自然言語処理 121－18
情報学基礎 47－10
（1997. 9. 12）

概念に基づく検索要求文の拡張

斉藤 公一　　森 辰則　　中川 裕志

横浜国立大学工学部

本研究では，概念に基づく検索要求文の拡張(Qiu & H.P.Frei, 1993)を日本語に対して適用した．検索要求文中の語を類似度に基づいて拡張し検索実験を行なった．語をそれが出現する文書の集合で索引づけすることを基本概念として特定分野における類似性シソーラスを構築した．それを利用して検索要求文の概念に基づいた拡張を日本語の特定分野記事のコレクションであるBMIR–J1に適用し，各種の比較実験を行なった．検索時に元の検索要求文に含まれず付け加えられた語を一語しか含まれない文書は採用しない制限を加えたところ，検索効率の改善が見られた．さらに，マニュアル文という小規模な文書においても本手法は内容検索において有効であることがわかった．

Concept Based Query Expansion

Koichi Saito , Tatsunori Mori and Hiroshi Nakagawa

Division of Electrical and Computer Engineering,

Faculty of Engineering,

Yokohama National University

{junkie@naklab,mori@forest,nakagawa@naklab}.dnj.ynu.ac.jp

In this paper, we apply the Concept Based Query Expansion (Qiu & H.P.Frei, 1993), which is based on an automatic built similarity thesaurus to two types of Japanese text, namely news paper articles, actually test collection BMIR-J1, and technical manuals. As for Japanese technical manuals, parts of the manual that are relevant to the input query are retrieved by IR methods. We modify the vector space model based retrieval method by expanding a query with additional terms. Moreover we discard answers which include not original query terms but only one addtional term. By this modification of retrieval algorithm, retrieval effectiveness of our method improves recall and precision rates in both BMIR-J1 and Japanese manuals.

1　はじめに

語のマッチングに基づく基本的な情報検索では，対応しきれない検索要求文が存在する．そのような検索要求文に対応するために，検索要求文の拡張が必要となる．

今までの研究では，もとの検索要求文中のある語に概念辞書やシソーラスにおいて類似する語を，拡張する語として選択していた(太田 奥村, 1997)(大井, 隅田, 飯田, 1997)(Harman, 1992)(Voorhees, 1994)．しかし，このような1つの語だけに注目し

た手法では，必ずしもより良い検索要求文を生成することができない．そこで，検索要求文をその文全体が表わす概念に基づいて拡張する必要が生じてくる．

本報告では，文書コレクション固有のシソーラスを一般的シソーラスによらず，各語をその語が現われる文書で特徴づける方法で構築する．つまり，2つの語が類似した文書の集合に出現するならば，それらの語は類似しているとみなして語の類似性シソーラスを構築する．そして，構築された類似性シソーラスを用いて，検索要求文の概念に基づいた拡張を日本語に適用する．

このような拡張を日本語テストコレクションであるBMIR-J1に適用する．さらに，マニュアル文という特定知識領域の小規模な文書において内容検索実験を行ない有用性を確認した．

2 類似性シソーラスの構築

本節では，文書によって特徴づけされたタームとタームの類似度に基づいて，類似性シソーラスを自動的に構築する手法(Qiu & H.P.Frei, 1993)について述べる．

まず，文書ベクトル空間(DVS)とは，各文書 $d_1, d_2, \cdots, d_n$ を次元とするn次元ベクトル空間である．DVSにおいて，ターム t_i は，n次元ベクトル

$$\vec{t_i} = (w(t_i, d_1), w(t_i, d_2), \cdots, w(t_i, d_n)) \qquad (1)$$

で表される．ただし，$w(t_i, d_k)$ は，以下の式(2)で表されるようなターム t_i の出現頻度から計算される重みである．正規化された $tf \cdot idf$ 重みを付ける方法とのアナロジーで，document frequency(df), inverse term frequency(itf) と df の最大値 (maxdf) によって，$w(t_i, d_k)$ を以下のように定義する．

$$w(t_i, d_k) = \frac{(0.5 + 0.5\frac{df(d_k, t_i)}{maxdf(t_i)}) \cdot itf(d_k)}{\sqrt{\sum_{j=1}^{n}((0.5 + 0.5\frac{df(d_j, t_i)}{maxdf(t_i)}) \cdot itf(d_j))^2}} \qquad (2)$$

この式において，

$df(d_k, t_i)$ はターム t_i が出現する文書 d_k の頻度

$itf(d_k) = \log(\frac{m}{|d_k|})$ は文書 d_k の inverse term frequency

m はコレクション中のタームの総数

$|d_k|$ は文書 d_k に現われるタームの種類

$maxdf(t_i)$ はターム t_i が出現する文書中の最大頻度

以上の定義を用いて，t_i と t_j の2つのターム間の類似度 $SIM(t_i, t_j)$ を次式で定義する．

$$SIM(t_i, t_j) = \vec{t_i} \cdot \vec{t_j} = \sum_{k=1}^{n} w(t_i, d_k) w(t_j, d_k) \qquad (3)$$

ただし，$\cdot$ は内積を表す

類似性シソーラスはすべてのタームの組 (t_i, t_j) の類似度を求めることによって構築される．直観的には，t_i と t_j が現われる文書の集合が類似していれば $SIM(t_i, t_j)$ が大きくなるということである．その結果 $SIM(t_i, t_j)$ は次の範囲に値を持つ対称行列となる．

$$0 \leq SIM(t_i, t_j) \leq 1 \qquad (4)$$

3 概念に基づく検索要求文の拡張

本節では，構築された類似性シソーラスを用いて検索要求文を拡張する手法及び，拡張された検索要求文での文書検索手法(Qiu & H.P.Frei, 1993)について述べる．

この手法の特徴は，検索要求文中のある一つの語に類似している語を拡張語とするのではなくて，検索要求文全体が表わす概念に近い語を拡張語とする．

3.1 DVSにおけるタームと検索要求文の関係

図1にDVSにおけるタームの散在する様子を示す．そこにおいて，t と t' という2つのタームを含むユーザの検索要求文は Q で示される．類似性シソーラスは，この特定コレクションDに関する

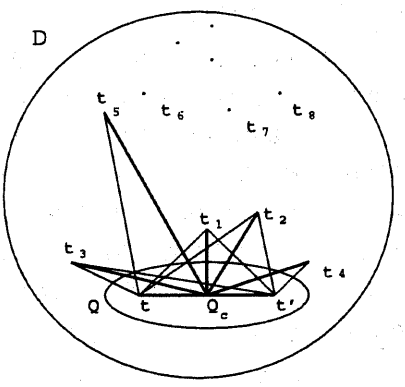

図 1: DVS におけるタームと検索要求文との関係

すべてのターム対の類似度を含んでいる．図 1で，2 つのタームがお互いに近いほど類似しているとする．t_3 は他のどれよりも t に類似しているし，t_4 は他のどれよりも t' に類似している．ここで，検索要求文 Q の表す概念を仮想ターム Q_c とする．Q_c は Q の重心である．すると，t_1 と t_2 は検索要求概念 Q_c に最も類似しているタームなので，t_3 と t_4 の代わりに付加される検索タームとして選ばれることになる．

この方法で，あるタームと検索要求文の表す概念間の類似度全体が評価できる．各々の検索要求ターム t_i は第 2節で述べたようにタームベクトル $\vec{t_i}$ によって定義される．q_i は検索要求 Q 中にターム t_i が現われれば正の値，現われなければ 0 という重み (スカラー量) である．言い替えれば，概念は検索要求 Q に対して Q に含まれるターム t_i を q_i で重みづけした総和となり，式 (5) で表される．これを検索要求概念ベクトル $\vec{Q_c}$ とする．

$$\vec{Q_c} = \sum_{i=1}^{m} q_i \vec{t_i} \tag{5}$$

通常のベクトル空間モデルでは，検索要求 Q をターム数次元のベクトルと考え，また，文書もターム数次元のベクトルとして表し，両者の内積によって Q と文書の類似度を計算する．

ところが，式 (5) に表されるように，ここでは，検索要求概念 Q_c は，$\vec{t_i}$(i=1 ～ m) の重みづけ総和である．各 $\vec{t_i}$ は，式 (1) より，文書数次元のベクトルである．よって，$\vec{Q_c}$ は文書数次元のベクトルである．この点が通常のベクトル空間モデルと根本的に異なる．

3.2 検索要求文とタームの類似度

あるタームと検索要求概念 Q_c との類似度を $Simqt(Q_c, t)$ で表すとすると式 (5) より，

$$\begin{aligned} Simqt(Q_c, t) &= \vec{Q_c} \cdot \vec{t} \\ &= (\sum_{i=1}^{m} q_i \vec{t_i}) \cdot \vec{t} \\ &= \sum_{i=1}^{m} q_i (\vec{t_i} \cdot \vec{t}) \\ &= \sum_{i=1}^{m} q_i SIM(t_i, t) \end{aligned} \tag{6}$$

検索要求文概念 Q_c に対するターム t の重み $weight_a$ を定義する．

$$weight_a(Q_c, t) = \frac{Simqt(Q_c, t)}{\sum_{i=1}^{m} q_i} \tag{7}$$

$$0 \leq weight_a(Q_c, t) \leq 1$$

3.3 検索要求文の拡張

$\vec{Q_c}$ に近いタームほど高い重みを持つベクトル $\vec{Q_e}$ を考える

$$\vec{Q_e} = (qe_1, qe_2, \ldots, qe_m) \tag{8}$$

ただし，

$$qe_j = \begin{cases} weight_a(Q, t_j) & t_j \text{ の } weight_a \text{ が} \\ & \text{上位 } r \text{ にある場合} \\ 0 & \text{それ以外} \end{cases} \tag{9}$$

また r はあらかじめ定められた加えられるタームの数である．

この $\vec{Q_e}$ によって，$\vec{Q_c}$ を拡張する．すなわち拡張された検索要求 Q_{new} は，

$$\begin{aligned} \vec{Q_{new}} &= \vec{Q_c} + \vec{Q_e} \\ &= \sum_{i=1}^{m} q_i \vec{t_i} + \sum_{i=1}^{m} qe_i \vec{t_i} \\ &= \sum_{i=1}^{m} (q_i + qe_i) \vec{t_i} \end{aligned} \tag{10}$$

ターム t_i の文書 d_k における頻度 $tf(t_j, d_k)$ を重みととする d_k の $\vec{t_j}$ の重みづけ総和によるベクトル表現 $\vec{d_k}$ が，式(11)で与えられる．このベクトルは，$\vec{Q_c}$ と同じく，文書数次元のベクトルである．

$$\vec{d_k} = \sum_{j=1}^{m} tf(t_j, d_k)\vec{t_j} \tag{11}$$

を考える．この時，$\vec{d_k}$ と $\vec{Q}_{new}$ の類似度は，式(6)により

$$\begin{aligned} \vec{Q}_{new} \cdot \vec{d_k} &= (\sum_{l=1}^{m}(q_l + qe_l)\vec{t_l})(\sum_{j=1}^{m} tf(t_j, d_k)\vec{t_j}) \\ &= \sum_{l=1}^{m}\sum_{j=1}^{m}(q_l + qe_l)tf(t_j, d_k)SIM(t_l, t_j) \end{aligned} \tag{12}$$

ただし，qe_l は，式(6)(8)により，SIM の関数である．式(12)の値によって，文書 d_k がランキングされる．よってこのランキングは，類似性シソーラスで定義される $SIM(t_i, t_j)$ を利用して計算される．

4 BMIR–J1 を用いた実験

BMIR–J1(福島, 小川, 石川, その他, 1996) に対して，この概念に基づく検索要求文拡張を適用し，検索効率から本手法の日本語文書における有用性を示す．

4.1 テストコレクション BMIR-J1

評価には，株式会社 日本経済新聞の協力によって，社団法人 情報処理学会・データベースシステム研究会・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が，1993年9月1日から12月31日の日本経済新聞記事を基に構築した情報検索評価用データベース(テスト版)を利用した．正解にはA,Bの2ランクがあるが，今回の評価では同一に扱った．なお，記事の情報は本文のみを利用し，タイトルや付与されているキーワード，記事の重要度等は使用していない．

4.2 試験タームの抽出

BMIR–J1の文書コレクション600記事(これらは名詞, 未定義語からなる10,215タームを含んでいる)から，重要と思われるタームを抽出してこれを試験用のタームのコレクションとする．形態素解析システムはJUMAN3.0(松本, 黒橋, 宇津呂, 妙木, 長尾, 1996)を用いた．重要語抽出の基準としては，TFとIDFを用いて算出された以下のような重要度を用いた．

$$\sum_{j=1}^{m} TF(t_i, d_j) \cdot IDF(t_i) \tag{13}$$

この式(13)の値が大きい順に定められた数のタームを抽出し，試験用タームコレクションとした．今回は，2000,2500語彙で試験を行なった．

4.3 概念ベースの構築

必要となるデータベースは以下の3種である．

1. タームから出現文書番号とその文書における出現回数がわかるもの
2. 文書番号からそこに出現するタームの種類がわかるもの
3. 2つのタームの組からその類似度がわかるもの

上記のうち，上の2つのデータベースは全タームに対して作成され，3つめは重要と思われるタームに対して作成した．

4.4 検索要求タームの重み

検索要求文 Q の概念を表すためには，検索要求ターム t_i にそれぞれ適切な重み q_i を設定しなければいけない．しかし，BMIR–J1のような自然言語による検索要求文のみから，ユーザの意思を的確に反映する重みを算出することは不可能である．

まず，タームの重みを(0,1)の二値にして実験を行なってみたが，良好な結果は得られなかった．そこで，我々はあるタームが検索要求文をどの程度特徴付けるのかという目盛として，そのタームの文書コレクション中におけるIDFを用いることにし

た．コレクション中で偏りをもって出現するタームは，平均的に出現するタームよりも検索要求文をより特徴づけると考えられるからである．

4.5 比較実験

今回比較に用いた方法は，検索要求文の拡張を行なっていないベースラインとして，標準的な $tf \cdot idf$ を重みとするベクトル空間法を用いた．各語彙数において，公平な比較をするために，$tf \cdot idf$ 法も式(13)で定めた数の語を用いて検索実験を行なった．

4.6 検索要求文拡張結果

検索要求文の拡張結果の例を表1に示す．なお，表の左上から順に $weight_a$ の大きいもの，つまり元の検索要求文の概念に近いタームから並べてある．元の検索要求文は「女性の雇用問題」であり，この文から抽出されるタームは「女性」「雇用」「問題」の3タームである．

表 1: 検索要求文の拡張結果の例: 女性の雇用問題

ターム	$weight_a$
女子	0.997905753479534
女性	0.997897074089773
雇用	0.997859774422078
主義	0.997652475609309
労組	0.997629969216605
業革	0.997595452690996
番頭	0.997552990141555
神	0.997539023654002
労働者	0.997511681616123
槙	0.997506910545145
KOA	0.997506834599689

4.7 検索結果

拡張された検索要求文での検索結果を図2，図3に示す．図において，“n word”とは，n語拡張した場合を示している．

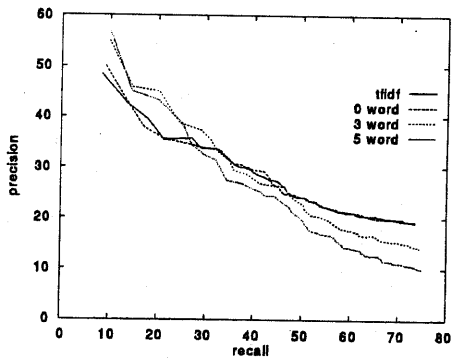

図 2: 語彙2000語での検索要求文の拡張による検索結果の相違

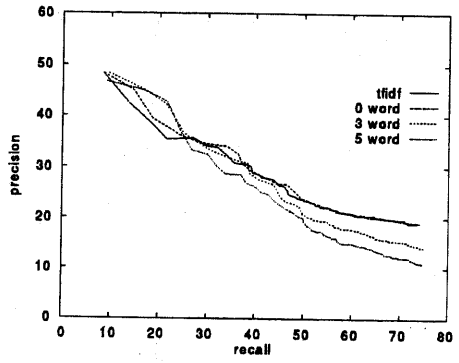

図 3: 語彙2500語での検索要求文の拡張による検索結果の相違

再現率の低いところでは，拡張された検索要求文で適合率が改善されている．しかし，再現率の高いところでは，元の検索要求文に劣っていることが確認された．この適合率の低下は元の検索要求文に追加された語が拾ってしまったノイズの文書によると考えられる．そこで，我々は(Qiu & H.P.Frei, 1993)の手法を改良し，元の検索要求文中の語が現われずに，追加された語が一語だけ現われるような文書はノイズであるとして，採用しないという制限を加えてみた．図4，図5にその結果を示す．

低再現率での適合率は向上し，高再現率でベースラインとほぼ同じ適合率を実現していることが確

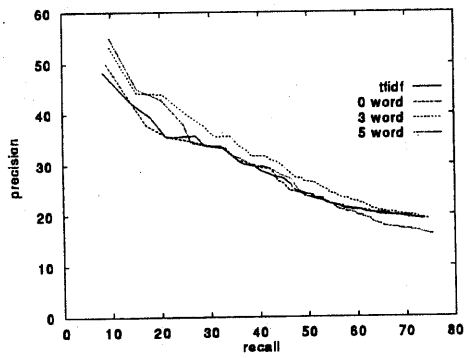

図 4: 拡張された語の一語でのマッチを制限した場合 (語彙 2000)

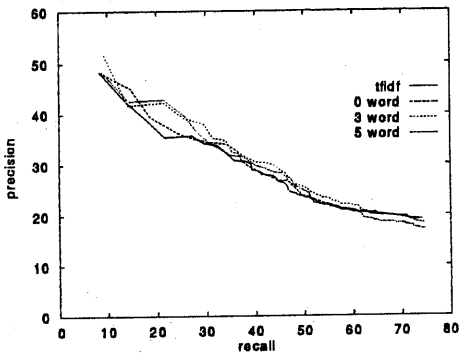

図 5: 拡張された語の一語でのマッチを制限した場合 (語彙 2500)

認できる.

5 マニュアル文への適用

より小さい，つまりターム数の少ない，文書空間において，この方法の有用性を確かめるために，日本語マニュアルの内容検索に本手法を適用した.

5.1 マニュアル文について

マニュアル文は新聞記事と違い，一つの完結した文書空間である．分野が強く限定されているので，意味の多義性などは少なく，より良いシソーラスが構築できると考えられる.

我々は一つのマニュアルについて一つのシソーラスを構築した．類似性シソーラスに登録された語彙はすべてのタームを選択した．また，検索の単位となる文書としては，マニュアルの著者が与えている構造，すなわちマニュアルの章ないし節のうちの最小の単位を用いた．つまり，質問文に対応する事柄の書かれている章ないし節が検索される．各質問に対する正解は学生３人によって人手で決定した.

実験には表 2に示す日本語マニュアル文を使用した.

表 2: 評価に用いたマニュアル

マニュアル	size (kB)	質問数 (個)	章節数 (個)	語彙数 (語)
日本語形態素解析システム JUMAN	31	20	39	596
構文解析システム SAX	29	24	33	557
家庭用ビデオデッキ	69	21	23	672
仮名漢字変換 FEP たまご	57	20	53	1244

5.2 実験結果

実験結果は以下のとおりである．４種類のマニュアルのうち３つで検索効率の改善が見られた．これは，本手法が短い文書においてもノイズを拾わずに適切なタームを拡張語として選択していると考えられる.

４本のマニュアルのうち３本で全体的なパフォーマンスの改善が見られた．さらに残りの１本についても，特定の再現率の範囲においては適合率の改善が見られた．これらは，小さな文書空間においても本手法が有効なことを示している．ただし，付加された語一語のマッチを採用しない方法では，高再現率での適合率の改善は見られなかった．これは，マニュアル文ではタームの意味が特定されているので，付加された語を原因とするノイズは少ないからと考えられる.

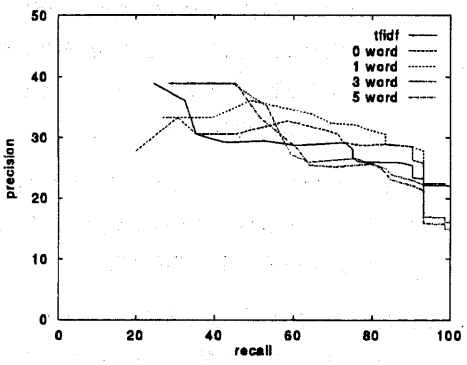

図 6: マニュアル"juman" の検索における適合率, 再現率

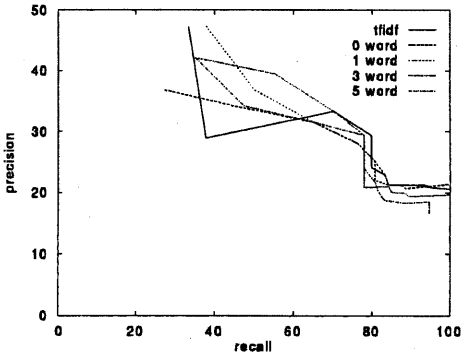

図 7: マニュアル"sax" の検索における適合率, 再現率

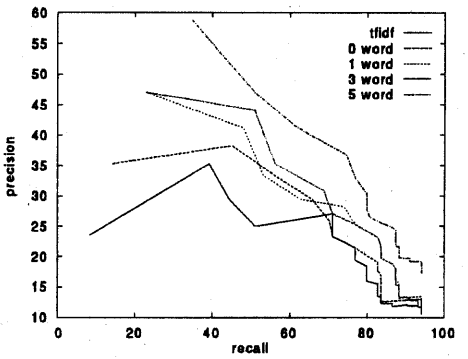

図 8: マニュアル" たまご" の検索における適合率, 再現率

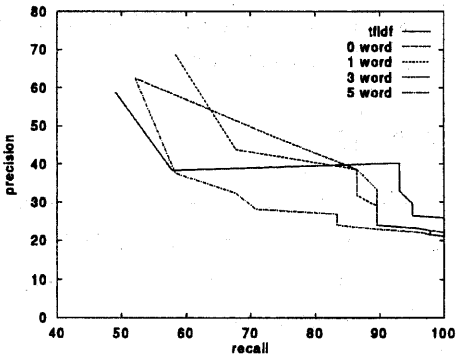

図 9: マニュアル" ビデオデッキ" の検索における適合率, 再現率

6 他研究との比較

6.1 EDR 電子化辞書を用いた方法との比較

(太田・奥村, 1997) の提案する EDR 電子化辞書を用いた方法を実装して比較してみたが, 検索効率の改善は見られていない (図 10). これは, 検索要求文の概念を考慮に入れていないからであると考えられる. さらに, 本手法は文書コレクション固有のシソーラスを構築するので, 一般的なシソーラスである EDR 電子化辞書を用いたのと違い, 文書コレクション固有の語や意味なども考慮に入れた検索がなされていると考えられる.

7 まとめ

本研究では, まず概念に基づく検索要求文の拡張について述べた. 次に, BMIR-J1 と日本語マニュアルを用いて拡張された検索要求文での, 検索実験を行なった. BMIR-J1 においては, 元の検索要求文にない語のみにマッチする検索結果をはずすことにより, 適合率, 再現率とも従来方法より改善することを示した.

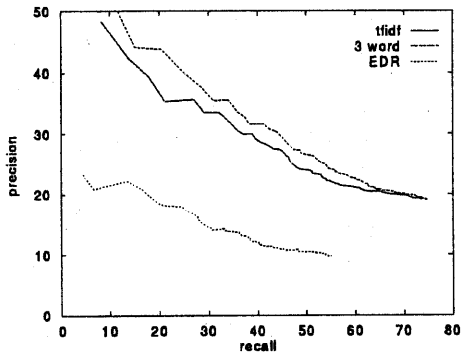

図 10: EDR 電子化辞書を用いたものとの比較

謝辞

本報告では，日本語情報検索システム評価用コレクション BMIR-J1 を用いた．

参考文献

Harman, D. (1992). Relevance Feedback and Other Query Modification Techniques. *Information Retrieval — Data Structures & Algorithms*, 241–263.

Qiu, Y. & H.P.Frei (1993). Concept Based Query Expansion. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrival*, pp. 160–169.

Voorhees, E. M. (1994). Query Expansion Lexical-Semantic Relations. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrival*, pp. 61–69.

大井耕三, 隅田英一郎, 飯田仁 (1997). “意味的類似性と多義性解消を用いた文書検索手法.” 自然言語処理 Vol.4 No.3, pp. 51–70.

福島俊一, 小川泰嗣, 石川徹也, その他 (1996). “日本語情報検索システム評価用コレクション 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. 電子情報通信学会, 言語理解とコミュニケーション研究会, 対話システム研究会.

太田千昌 奥村学 (1997). “EDR 電子化辞書を用いたクエリー拡張による検索支援.” 言語処理学会第 3 回年次大会発表論文集, pp. 373–376.

松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木裕, 長尾真 (1996). 日本語形態素解析システム JUMAN. 京都大学工学部 長尾研究室, 奈良先端科学技術大学院大学 松本研究室.