位置情報と分野情報を用いた情報検索

村田　真樹[†]　馬　　青[†]　内元　清貴[†]
小作　浩美[†]　内山　将夫[†]　井佐原　均[†]

われわれの情報検索の方法では基本的に，確率型手法の一つの Robertson の 2-ポアソンモデルを用いている．しかし，この Robertson の方法では検索のための手がかりとして当然用いるべき位置情報や分野情報などを用いていない．それに対しわれわれは位置情報や分野情報などをも用いる枠組を考案した．IREX のコンテストでは，この枠組に基づくシステムを二つ提出していたが，記事の主題が検索課題に関連している記事のみを正解とする A 判定の精度はそれぞれ 0.4926 と 0.4827 で，参加した 15 団体，22 システムの中では最もよい精度であった．本論文ではこのシステムの詳細な説明を行なうとともに，種々のパラメータを変更した場合の詳細な対照実験を記述した．この対照実験で位置情報や分野情報の有効性を確かめた．

キーワード:　*2-poisson* モデル，位置情報，分野情報，ラティス

Information Retrieval Using Location and Category Information

Masaki Murata[†] , Qing Ma[†] , Kiyotaka Uchimoto[†] , Hiromi Ozaku[†] ,
Masao Utiyama[†] and Hitoshi Isahara[†]

Robertson's 2-poisson model to retrieve information does not use location information and category information. We constructed a framework using location information and category information in a 2-poisson model. We submitted two systems based on this framework at the IREX contest. Their precision in the A-judgement measure was 0.4926 and 0.4827, the highest values among the 15 teams and 22 systems that participated in the IREX contest. This paper describes our systems and comparative experiments when various parameters are changed. These experiments confirmed the effectiveness of location information and category information.

KeyWords:　*2-poisson model, location information, category information, lattice*

1　はじめに

近年の WWW (World Wide Web) などのインターネットの発展や電子化文書の増加により情報検索 (徳永 1996; 小川 1996; 藤田 1999) の研究は盛んになっている．これを背景に日本で情報検索コンテスト IREX が行なわれた．われわれはこのコンテストに二つのシステムを提出していたが，記事の主題が検索課題に関連している記事のみを正解とする A 判定の精度はそれぞれ 0.4926 と 0.4827 で，参加した 15 団体，22 システムの中では最もよい精度であった．本論文は，この二

† 郵政省 通信総合研究所, Communications Research Laboratory, Ministry of Posts and Telecommunications

つのシステムの詳細な説明と，これを用いた詳細な実験結果を記述するものである．

われわれの情報検索の方法では基本的に，確率型手法の一つの Robertson の 2-ポアソンモデル (Robertson and Walker 1994) を用いている．しかし，この方法では検索のための手がかりとして当然用いるべき位置情報や分野情報などを用いていない．それに対しわれわれは 2-ポアソンモデルにおいて位置情報や分野情報，さらに種々の詳細な情報などをも統一的に用いる枠組を考案し，これらの情報の追加により精度向上を実現できることを実験により確かめている．また，2-ポアソンモデルを用いる際にはまず，どのようなものをキーワードとするかを定める必要がある．本研究では，キーワードの抽出方法について 4 つのものを示し，それらの比較実験を行なっている．

2　情報検索の方法

2.1　問題設定

本研究での情報検索の問題設定は，日本で開催された情報検索コンテスト IREX(Sekine and Isahara 1999) のものと全く同じである．本研究で検索の対象とするデータは，IREX で用いられた毎日新聞 94 年 95 年の二年分の新聞記事データである．このデータに対して日本語文で記述された検索要求を満足する文書を検索する．

日本語文で記述された検索要求の例を以下に示す (IREX の予備試験の課題より)．

<TOPIC>
<TOPIC-ID>1001</TOPIC-ID>
<DESCRIPTION> 企業合併 </DESCRIPTION>
<NARRATIVE> 記事には企業合併成立の発表が述べられており、その合併に参加する企業の名前が認定できる事。また、合併企業の分野、目的など具体的内容のいずれかが認定できる事。企業合併は企業併合、企業統合、企業買収も含む。 </NARRATIVE>
</TOPIC>

ここで TOPIC-ID で囲まれた数字は設問番号を意味し，DESCRIPTION が検索課題を端的に示すフレーズ，NARRATIVE が検索要求を厳密に規定する説明文となっている．これをシステムがうけとり，例えば以下のような記事を検索対象としてかえせばよい．

<DOCNO>950217091</DOCNO>
<SECTION> 経済 </SECTION>
<AE> 無 </AE>
<WORDS>288</WORDS>
<HEADLINE> キグナス石油精製を東燃が１００％子会社化 </HEADLINE>
<TEXT>
東燃は十六日、系列のキグナス石油精製（資本金十億円、本社・川崎市、森利英社長）を

一〇〇％子会社化すると発表した。同社は東燃が七割、ニチモウが三割出資しており、東燃はニチモウが所有する全株式六十万株を百二十五億円で買収する。

東燃は石油精製専業大手で、設備シェアは一九九三年度末で八％。キグナス石油精製を加えると九・四％にシェアがアップする。

ニチモウは山口県の工場閉鎖などに伴う経費ねん出のため株式譲渡を決断した。

石油業界は来年春の特定石油製品輸入暫定措置法（特石法）廃止をにらみ、コスト削減と効率化を進めており、グループ企業統合を含む再編の動きがいよいよ本格化してきた。

</TEXT>

</DOC>

この記事のように SECTION には経済面などの新聞の誌面情報が， HEADLINE には記事のタイトルが， TEXT には記事の本文がある．IREX では，各設問に対し 300 個の記事を順位つきで提出することになっており，各設問おおよそ 50 個程度ある正解を上位にたくさん含むような結果を提出すれば，よりよい精度が得られるようになっている．情報検索の精度の評価には， TREC の trec_eval というツール (trec_eval 1992) を利用し，コンテストの評価では trec_eval で得られる評価値のうち R-Precision という評価値 (正解記事数分だけ検索した時に正解の記事が含まれている割合) が用いられている．

2.2 検索方法の概略

先に述べたとおり，われわれの情報検索の方法では基本的に，確率型手法の一つの Robertson の 2-ポアソンモデル (Robertson and Walker 1994) を用いる[1]．Robertson らの方法とは，各記事毎に以下の式で与えられる Score を算出し， Score の上位のものを検索結果として出力する方法である．(以下の Score(d) は記事 d の Score．)

$$\mathrm{Score(d)} = \sum_{\substack{\text{キーワード t}\\ \text{で和をとる}}} \left(\frac{\mathrm{TF(d,t)}}{\frac{\mathrm{length(d)}}{\Delta} + \mathrm{TF(d,t)}} \times \log \frac{\mathrm{N}}{\mathrm{DF(t)}} \right) \tag{1}$$

ただし，ここでのキーワードとは検索要求に出現していたキーワードである．$TF(d,t)$ はキーワード t の記事 d での出現回数である．$DF(t)$ は全データベースでのキーワード t が出現している記事の数である．N は全データベースに存在する記事の数．$length(d)$ は記事 d の長さ (文字列単位) である．Δ は全データベースでの記事の長さの平均である．また，この式の $\frac{TF(d,t)}{\frac{length(d)}{\Delta} + TF(d,t)}$ を TF に関する項として TF 項，$log\frac{N}{DF(t)}$ を IDF(DF の逆数) に関する項として IDF 項と呼ぶことにする．この式の TF の項が一般のベクトル空間法で重みの一部として

1 IREX のコンテストでは上位三団体のシステムはすべて Robertson らの方法に基づくものであったので，この手法を検索の基本部分に使用することは現状では最善と思われる．また，この手法の有効性について文献 (村田, 内元, 小作, 馬 1999) においてより詳しく論じている．

用いられる TF と異なり，$\frac{TF}{\frac{length}{\Delta}+TF}$ のようになっている．Score をキーワードによる加点として扱うとき，TF の影響が大きい場合だと一つでも TF の値が大きいキーワードがあれば，他のキーワードがほとんど存在していないという関連性が低いときでも十分大きな得点を取ってしまう．この式はこのことを防ぐのに役に立っている．この式では TF が無限になってもたかだか 1 の値を持つにすぎない．このため全キーワードがまんべんなく評価されることになる．さらに，$\frac{length}{\Delta}$ という複雑な部分があるが，これは長い記事ほど TF の値が大きくなるのでそれを補正するための項である．

われわれの方法は，この式 (1) にいくつかの補強項をつけたものであり，以下の式で表現される．

$$\mathrm{Score(d)} = K_{分野}(d)\left\{\sum_{キーワードt}(TF項(d,t) \times IDF項(t) \times K_{詳細}(d,t) \times K_{位置}(d,t)) + \frac{length(d)}{length(d)+\Delta}\right\} \quad (2)$$

この式の TF 項 IDF 項 は，式 (1) と同じである．この式の $\frac{length(d)}{length(d)+\Delta}$ $(= K_{記事長}(t))$ は記事長が長いほど値が大きくなる項で，他の情報がまったく同じならば記事長が長ければ長いほど検索要求を含みやすいと考えて作ったものである．$K_{分野}$，$K_{詳細}$，$K_{位置}$ は精度向上のために追加した補強項である．$K_{分野}$ は新聞の紙面情報 (分野情報) を利用する項で，$K_{位置}$ は記事中でのキーワードの位置で重みを変更するものである．タイトルにあれば大きい値とし，記事中での位置が最初のものを加点し，後ろのものを減点するということをしている．$K_{詳細}$ は，さらに詳細な項でキーワードが固有名詞ならば加点したり，キーワードが「事」「認定」「記事」「言及」などの不要な単語の場合減点したりする項である．次節ではこれらの補強項の詳細な説明を行なう．

2.3　補強項の説明

式 (2) のようにわれわれは補強項として $K_{位置}$，$K_{分野}$，$K_{詳細}$ の三つを用いている．ここではこれらを詳細に説明する．

(1)　位置情報の利用 ($K_{位置}$)

特に新聞記事でいわれることだが，タイトルや記事の最初の文はその記事のおおまかな内容を示すことが多い．このため，そのような位置に現れるキーワードを重視することで情報検索の精度を向上させることができる (新谷, 角田, 大石, 長尾 1997)．この $K_{位置}$ はそのためのもので，記事中でそのキーワードが初めて出現している位置で重みを変更するものである．タイトルにあれば大きい値とし，記事中での位置も最初のものを加点し，後ろのものを減点するということをする．この項は以下の式で表される．

$$\mathrm{K}_{\text{位置}}(\mathrm{d},\mathrm{t}) = \begin{cases} k_{\text{位置},1}\ (\text{キーワード}t\text{が記事}d\text{でタイトルに出現}) \\ 1 + k_{\text{位置},2}\dfrac{(length(d) - 2 * P(d,t))}{length(d)}\ (\text{それ以外}) \end{cases} \quad (3)$$

この式で length(d) は記事 d の長さで P(d,t) はキーワード t の記事 d での位置を意味する．$k_{\text{位置},1}$，$k_{\text{位置},2}$ は実験で定める定数である．一記事中に同じキーワードが複数出現する場合は最も初めに出現したもののみを用いる．

(2) 分野情報 (紙面情報) の利用 ($K_{\text{分野}}$)

$K_{\text{分野}}$ は新聞の紙面情報 (分野情報) を利用する項である．これは，関連性フィードバック (Salton and Buckley 1997) のようなことをするもので，一度この項を 1 として検索を行ない，その検索結果における上位 100 個において紙面情報の統計をとり，その統計結果に基づき上位に出現しやすい面を特定し，それと同じ面に書いてある記事の得点を増加させることで Score を再計算するものである．例えば，一回目の検索で上位 100 個には経済面が集中していたとすると，経済面の記事には加点し，そうでない記事は減点するといったことを行なう．この項 ($K_{\text{分野}}$) は以下の式で表される．

$$\mathrm{K}_{\text{分野}}(\mathrm{d}) = 1 + \mathrm{k}_{\text{分野}}(\text{割合A(d)} - \text{割合B(d)})/(\text{割合A(d)} + \text{割合B(d)}) \quad (4)$$

だだし，割合$A(d)$は一回目の検索結果の上位 100 個における記事 d が該当する面の割合[2]で，割合$B(d)$は全記事での記事 d が該当する面の割合を意味する．この式の値は，割合Aが大きく (該当記事の面が一回目の検索結果でよく出現していて)，割合Bが小さい (該当記事の面が全記事ではそれほど出現していない) 場合に大きくなるようになっている．$k_{\text{分野}}$ は実験で定める値である．

(3) その他の情報の利用 ($K_{\text{詳細}}$)

$K_{\text{詳細}}$ は，さらに詳細な項でキーワードが固有名詞ならば加点したり，キーワードが「事」「認定」「記事」「言及」などの不要な単語の場合減点したりするもので，以下の式によって表される．(本節では表記の簡単化のため，記事，キーワード用の変数 d,t を省略して記述している．)

$$\begin{aligned} \mathrm{K}_{\text{詳細}} \quad = \quad & K_{\text{タイトル}} \times K_{\text{固有}} \times K_{\text{など}} \times K_{\text{数字}} \\ & \times K_{\text{ひらがな}} \times K_{\text{neg}} \times K_{\text{不要語}} \end{aligned} \quad (5)$$

この式の各項の説明を以下に記述する．

- $K_{\text{タイトル}}$

2 このときの割合の算出は，上位のものに重みを傾斜的に付加している．順位 x の記事に対して (150 - x)/100 の重みを頻度にかけてから割合の算出を行なっている．

該当するキーワードが検索要求のタイトル DESCRIPTION から得られたものの場合は $k_{タイトル}$ の値とし，そうでない場合 1 とする．この項は検索要求のタイトル DESCRIPTION から得られたキーワードを重要と考えて加点するためのものである．

- $K_{固有}$
  該当するキーワードが固有名詞の場合，$k_{固有}$ の値とし，そうでない場合 1 とする．検索要求から得られたキーワードが固有名詞ならば，そのキーワードを重要と考えて加点する．
- $K_{など}$
  該当するキーワードが検索要求において「など」の直前にあった場合，$k_{など}$ の値とし，そうでない場合 1 とする．検索要求から得られたキーワードが検索要求において「など」の直前にあった場合，特殊化されたキーワードであると考えて固有名詞と同様に加点する．
- $K_{数字}$
  該当するキーワードが数字だけで構成されている場合，$k_{数字}$ の値とし，そうでない場合 1 とする．検索要求から得られたキーワードが数字だけで構成されている場合は情報量が少なくあてにならないキーワードであると考えて減点する．
- $K_{ひらがな}$
  該当するキーワードがひらがなだけで構成されている場合，$k_{ひらがな}$ の値とし，そうでない場合 1 とする．検索要求から得られたキーワードがひらがなだけで構成されている場合は情報量が少なくあてにならないキーワードであると考えて減点する．
- K_{neg}
  該当するキーワードが NEG のタグで囲まれた部分からのみ抽出されている場合，k_{neg} の値とし，そうでない場合 1 とする．検索要求の中には下記のように「〜は除く」という表現には NEG のタグがふられている．

<TOPIC>
<TOPIC-ID>1003</TOPIC-ID>
<DESCRIPTION> 国連軍の派遣 </DESCRIPTION>
<NARRATIVE> 平和維持活動など国連の活動における国連軍の派遣について述べられている記事。派遣の目的または対象地域が記事から明示的に分る事。 <NEG> 日本の自衛隊を国連に派遣するかどうかという問題のみに関する記事は除く。 </NEG></NARRATIVE>
</TOPIC>

K_{neg} は，NEG のタグで囲われている部分のみから抽出されたキーワードは逆に不要な記事を取ってくる可能性が高いと考えて減点する．本論文での実験では k_{neg} の値として 0 を用いている．これは，NEG のタグで囲われた部分のキーワードの得点は加算も減点もしないという，その部分の情報は全く利用しないという状態を意味する[3]．

- $K_{\text{不要語}}$
  該当するキーワードが検索要求において「事, 認定, 記事, 言及, 対象, 場合, 具体的内容」の場合を $k_{\text{不要語},1}$ とし，そうでなく「分野, 目的, 具体的, 具体, 的, 内容, いずれ, 結果, 問題, 場合, 影響, 可能性, 可能, 性, 指摘, 対策」の場合を $k_{\text{不要語},2}$ とし，それらでない場合 1 とする．「事, 認定, 記事, 言及」といった今回の検索要求固有の表現で検索内容に関わらない表現は不要なキーワードとし，減点するためのものである．この不要語のリストは IREX の予備試験のときに作成した．

ここで，$k_{\text{タイトル}}$ などの各定数は，実験で定めるものとする．

2.4 キーワードの抽出方法

本節では検索要求文からのキーワードの抽出方法について述べる．

キーワードの抽出方法については，いくつかの異なる方法を考えている．これらについて以下で説明する．

(1) 最も短いキーワードのみを利用する方法

これは最も単純な方法である．検索要求文を単語列に分割し，その単語列のそれぞれの単語をキーワードとする方法である．われわれのシステムでは JUMAN(黒橋 長尾 1998) でまず形態素列に分割し，さらに得られた形態素を辞書を用いて細分割するということを行なっている．これは，JUMAN では形態素解析の精度向上のために，複合語のような長い形態素が登録されておりそれが単一の形態素として出力されるためである．例えば，「国連軍」という語を JUMAN に入力しても「国連軍 (名詞)」と出力されるだけで細分割を行なわない．これでは「国連」や「軍」がキーワードとならず情報検索では検索洩れの大きな原因となる．そこでわれわれのシステムでは JUMAN の結果をさらに辞書を参照して細分割するようにしている．いまのところ，簡単のため，二分割を繰り返すアルゴリズムを利用しており，「国連軍」だと「国連」「軍」が辞書にあれば分割するということを行なっている．辞書としては EDR の単語辞書 (日本電子化辞書研究所 1993) を利用している．

例として以下の検索要求からキーワードを取り出すこととする．

3 $k_{\mathrm{neg}} = -1$ つまり，NEG のタグで囲われた部分のキーワードがあると Score をその分下げるという条件や，$k_{\mathrm{neg}} = 0.5$ つまり，NEG のタグで囲われた部分のキーワードは重みを半分にするというものでも，予備試験，本試験のデータで実験を行なったが，R-Precision の精度は微妙に下がった (0.01 未満).

<TOPIC>
<TOPIC-ID>1001</TOPIC-ID>
<DESCRIPTION>企業合併 </DESCRIPTION>
<NARRATIVE>記事には企業合併成立の発表が述べられており、その合併に参加する企業の名前が認定できる事。また、合併企業の分野、目的など具体的内容のいずれかが認定できる事。企業合併は企業併合、企業統合、企業買収も含む。
</NARRATIVE>
</TOPIC>

まず，DESCRIPTION から「企業」「合併」というキーワードが得られる．また，NARRATIVE からは「記事,企業,合併,成立,発表,合併,参加,企業,名前,認定,事,合併,企業,分野,目的,具体的,内容,いずれ,認定,事,企業,合併,企業,併合,企業,統合,企業,買収」というキーワードが得られる．このとき，助詞の「に」「は」など名詞・未定義語以外の単語はキーワードとしては不適切としてすべて省き，接尾辞の形態素は直前の形態素に接合させるなどの処理をしている[4]．「最も短いキーワードのみを利用する方法」とは以上の単語をキーワードとして利用するものである．

(2) あらゆるパターンのキーワードを利用する方法

「最も短いキーワードのみを利用する方法」では，細分割されすぎていて例えば，「企業合併」というキーワードは用いず，「企業」「合併」と分離したキーワードしか用いないようになっている．それよりは，「企業合併」というものもキーワードとして用いた方がよいと考え，短いものも長いものもすべてキーワードとすることを考える．これを「あらゆるパターンのキーワードを利用する方法」と呼ぶことにする．例えば，「企業合併成立」が入力されると，「企業」「合併」「成立」という短いものから「企業合併」「合併成立」という中くらいのものと「企業合併成立」という一番長いものまですべてキーワードとして扱う方法である．この方法ならば，あらゆる長さのキーワードを利用しておりよいのではないかと考えた．しかし，「企業の合併が成立」からは「企業」「合併」「成立」の三つしか得られないのに対し，「企業合併成立」という表現からは六つのキーワードが得られ，若干不公平ではないかと考えた．これを正規化するために種々の方法を考えたが，予備試験でのデータでの実験では $\sqrt{\frac{n(n+1)}{2}}$ で各キーワードの重みを割ると精度が良かったのでそのように正規化することにした．ただし，n は連続している単語の数を示す．例えば「企業合併成立」の例だと n=3 となる．

(3) ラティスを利用する方法

「あらゆるパターンのキーワードを利用する方法」は，あらゆるパターンのキーワード

4 この他に「文献,研究,論文,提案,処理,こと,とき,もの,の」などの不要語を省いている．これは NTCIR のコンテスト (NACSIS 1999) の予備試験のデータにおいてシステムを構築する際に不要と思われた単語である．$K_{\text{不要語}}$ と混同しないようにしてほしい．$K_{\text{不要語}}$ とはまったく別処理である．

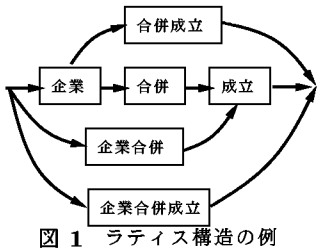

図 1　ラティス構造の例

を利用するという発想はよいが理由をつけづらい $\sqrt{\frac{n(n+1)}{2}}$ というアドホックな式で正規化する必要がある．そこで，なるべくあらゆるパターンのキーワードを利用しつつなおかつ整合性が保てるように，キーワードへの分割の曖昧性をラティス構造に保存し，先に示した検索で用いる式 (2) の Score の値が最も大きくなるようなパスでキーワードを分割してキーワードを抽出する方法を考えた．(この方法は，小澤らの論文 (小澤, 山本, 山本, 梅村 1999) の 3.1 節の「類似度を最大とする単語分割」とほぼ同じ考え方である．違いは検索方法の基本式が異なることや，分割に形態素解析システムを用いていないことで，評価関数の値が最も大きくなるように分割するという意味では全く同じである．)

例えば，「企業合併成立」の場合だと図 1のラティス構造を得る．図のように 4 種類の経路があるので分割の曖昧性として 4 種類ある．この 4 種類の分割を行ないそれぞれ式 (2) の Score を求め，最も値の大きい経路に分割する．ある記事では「企業」「合併」「成立」と分割されたものしかなく，分割して計算した方が Score が大きかったり，またある記事では「企業合併」というものが出現しており TF 項 × IDF 項 の値が「企業合併」の方が極端に大きく「企業合併」「成立」と分割した方が Score が大きくなりそのように分割するといったことになる．また，「企業合併」というものが出現していても「企業合併」の TF 項 × IDF 項 の値がそれほど大きくなく「企業」と「合併」の TF 項 × IDF 項 の和の方が大きい場合は，「企業」と「合併」と分割したものをキーワードとするといったことにもなり，そのときそのときの状況に応じたキーワード分割が可能となる．この方法ならば，「あらゆるパターンのキーワードを利用する方法」のようなアドホックな正規化を行なう必要がない．

(4)　down-weighting(Fujita 1999) を利用する方法

この方法は IREX のコンテストで他のチームが提案していた方法で，コンテストの終了後，利用を検討したものである．この方法は，「あらゆるパターンのキーワードを利用する方法」において，最も短いキーワードはそのままの重みで用い，それよりも長いキーワードは重みが小さくなるように重みづけして用いる方法である．本研究では，あるキーワードが短いキーワード x 個から構成されるとき，そのキーワードに $k_{down}{}^{x-1}$

の重みをかけることにした．ただし，k_{down} は実験で定める定数である．この方法は基本的には最も短いキーワードを用いるが，それよりも長いキーワードについての影響も少々考慮するといったものとなっている．

キーワードの抽出方法として以上の四つを示したが，さらにこれらを式 (2) で使う際にいくつかの選択肢が残されている．

例えば，先ほどの「企業合併」の例の NARRATIVE からは「記事, 企業, 合併, 成立, 発表, 合併, 参加, 企業, 名前, 認定, 事, 合併, 企業, 分野, 目的, 具体的, 内容, いずれ, 認定, 事, 企業, 合併, 企業, 併合, 企業, 統合, 企業, 買収」というキーワードが得られるが，ここでは「企業」という表現が数多く出現する．式 (1) の説明で TF の影響が強い場合の弊害を説明したが，この場合検索要求文側での TF の影響が強いという弊害が出る恐れがある．このために，このキーワード列をそれぞれ同じ種類のものはまとめ，それぞれの種類ごとに一回ずつしか出現していなかったというようにすることも可能である．Robertson らはこの場合のことも考え，以下のような評価式も利用している．(ここでは検索要求文 q における Score(d) を Score(d,q) と表記している．)

$$Score(d,q) = \sum_{\substack{\text{キーワード}\\ \text{で和をとる}\ t}} \left(\frac{TF(d,t)}{\frac{length(d)}{\Delta} + TF(d,t)} \times log\frac{N}{DF(t)} \times \frac{TFq(q,t)}{TFq(q,t)+kq} \right) \tag{6}$$

ただし，TFq(q,t) は検索要求文 q でのキーワード t の出現頻度である．kq は実験で定める定数である．kq が 0 のとき，キーワードをすべてそれぞれの種類ごとに一回ずつしか出現していないかのように扱うことと等価で，kq が ∞ のとき，キーワードをそのまま出現した個数で扱うことと等価となる[5]．

さらに，検索要求文においても IDF 項を考慮することが可能で，以下のような式も考えられる．

$$Score(d,q) = \sum_{\substack{\text{キーワード}\\ \text{で和をとる}\ t}} \left(\frac{TF(d,t)}{\frac{length(d)}{\Delta} + TF(d,t)} \times log\frac{N}{DF(t)} \times \frac{TFq(q,t)}{TFq(q,t)+kq} \times log\frac{Nq}{DFq(t)} \right) \tag{10}$$

5 kq を∞にする際，kq は定数のため式 (6) に掛けても Score の順序関係は変わらないので，式 (6) に kq をかけてから kq を∞にする．そうすると，式 (6) の最終項は以下のように計算される．

$$\lim_{kq\to\infty} \frac{TFq(q,t)\times kq}{TFq(q,t)+kq} \tag{7}$$

$$= \lim_{kq\to\infty} \frac{TFq(q,t)}{TFq(q,t)/kq+1} \tag{8}$$

$$= TFq(q,t) \tag{9}$$

つまり，TFq(q,t) となり，キーワードをそのまま出現した個数で扱うことと等価となる．

ただし，Nq は検索要求の個数で DFq(t) はキーワード t がいくつの検索要求に出現しているかの個数を意味する．多くの検索要求に出現するキーワードほど，「記事」「認識」などの不要語である可能性があり，この項を利用することによりこれらの不要語を減点する効果がある．

3　IREX コンテストの本試験に提出した二つのシステムの説明とその実験結果

われわれは IREX のコンテストとして，二つのシステムを提出する[6]際にキーワードの抽出方法として「あらゆるパターンのキーワードを利用する方法」と「ラティスを利用する方法」の二つの方法を提出することにして[7]，それぞれの方法において予備試験のデータでの精度が最も高くなるように種々の補強項を設定した．その結果，結局以下の二つのシステムを提出することにした．

(1)　システム A

キーワードの抽出方法としては「ラティスを利用する方法」を採用．位置情報は，$k_{位置,1} = 1.35$，$k_{位置,2} = 0.125$ として利用．分野情報は予備試験データでは大きな精度向上につながらなかったため利用せず．詳細情報は，$k_{タイトル} = 1.5$，$k_{固有} = 2$，$k_{など} = 1$，$k_{数字} = 0.5$，$k_{ひらがな} = 0.5$，$k_{neq} = 0$，$k_{不要語,1} = 0$，$k_{不要語,2} = 0.5$ として用いた．また，検索要求側の TF 項，つまり，式 (6) の TFq 項としては，DESCRIPTION と NARRATIVE でのキーワードを全く異なるキーワードとして扱って $k_q = 0.1$ として利用した．IDFq 項は利用していない．

(2)　システム B

キーワードの抽出方法としては「あらゆるパターンのキーワードを利用する方法」を採用．位置情報は，$k_{位置,1} = 1.3$，$k_{位置,2} = 0.15$ として利用．分野情報は，$k_{分野} = 0.1$ として利用．詳細情報は，$k_{タイトル} = 1.75$，$k_{固有} = 2$，$k_{など} = 1.7$，$k_{数字} = 0.5$，$k_{ひらがな} = 0.5$，$k_{neq} = 0$，$k_{不要語,1} = 0$，$k_{不要語,2} = 0.5$ として用いた．また，検索要求側の TF 項，つまり，式 (6) の TFq 項としては，DESCRIPTION と NARRATIVE でのキーワードを全く異なるキーワードとして扱って $k_q = 0$ として利用した．IDFq 項は利用していない．

われわれはこの条件のものを提出した．

コンテストでは 15 団体から 22 システムの結果が提出された．そのすべてのシステムの精度 (R-Precision) を表 1にあげる．表の一番左の列は各システムの名前で，この表においてわれわれのシステム A およびシステム B は，1135a と 1135b に相当する．また，表の A 判定，B 判定とは，IREX 実行委員会が定めた判定基準で，A 判定とは「記事の主題が検索課題に関連してい

6 IREX では二つまでのシステムを提出してよいことになっていた．

7 「あらゆるパターンのキーワードを利用する方法」と「ラティスを利用する方法」の二つの方法を利用して，「最も短いキーワードのみを利用する方法」を用いなかったのは，「最も短いキーワードのみを利用する方法」は単純な方法であまりよくないだろうと考えていたためである．また，「down-weighting を利用する方法」は IREX のコンテストで他のチームが提案していた方法で，コンテストの終了後に利用を検討したもので，このときは利用できなかった．

表 1　各システムの精度 (R-Precisions of all the systems)

システム名	A 判定	B 判定
1103a	0.4505	0.4888
1103b	0.4657	0.5201
1106	0.2360	0.2120
1110	0.3329	0.4276
1112	0.2790	0.3343
1120	0.2713	0.3339
1122a	0.3808	0.4689
1122b	0.4034	0.4747
1126	0.0966	0.0891
1128a	0.3384	0.3897
1128b	0.3924	0.4175
1132	0.0602	0.0791
1133a	0.2383	0.2277
1133b	0.2457	0.2248
1135a	0.4926	0.5119
1135b	0.4827	0.4878
1142	0.4455	0.4929
1144a	0.4658	0.5510
1144b	0.4592	0.5442
1145a	0.3352	0.3424
1145b	0.2553	0.2935
1146	0.2220	0.2742

る」記事のみを正解とするものであり，B 判定とは「主題ではないが記事の一部が関連する，または，なんらかの関連がある」記事をも正解とするものである．われわれのシステムは B 判定では他のシステムに及ばないところがあったが，A 判定ではシステム A,B ともに他のものよりもよい精度であった．この結果により，われわれの手法は相対評価としてそれなりによい方法なのではないかと思われる．

4　各種の情報・手法の評価実験

本節ではわれわれのシステムで用いていた様々な手法の有効性を調べるために行なった，いくつかの実験について記述する[8]．本節の実験結果では R-Precision の他に，trec_eval ツールの Average Precision (正解記事を上位から取ったたびに求めた適合率の平均) も示す．また，本比較実験の実験結果では t 検定[9]を行なっている．実験結果の各表 (表 2〜表 4) の "#" の記号のついている手法は比較の基準となる手法で，"*" のついている手法は基準の手法に対して t 検定による片側検

8 本論文の主張と直接関係のない，ここにあげなかったいくつかの実験 (細分割の有効性の確認など) が存在する．これらの実験については，文献 (村田他 1999) において詳細に記述している．

9 比較する二つの手法の各課題での精度差の分布が t 分布に従うことを仮定して，その分布で 0 よりも小さい部分が何パーセントであるかを調べることにより検定を行なった．このような検定の場合，精度差の平均が大きくても (二つの手法に大きい精度差がある場合でも) 精度差の分散が大きければ検定結果として有意差が出ない場合がある．

表 2　キーワード抽出方法の比較 (Comparison of how to extract keywords)

(a) 補強項をすべて用いた場合

キーワード抽出方法	本試験のデータでの精度				予備試験のデータでの精度			
	R-Precision		Average precision		R-Precision		Average precision	
	A 判定	B 判定	A 判定	B 判定	A 判定	B 判定	A 判定	B 判定
最も短いものだけ	0.5012	0.5205**	0.4935**	0.4764*	0.4412	0.5442	0.4546	0.5151
あらゆるパターン #	0.4827	0.4878	0.4553	0.4453	0.4373	0.5573	0.4576	0.5317
ラティスの利用	0.4926	0.5119	0.4808	0.4698	0.4599	0.5499	0.4638	0.5170
down-weight ($k_{down} = 0.01$)	0.5006	0.5217	0.4935	0.4778	0.4412	0.5445	0.4546	0.5157
down-weight ($k_{down} = 0.1$)	0.4997	0.5233	0.4939	0.4809	0.4478	0.5504	0.4563	0.5185

(b) 補強項をすべて削除した場合

キーワード抽出方法	本試験のデータでの精度				予備試験のデータでの精度			
	R-Precision		Average precision		R-Precision		Average precision	
	A 判定	B 判定	A 判定	B 判定	A 判定	B 判定	A 判定	B 判定
最も短いものだけ	0.4744	0.4897	0.4488*	0.4487*	0.3900	0.5082	0.3850	0.4468
あらゆるパターン #	0.4445	0.4665	0.4172	0.4180	0.3965	0.4981	0.3960	0.4444
ラティスの利用	0.4711	0.4884	0.4436	0.4448	0.4009	0.5069	0.3884	0.4469
down-weight ($k_{down} = 0.01$)	0.4760	0.4896	0.4492	0.4494	0.3940	0.5082	0.3850	0.4470
down-weight ($k_{down} = 0.1$)	0.4816	0.4986	0.4545	0.4568	0.4003	0.5076	0.3860	0.4498

のついている手法を基準として，“*” は t 検定の片側検定で有意水準 5% で有意に優れていることを意味し，“**” は有意水準 1% で有意に優れていることを意味する．

定で有意水準 5% で有意に優れていることを意味し，“**” のついている手法は有意水準 1% で有意に優れていることを意味する．また，この t 検定は標本数の少ない予備試験のデータでは行なっていない．(課題数は予備試験で 6 題，本試験で 30 題)

4.1　キーワード抽出方法の比較

キーワード抽出方法として，2.4節において以下の四つを示した．

(1)　最も短いキーワードのみを利用する方法

(2)　あらゆるパターンのキーワードを利用する方法

(3)　ラティスを利用する方法

(4)　down-weighting を利用する方法

このうち本試験に提出したものは，「あらゆるパターンのキーワードを利用する方法」と「ラティスを利用する方法」であった．本試験での精度では，若干ではあるが「ラティスを利用する方法」の方が「あらゆるパターンのキーワードを利用する方法」よりも良かった．

次に「最も短いキーワードのみを利用する方法」がどのくらいの精度となるかを調べるために，この方法でも本試験のデータで試してみた．そのときの各補強項の設定は「あらゆるパターンのキーワードを利用する方法」と全く同じものを用いた．これで本試験のデータで実験してみるとA 判定の R-Precision は 0.5012 で他の方法に比べてよい値であった．「最も短いキーワードのみを利用する方法」はあまりよくないだろうと考えあまり試していなかったが，この方法でも高い精

度が出せることがわかる．

また，本研究では「down-weighting を用いる方法」でも実験を試みた．ここでは k_{down} としては 0.1, 0.01 の二つのもので実験してみた．各補強項の設定は「あらゆるパターンのキーワードを利用する方法」と全く同じものを用いた．これで本試験のデータで実験してみると A 判定の R-Precision は 0.5006, 0.4997 でなかなかよい値であった．

上記五つの場合 (四手法で，「down-weighting を用いる方法」だけ $k_{down} = 0.1$ と $k_{down} = 0.01$ の二つの場合) の精度を表にまとめておくと表 2(a) のようになる．さらに，同様の実験を補強項をすべて削除した設定でも行なった．これを表 2(b) に示す．

「あらゆるパターンのキーワードを利用する方法」はアドホックな式による正規化を行なう必要性があるうえに本試験での結果では他手法に劣っているため，この方法は他の方法よりもよくない方法だと考えられる．なお，「あらゆるパターンのキーワードを利用する方法」は「最も短いキーワードのみを利用する方法」と比べ，t 検定でもいくつか有意に劣っていることが示されている．他の手法間では t 検定で有意な差は見受けられなかった．

「down-weighting を用いる方法」は補強項をすべて削除した場合，他の手法よりも高い精度を得るが，補強項を用いる場合はそれほどの効果はない．t 検定でも他の手法と有意差が見られなかったので，確実に精度向上に寄与する情報ということではない．しかし，補強項を用いない場合のように解析に用いる情報が少ない場合は精度向上が大きい．

「最も短いキーワードのみを利用する方法」だけが「あらゆるパターンのキーワードを利用する方法」と有意差が見られ，他の手法では有意差が見られなかったので，「最も短いキーワードのみを利用する方法」は安定して良好な結果を与える堅実な方法であると思われる．「ラティスを用いる方法」と「down-weighting を用いる方法」は，検定で「あらゆるパターンのキーワードを利用する方法」と有意差が出なかったためになんらかの欠点を持っていると思われる．「ラティスを用いる方法」では用いられるキーワードが状況に応じて容易に変わってしまう問題，また，「down-weighting を用いる方法」は重みを下げているとはいえ余分にキーワードを用いる問題がある．とはいえ最も短いキーワードのみを利用するよりは，もう少し長いキーワードも利用した方が望ましいのは間違いないとも思われ，このあたりはさらに研究を進める必要がある[10]．

4.2　補強項の有効性

本研究で用いた補強項は主に以下の三つに分類できる．

(1)　$K_{位置}$ (位置情報の利用)

10 本研究では，「最も短いキーワードのみを利用する方法」だけでなく，「最も長いキーワードのみを利用する方法」でも実験を行なった (つまり，連続するキーワードをすべてつなげたもののみをキーワードとする方法)．この方法では，キーワードが長くなることによりキーワードのヒット率が下がり，再現率の大幅な低下により元より精度がかなり悪くなると予想される方法である．全補強項を用いた状況で実験を行ない，本試験データで A 判定の Recall-Precision は 0.4128 の精度を得た．「最も短いキーワードのみを利用する方法」の場合が 0.5012 であったことから，大きい精度低下があることがわかる．

表 3　補強項の比較 (Comparison of extended numerical terms)

(a)「ラティスを利用する方法」での精度比較

補強項の有無			本試験のデータでの精度				予備試験のデータでの精度			
			R-Precision		Average precision		R-Precision		Average precision	
$K_{位置}$	$K_{分野}$	$K_{詳細}$	A 判定	B 判定	A 判定	B 判定	A 判定	B 判定	A 判定	B 判定
有	有	有	0.5031	0.5161	0.4888*	0.4745	0.4495	0.5471	0.4625	0.5202
有	有	無	0.4764	0.4935	0.4619	0.4375	0.4092	0.5086	0.4207	0.4624
有	無	有	0.4926	0.5119	0.4808*	0.4698	0.4599	0.5499	0.4638	0.5170
無	有	有	0.4998*	0.5301**	0.4731*	0.4856**	0.4421	0.5618	0.4383	0.5171
有	無	無	0.4932	0.4984	0.4735*	0.4519	0.4208	0.5083	0.4326	0.4638
無	有	無	0.4931	0.5084*	0.4654*	0.4634*	0.4085	0.5134	0.3945	0.4554
無	無	有	0.4979*	0.5277**	0.4673*	0.4829**	0.4407	0.5603	0.4391	0.5127
無	無	無 #	0.4711	0.4884	0.4436	0.4448	0.4009	0.5069	0.3884	0.4469

(b)「最も短いキーワードのみを利用する方法」での精度比較

補強項の有無			本試験のデータでの精度				予備試験のデータでの精度			
			R-Precision		Average precision		R-Precision		Average precision	
$K_{位置}$	$K_{分野}$	$K_{詳細}$	A 判定	B 判定	A 判定	B 判定	A 判定	B 判定	A 判定	B 判定
有	有	有	0.5012	0.5205*	0.4935**	0.4764	0.4412	0.5442	0.4546	0.5151
有	有	無	0.4867	0.4976	0.4704*	0.4464	0.4126	0.5136	0.4220	0.4649
有	無	有	0.5017	0.5094	0.4850*	0.4740	0.4410	0.5517	0.4556	0.5094
無	有	有	0.4991	0.5264**	0.4759*	0.4841**	0.4213	0.5616	0.4340	0.5095
有	無	無	0.4883	0.4952	0.4647*	0.4444	0.4247	0.5076	0.4200	0.4614
無	有	無	0.4824*	0.4990*	0.4537	0.4509	0.3927	0.5119	0.3901	0.4517
無	無	有	0.4970	0.5242**	0.4693*	0.4804*	0.4198	0.5595	0.4332	0.5070
無	無	無 #	0.4744	0.4897	0.4488	0.4487	0.3900	0.5082	0.3850	0.4468

(2)　$K_{分野}$ (分野情報の利用)

(3)　$K_{詳細}$ (種々の詳細な情報の利用)

($K_{詳細}$ には，式 (2) の記事長の項 $K_{記事長} = \frac{length}{length+\Delta}$ を含めて扱っている．)

上記の三つ補強項の有効性を確かめるために，これら三つをそれぞれ用いる場合用いない場合の合計 8 種類の実験を行なった．この実験は，「ラティスを利用する方法」と「最も短いキーワードのみを利用する方法」の二つの方法で行なった．その結果を表 3に示す．

表の一番下の行が補強項を全く用いないもので，表の一番上の行が補強項を全て用いるものだが，それらを見比べると $0.027 \sim 0.045$ という精度向上が実現できていることがわかる．(例えば，「最も短いキーワードのみを利用する方法」での A 判定の Average precision は，0.4488 から 0.4935 に 0.0447 の精度向上がある．) このことから，本研究で用いた補強項の情報が総合的に有効であり，確率型情報検索手法に位置情報や分野情報などを追加することで精度向上を実現できている．

また，補強項単独でも，補強項それぞれを一つ用いたものは補強項を一つも用いないものよりも精度がよく，それぞれの補強項が有効なことがわかる．検定結果も各補強項ともそれぞれいずれかの評価値では有意差が見受けられる．これらのことにより，位置情報や分野情報などが単独でも

表 4　詳細項の比較 (Comparison of detailed numerical terms)

	本試験のデータでの精度				予備試験のデータでの精度			
	R-Precision		Average precision		R-Precision		Average precision	
詳細項の有無	A 判定	B 判定	A 判定	B 判定	A 判定	B 判定	A 判定	B 判定
すべて無 #	0.4744	0.4897	0.4488	0.4487	0.3900	0.5082	0.3850	0.4468
$K_{\text{タイトル}}$ 有	0.4878	0.5125**	0.4614*	0.4674**	0.4136	0.5336	0.3930	0.4635
$K_{\text{固有}}$ 有	0.4746	0.4940	0.4481	0.4523	0.4031	0.5330	0.4172	0.4765
$K_{\text{など}}$ 有	0.4630	0.4765	0.4384	0.4303	0.3973	0.5097	0.3859	0.4487
$K_{\text{数字}}$ 有	0.4744	0.4897	0.4488	0.4487	0.3900	0.5082	0.3847	0.4465
$K_{\text{ひらがな}}$ 有	0.4744	0.4897	0.4488	0.4487	0.3942	0.5074	0.3854	0.4470
K_{neg} 有	0.4874	0.5037*	0.4603	0.4628*	0.4019	0.5134	0.3967	0.4554
$K_{\text{不要語}}$ 有	0.4713	0.4941	0.4507	0.4548**	0.3968	0.5295	0.3985	0.4629
$K_{\text{記事長}}$ 有	0.4775	0.4880	0.4472	0.4492	0.3945	0.5038	0.3809	0.4448

有効な情報であることがわかる．

本研究の最も大きい主張点は，確率型情報検索手法に，情報検索において当然用いるべき位置情報や分野情報などを追加して用いることで精度向上を実現することであったが，以上の結果によりこのことが実現できることが確かめられた．

本試験に提出したときの「ラティスを利用する方法」のシステムでは，分野情報を利用した場合の予備試験での精度向上がそれほどでもなかったため分野情報を用いていなかったが，本試験のデータでは分野情報を用いて 0.01 の精度向上がある．予備的な試験ではそれほど有効そうでない情報であっても，少しでも精度向上が期待できそうな情報ならば結果的には用いた方がよいと考えられる情報もある．

また，位置情報を用いるとB判定の精度が下がる傾向がある．これはB判定が「主題ではないが記事の一部が関連する，または，なんらかの関連がある」記事をも正解とするものであるためで，位置情報を用いると記事のタイトルや前の方の位置にあるキーワードの重みを大きくするために，記事の主題部分以外に検索要求を満足することが書かれている記事を拾いにくくなっているためと思われる．検定結果でも位置情報を用いる方法ではB判定で有意差がでなくなっている．

4.3　詳細項の有効性

ここでは詳細項 $K_{\text{詳細}}$ の各項と記事長の項 $K_{\text{記事長}}$ の有効性を調べる．この実験では簡単のため「最も短いキーワードのみを利用する方法」のみで行なった．各詳細項のパラメータ設定は先のとおりである．実験としては，三つの補強項 ($K_{\text{位置}}$,$K_{\text{分野}}$,$K_{\text{詳細}}$) すべてを削除したものと，その状況で八つの各詳細項 ($K_{\text{タイトル}}$,$K_{\text{固有}}$, $K_{\text{など}}$,$K_{\text{数字}}$, $K_{\text{ひらがな}}$,K_{neg},$K_{\text{不要語}}$,$K_{\text{記事長}}$) をそれぞれ一つだけ追加する実験の合計 9 個の実験を行なった．これを表 4に示す． $K_{\text{など}}$,$K_{\text{数字}}$,$K_{\text{ひらがな}}$,$K_{\text{不要語}}$ については精度が低下か変わらないかであまりよい効果がなかった．効果の大きい項としては，$K_{\text{タイトル}}$ と K_{neg} があげられる．このことは元より予想されることだが，検索要求のタイトル部

分 (DESCRIPTION) が重要ということと NEG のタグで囲まれた部分を削除するべきということは，実験においても確認されたことになる．

4.4 まとめ

以上の実験をまとめると以下のようになる．

- 四つのキーワード抽出法の比較
  キーワード抽出法として四手法を実験で試した．現状では，「位置情報・分野情報」を併用する場合「最も短いキーワードを利用する方法」が最も精度が良かった．しかし，はっきりした精度差ではなく，各手法とも調査を続ける必要がある．
- 位置情報・分野情報の利用の有効性の確認
  キーワードの位置情報と，記事の分野情報を利用して精度向上を実現できることがわかった．
- 詳細情報の利用
  実験で特に効果のあった詳細情報は，DESCRIPTION (情報要求を端的に示すフレーズ) 中のキーワードの重みを他のキーワードよりも大きくすることと，NEG のタグ (検索要求の中の「〜は除く」という表現を囲んだもの) 中のキーワードを利用しないことにすることの二つであった．

5 おわりに

われわれの情報検索の方法では基本的に，確率型手法の一つの Robertson の 2-ポアソンモデルを用いている．しかし，この Robertson の方法では検索のための手がかりとして当然用いるべき位置情報や分野情報などを用いていない．それに対しわれわれは位置情報や分野情報，さらに種々の詳細な情報をも統一的に用いることができる枠組を考案した．IREX のコンテストでは，この枠組に基づくシステムを二つ提出していたが，A 判定の精度は 0.4926 と 0.4827 で，参加した 15 団体，22 システムの中では最もよい精度であった[11]．この結果により，われわれの手法は相対評価としてそれなりによい方法と思われる．

また，本システムで用いた各手法の有効性を確かめる比較実験を行ない，種々の手法の有効性や傾向を調べた．その結果，Robertson の方法では用いられていなかった位置情報や分野情報などを用いることで精度向上が実現できること，「最も短いキーワードのみを利用する方法」でもよい精度が得られること，NEG のタグで囲まれた部分に存在するキーワードは利用しないことがよいことなどがわかった．

本研究でも様々な情報を用いたが，情報検索の手法として，関連性フィードバック (酒井,

11 本論文では記述しなかったが，IREX のコンテストで用いられたデータで，コンテスト後各種パラメータを調節することでさらに精度向上を実現している．詳細は文献 (村田他 1999) を参照のこと．

Gareth J.F. Jones, 梶浦, 住田 1999) や共起情報の利用 (高木 木谷 1996) などとまだまだ有用そうな手法がいろいろと存在する．関連性フィードバック一つをとっても，パッセージレベルの情報を利用した方が精度がよい (佐藤, 伊藤, 野口 1999) ということが言われていたりして様々な要因が絡んでいる難しそうな研究のようであるが，今後はこのあたりの情報について深く研究していく予定である．

謝辞　通商産業省電子技術総合研究所の高橋直人氏には新聞の面情報の利用についてコメントをいただいた．ここに深く感謝いたします．また，本研究では IREX のデータを利用しており，検索結果の集計をされた方々を含め，IREX の運営に携わった方々に対してもここに深く感謝いたします．

参考文献

新谷研, 角田達彦, 大石巧, 長尾真 (1997). “単語の共起頻度と出現位置による新聞の関連記事の検索手法.” 情報処理学会論文誌, **38** (4), 855–862.

藤田澄男 (1999). “自然言語処理を利用した情報の検索・分類へのアプローチ.” 情報処理学会誌, **40** (4), 352–357.

Fujita, S. (1999). “Notes on Phrasal Indexing JSCB Evaluation Experiments at IREX-IR.” IREX ワークショップ予稿集, 45–51.

黒橋禎夫 長尾真 (1998). 日本語形態素解析システム JUMAN 使用説明書 version 3.6. 京都大学大学院工学研究科.

村田真樹, 内元清貴, 小作浩美, 馬青 (1999). “IREX コンテストにおける確率型手法による情報検索.” IREX ワークショップ予稿集.

NACSIS (1999). “NTCIR Workshop 1.” *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition.*

日本電子化辞書研究所 (1993). “EDR 電子化辞書仕様説明書.”.

小川泰嗣 (1996). “情報検索の最近の動向.” 「インターネットと情報検索」講習会資料, 1–16.

小澤智裕, 山本幹雄, 山本英子, 梅村恭司 (1999). “情報検索の類似尺度を用いた検索要求文の単語分割.” 言語処理学会 第 5 回年次大会, 305–308.

Robertson, S. E. and Walker, S. (1994). “Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retreival.” In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

酒井哲也, Gareth J.F. Jones, 梶浦正浩, 住田一男 (1999). “確率モデルに基づく日本語情報フィルタリングにおけるフィードバックによる検索条件展開および検索精度評価.” 情報処理学会論

文誌, **40** (5), 2429–2438.

Salton, G. and Buckley, C. (1997). "Improving Retrieval Performance by Relevance Feedback." In Jones, K. S. and Willett, P. (Eds.), *Readings in Information Retreival.* Morgan Kaufmann Publishers.

佐藤光弘, 伊藤快, 野口直彦 (1999). "松下電器産業における IR タスクへの取り組み." IREX ワークショップ予稿集, 69–74.

Sekine, S. and Isahara, H. (1999). "IREX Project Overview." *Proceedings of the IREX Workshop*, 7–12.

高木徹 木谷強 (1996). "単語出現共起関係を用いた文書重要度付与の検討." 情報処理学会 情報学基礎研究会 FI-41-8, 61–68.

徳永健伸 (1996). "情報検索と自然言語処理." 言語処理学会第 2 回年次大会チュートリアルプログラム, 60–69.

trec_eval (1992). "ftp://ftp.cs.cornell.edu/pub/smart.".

略歴

村田真樹: 1993 年京都大学工学部卒業. 1995 年同大学院修士課程修了. 1997 年同大学院博士課程修了, 博士（工学）. 同年, 京都大学にて日本学術振興会リサーチ・アソシエイト. 1998 年郵政省通信総合研究所入所. 研究官. 自然言語処理, 機械翻訳の研究に従事. 言語処理学会, 情報処理学会, ACL, 各会員.

馬 青: 1983 年北京航空航天大学自動制御学部卒業. 1987 年筑波大学大学院理工学研究科修士課程修了. 1990 年同大学院工学研究科博士課程修了. 工学博士. 1990 ～ 93 年株式会社小野測器勤務. 1993 年郵政省通信総合研究所入所, 主任研究官. 人工神経回路網モデル, 知識表現, 自然言語処理の研究に従事. 日本神経回路学会, 言語処理学会, 電子情報通信学会, 各会員.

内元清貴: 1994 年京都大学工学部卒業. 1996 年同大学院修士課程修了. 同年郵政省通信総合研究所入所, 郵政技官. 自然言語処理の研究に従事. 言語処理学会, 情報処理学会, ACL, 各会員.

小作浩美: aiu.

内山将夫: aiu.

井佐原均: 1978 年京都大学工学部電気工学第二学科卒業. 1980 年同大学院修士課程修了. 博士（工学）. 同年通商産業省電子技術総合研究所入所. 1995 年郵政省通信総合研究所関西支所知的機能研究室室長. 自然言語処理, 機械翻訳の研究に従事. 言語処理学会, 情報処理学会, 人工知能学会, 日本認知科学会, ACL, 各会員.

(1995 年 5 月 6 日 受付)
(1995 年 7 月 8 日 再受付)
(1995 年 9 月 10 日 採録)