ組の得点式である．これは式 (6) において，一致レベルを表す得点要素 $LD(TR)$ による重みを全て 1 としたものと等価である.

概念語間の関係に対する総得点　文書の各要素中でマッチした全ての得点単位（*ST* の場合は三つ組，*CO* の場合は二つ組）の得点を使い，概念語間の関係に関する問合せと文書要素 b との類似度を以下の式 (8) で与える.

$$Sr_b = \sum_{j=1}^{m} \max\{Sd(TR) : TR \in Rel_j\} \qquad (8)$$

ここで，Rel_j は問合せ中の j 番目の三つ組（または二つ組）で，m はその総数である．関数 max は問合せに含まれるそれぞれの三つ組（または二つ組）に対して，文書要素中で最も高得点でマッチしたものを選択する関数である．この得点付けによって，一つの三つ組（または二つ組）が繰り返しマッチすることによって得点が高くなり過ぎたり，いくつかの文に分散している重要な三つ組（または二つ組）を見落としたりすることを防ぐことができる [1].

3 実験と評価

3.1 実験条件

評価実験には情報検索システム評価用テストコレクション NTCIR–1(本格版) および NTCIR–2(本格版) を利用した．これは，日本語の単言語検索および日英の言語横断検索に対するものとしては最大規模である．本研究では，日本語の単言語検索に関する部分を用いて実験と評価を行なった．コレクションには約 73 万件の文書，49 件の検索課題および，それに対する適合度判定セットが含まれている．適合度判定は，高適合 (S)，適合 (A)，部分的適合 (B)，不適合 (C) の 4 段階判定で行なわれている.

本研究では，検索課題を一文で表現した「検索要求」を問合せとして用い適合度判定 S と A の文書の両方を正解として評価した．検索対象は文書中のタイトル及びアブストラクトとし，式 (5) において $w_t = w_a = 1$，それ以外の重みは 0 とした．また，$xw(0 \leq xw \leq 1)$ を独立したパラメタとして $xr = 1 - xw$ と定義する．したがって，本手法は $xw = 1$ の時に TF–IDF を基にした baseline と等価となる．また，*Exact Match* の重み we と *Category Match* の重み wc をともに 1 とし，*Wild Match* の重み $ww(0 \leq ww \leq 1)$ をシステムを特徴付けるもう一つのパラメタとして利用する．$ww = 1$ の時は，*ST* と *CO* は等価となる.

3.2 実験結果と分析

単語得点の式として，式 (3) を用いて行なった検索実験の結果を表 2 にまとめた．*ST* と *CO* の結果は，二つのパラメタ xw と ww を最適化して得られた結果である．この結果から，式 (3) による得点付けを baseline とした場合には *ST* や *CO* の効果は 10%近くあることが分かる.

また，両手法による平均適合率の baseline からの差を問合せ毎に示したのが図 3 である．これを見ると非常に多くの問合せで精度の大幅な改善が達成されていることが分かる．NTCIR–1 での評価で有効であった本手法が NTCIR–2 を含めた評価においても有効であることが示されたと言える.

一方，式 (4) を用いた実験の結果が表 3 である．baseline を比較しても 27.6%もの精度の向上が得られ，式 (4) が非常に有効な重み付けであることが分

表 2: 式 (3) を baseline とした時の 11 点平均適合率

method	11-pt. ave.	gain
baseline	0.2384	–
CO	0.2615	9.7%
ST	0.2640	10.7%

表 3: 式 (4) を baseline とした時の 11 点平均適合率

method	11-pt. ave.	gain
baseline	0.3042	–
CO	0.3081	1.28%
ST	0.3085	1.41%

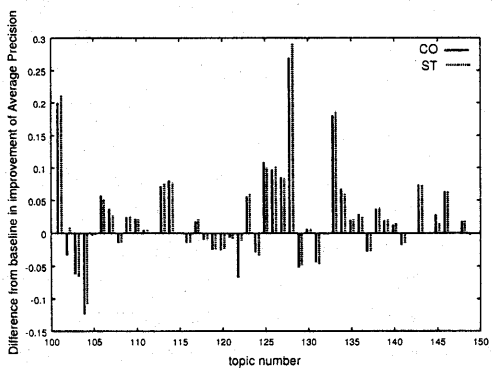

図 3: 平均適合率の baseline との差，式 (3) の場合

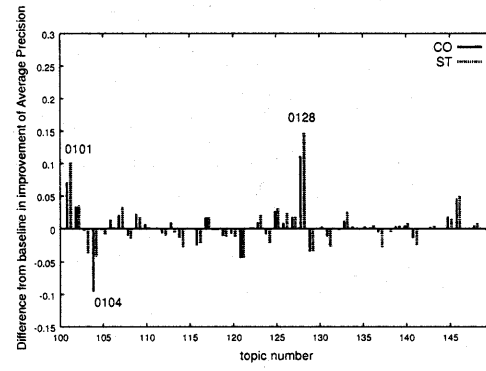

図 4: 平均適合率の baseline との差，式 (4) の場合

かった．しかしながら，これに伴って *CO* 及び *ST* の寄与は 1%あまりと非常に小さくなり，両者の間の差もほとんどなくなった．この結果は，Mitra らの実験 [2] を日本語のコレクションにおいて確かめた形となっている．すなわち，baseline の向上に伴って単語間の関係を考慮する効果は小さくなるというものである．また，構文的フレーズを利用する手法に相当する *ST* と統計的フレーズを利用する手法に相当する *CO* との差がほとんど無くなっている点も Mitra らの結果と対応している．

次に，個々の問合せの baseline に対する平均適合率の差を図 4 に示す．全体の平均が示す通り，多くの問合せではほとんど精度の向上はない．しかし，いくつかの問合せでは大きく精度が向上している場合や，逆に大きく精度が落ちている場合が見られる．以下ではこのうち三つの特徴的な問合せ 0128，0101，0104 について，baseline と *CO* との比較という観点から分析した結果を示す．

まず，各問合せに対して baseline によって検索された上位 500 件の文書を TF–IDF 得点を横軸に，*CO* による関係得点を縦軸にとりプロットした．図 5，図 6 及び図 7 は，それぞれ問合せ 0128，0101 及び 0104 の場合のプロットである．各文書は正解判定の違いによって，グラフ左上に示されたマークでプロットされている．式 (5) の変数 xw に対して，グラフ上で傾き $xw/(xw-1)$ の直線を原点に向かって動かしたときに通過した文書を順番に出力したものがこのパラメタにおける検索出力になる．したがって，正解文書がグラフの上方に，不正解文書が下方にプロットされるような得点付けが検索精度を向上させる．

適切な共起の利用 (問合せ 0128 の場合) 最も大きな精度向上を達成できた問合せは，0128 の「コアグラーゼ陰性ブドウ球菌による感染症」であった．図 5 を見ると，正解と不正解の分離がはっきりしている．最も高い関係得点を得ている文書 1 はタイトルとアブストラクトに「コアグラーゼ陰性ブドウ球菌感染症」という表現が存在した．また，文書グループ 2 は「コアグラーゼ陰性ブドウ球菌」がタイトルとアブストラクトに，文書グループ 3 は同じ表現がアブストラクトのみに出現していた．これに対して，グループ 4 は「黄色ブドウ球菌」についての文書であった．「黄色ブドウ球菌」は「コアグラーゼ陽性ブドウ球菌」であり，「コアグラーゼ陰性ブドウ球菌」と対をなす概念であるため不正解となる．「黄色ブドウ球菌」には「コアグラーゼ陽性」という意味が含

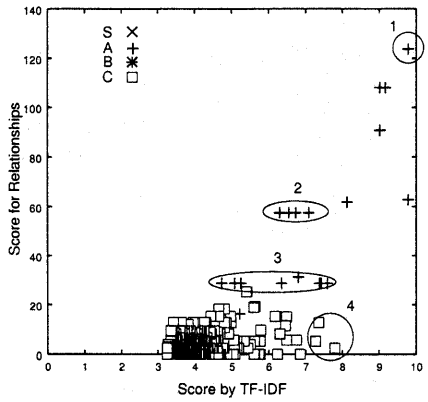

図 5: 文書の単語点と関係点の関係 (問合せ 0128)

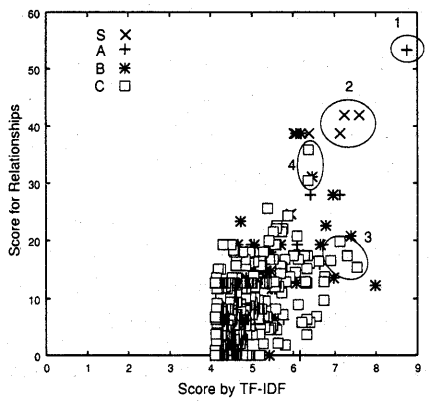

図 6: 文書の単語点と関係点の関係 (問合せ 0101)

まれているため「コアグラーゼ」という単語と一文内で共起しにくいと考えられる.

この問合せの場合には「コアグラーゼ」と「ブドウ球菌」という共起を適切に利用した効果が検索精度に現れていると言える.

言い替え, 照応, 否定形の影響 (問合せ 0101) 問合せ 0101 の「遺伝子工学的手法による B 型肝炎ワクチンの開発」も精度が大きく向上している. 図 6 におけるグループ 1 の文書には「遺伝子工学的手法」と「B 型肝炎ワクチン」が共起していた. グループ 2 とグループ 3 を比較すると同程度の単語得点でありながら正解文書のグループ 2 のほうが関係得点が高い. これはどちらのグループも遺伝子工学的手法によるワクチンの開発について述べているが, 対象

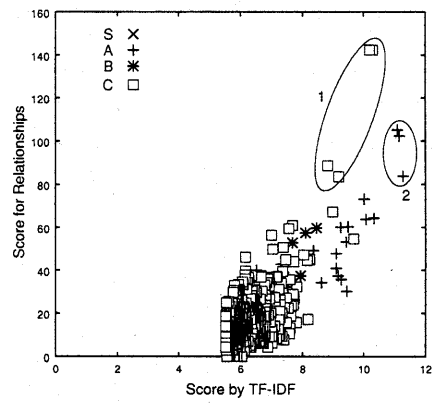

図 7: 文書の単語点と関係点の関係 (問合せ 0104)

とする病気がグループ 2 は「B 型肝炎」であるのに対して, グループ 3 はそれ以外の「ジフテリア」や「破傷風」であったためである. グループ 3 の不正解文書中には低頻度語の「ワクチン」が頻出しているが, B 型肝炎との共起が無いことによって得点を下げることに成功している.

一方, グループ 2 が S 判定ながら得点が低い理由は,「遺伝子工学的手法」が「遺伝子操作」と言い替えられていることや,「遺伝子操作」を指している「この方式」を字面でしか理解できないことがあげられる. これら言い替えと照応の理解が可能となれば, さらに精度が向上すると考えられる.

精度の劣化のもう一つの原因は, グループ 4 の不正解文書である. これらは「遺伝子工学的手法」を用いた「非 A 非 B 型肝炎」に関する文書であった. 「非」などの否定形の挿入により, 内容は全く逆になるが共起だけを見た場合は同等である. これは, 表現の一部の違いが大きな違いとなる場合への対処が必要であることを示している.

文書要素の違いの考慮の影響 (問合せ 0104) 問合せ 0104 の「モノクローナル抗体を利用した肺小細胞癌の診断と治療」は精度が大きく落ちているが, この場合の原因は図 7 のグループ 1 の文書である. これらは「肺非小細胞癌」についての文書であり, 0101 の場合と同様に「非」という否定形の挿入という問題であった.

一方で, 正解にもかかわらず関係得点が低いグループ 2 には, タイトルが「肺癌」とあるべきところを「肝癌」と間違っている文書, タイトルは「キ

メラ抗体」としているが研究の途中経過報告であるため，アブストラクトは「モノクローナル抗体」のみについてしか書かれていない文書などがあった．このようにタイトルとアブストラクトが一致しない文書が悪影響を及ぼしているが，これはタイトルとアブストラクトを別々に扱う場合には注意を要することを示している．

4 まとめと今後の課題

本論文では，単語間の関係を利用した二つの情報検索手法を提案し，情報検索システム評価用テストコレクション NTCIR–1 及び NTCIR–2 を用いて単語間の関係の情報検索における効果を評価した．二つの検索手法とは，係受け関係を用いた手法 *ST* と単語の順序付共起関係を用いた手法 *CO* である．

TF–IDF ベースの baseline として，本研究でこれまで利用して来た一般的な重み付け関数と Robertson らの BM11 の単純化による関数の二種類で実験し，それぞれの baseline 対する *CO* および *ST* の効果を評価した．従来からの関数を用いた場合の検索実験では，検索精度は低いがそれに対する *CO* 及び *ST* の 11 点平均適合率は 10%程の向上となった．一方，BM11 を単純化した関数による baseline を利用すると検索精度は大幅に向上し，本手法の *CO* と *ST* の寄与は 1%程度と非常に低くなった．このような baseline の向上に伴うフレーズの効果の減少は Mitra らによって報告されているが，本実験結果は日本語においても同様の傾向があることを示すものである．また，*ST*，*CO* の提案手法間の差は極わずかであり，現状ではインデクス作成や検索のコストを考えて *CO* が実用的な手法と言わざるを得ない．

しかしながら，個々の問い合わせを見るとその振舞は様々であった．いくつかの問い合わせを分析した結果，フレーズが一部を除いてほぼ同じような表現である場合に，それらの違いを明確に判定できることが精度の向上につながることが分かった．これは，より厳密な係受け解析を行なうことの正当性を主張するものであり，*ST* の有効性を示唆するものである．また，言い替えや照応解析などが可能となれば，さらに本手法の効果も上がる可能性も見られた．

問い合わせの中には *CO* と *ST* で異なる振舞するものもあり，今後はこれらを中心により詳細な分析をすすめ，本手法の適用方法についての研究を行なう必要がある．

謝辞

本研究は日本学術振興会未来開拓事業 JSPS-RFTF96P00602 の援助を受けた．

また，NTCIR–1(本格版) 及び NTCIR–2(本格版) を使用した．これは国内学会[4] の提供する学会発表データベースの一部，および科研費データベースの一部を利用して国立情報学研究所によって作成された．

参考文献

[1] Atsushi Matsumura, Atsuhiro Takasu, and Jun Adachi. The effect of information retrieval method using dependency relationship between words. In *RIAO'2000 Conference Proceedings*, Vol. 2, pp. 1043–1058, April 12–14 2000.

[2] Mandar Mitra, Chris Buckley, Amit Singhal, and Claire Cardie. An analysis of statistical and syntactic phrases. In *Proceedings of the RIAO'97 Conference*, pp. 200–214, June 1997.

[3] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of ACM SIGIR'92*, pp. 232–241, 1992.

[4] Alan F. Smeaton and Fergus Kelledy. User-chosen phrases in interactive query formulation for information retrieval. In *Proceedings of the 20th BCS-IRSG Colloquium*, pp. 200–214, April 1998.

[5] 石崎雅人. 複合名詞における語と語の関係について. 第 37 回情報処理学会全国大会, 1988. 3C–2.

[6] 宮崎正弘. 係受け解析を用いた複合語の自動分割法. 情報処理学会論文誌, Vol. 25, No. 6, pp. 970–979, 1984.

[7] 松村敦, 高須淳宏, 安達淳. 単語間の係受け関係を用いた情報検索手法の評価. 情報処理学会論文誌：データベース, Vol. 41, No. SIG1(TOD5), pp. 22–30, 2000.

[8] 池田和幸, 安達淳. 単語間の係受け情報を用いた文献検索手法. 第 54 回情報処理学会全国大会, 1997. 4K–2.

[4] http://research.nii.ac.jp/ntcir/acknowledge/thanks1-ja.html