TREC-7参加報告

白井 清昭[†]，Rila Mandala[†]，徳永 健伸[†]，田中 穂積[†]，奥村 明俊[‡]，佐藤 研治[‡]

[†] 東京工業大学 大学院情報理工学研究科
〒 152–8552 東京都目黒区大岡山 2–12–1
Tel: 03–5734–2837
E-mail: {kshirai,rila,take,tanaka}@cs.titech.ac.jp

[‡] 日本電気株式会社 C & C メディア研究所
〒 216–8555 神奈川県川崎市宮前区宮崎 4–1–1
Tel: 044–856–2152
E-mail: {okumura,satoh}@hum.cl.nec.co.jp

概要:

我々は，1998 年 11 月に開催された情報検索のコンテスト The 7th Text REtrieval Conference(TREC-7) に参加した．本論文では，TREC-7 のメインタスクである ad hoc task 及び 7 つのサブタスクの概要について説明し，これらのタスクの評価結果について報告する．また，会議では次回のコンテスト TREC-8 に関する議論も行われたので，その内容についても報告する．さらに，我々が TREC-7 で用いた，既存のシソーラス (WordNet) とコーパスから自動構築されたシソーラスの両方を利用して質問拡張を行う手法について述べる．

キーワード: TREC-7，情報検索，コンテスト

Report on TREC-7

SIRAI Kiyoaki[†], Rila Mandala[†], TOKUNAGA Takenobu[†], TANAKA Hozumi[†], OKUMURA Akitoshi[‡], SATOH Kenji[‡]

[†] Graduate School of Information Science and Engineering, Tokyo Institute of Technology
2–12–1, Ookayama, Meguro-ku, Tokyo 152–8552, Japan
Tel: +81–3–5734-2837
E-mail: {kshirai,rila,take,tanaka}@cs.titech.ac.jp

[‡] NEC Corporation, C & C Media Research Laboratories
4–1–1, Miyazaki, Miyamae-ku, Kawasaki, Kanagawa 216–8555, Japan
Tel: +81–44–856–2152
E-mail: {okumura,satoh}@hum.cl.nec.co.jp

Abstract:

We participated in the 7th Text REtrieval Conference (TREC-7) held in November, 1998. In this paper, we outline the ad hoc task, which is the main task of TREC-7, and 7 sub tasks. We also describe the results of the various participants on these tasks and discuss the forthcoming TREC-8. Furthermore, we describe our method for TREC-7, using query expansion based on both an existing thesaurus (WordNet) and thesauri acquired automatically from a corpus.

Keywords: TREC-7, information retrieval, contest

1 はじめに

1998 年 11 月 9 日から 11 日にかけて，アメリカはワシントン DC 郊外にある National Institute of Standards and Technology (NIST) において，The 7th Text REtrival Conference (TREC-7) が開催された [1, 2]．TREC は，NIST と Defence Advanced Research Projects Agency (DARPA) をスポンサーとする情報検索のコンテストであり，Tipster プログラムの一環として行われている．今回は第 7 回目のコンテストにあたる．

NEC・東工大グループは，TREC のメインタスクである ad hoc task に参加した．本論文では，TREC-7 の概要について報告し，また当グループが用いたテキスト検索手法について概説する．

2 TREC-7 の概要

2.1 タスクの概要

TREC-7 では以下の 8 つのタスクが行われた．ここでは，TREC-6 からの変更点も含めて，それぞれのタスクの概要について述べる．

- ad hoc task

  TREC のメインタスクであり，取り出すべき文書の内容を記述したトピックが与えられ，文書集合の中からそのトピックの記述に該当する文書を取り出し，その精度を競うタスクである．いわゆるテキスト検索に相当する．ad hoc task には，関連文書の検索を完全に自動で行う automatic と，ユーザとのインタラクションなども利用する manual の 2 種類のタスクがあり，両者は別々に評価される．

  参加者には検索課題として 50 個のトピックが与えられる．トピックは，title, description, narrative の 3 つの部分より構成され，検索するべき文書の内容がこの順序でより詳しく記述されている．title に含まれる単語は description の中にも必ず含まれるようになったことと，検索された文書がそのトピックと関連があるか否かの判断基準のみを記述するように narrative の内容が限定されたことが TREC-6 からの変更点である．トピックの例を図 1 に，50 個のトピックの長さの最小値，最大値，平均値を表 1 に示す．一方，検索対象となる文書集合の大きさは約 2GB であり，Foreign Broadcast Information System，Federal Register(以上 2 つは米国政府による報告書), Financial Times，LA Times (以上 2 つは新聞記事) から構成されている．

  参加者は，文書集合からトピックに関連する文書をランキングをつけて取り出し，その上位 1000 個の文書番号を提出する．検索対象となる文書集合と過去の TREC における評価用データが事前に配布され，それらをもとにシステムの開発が行われる．また，各参加者は異なる手法による検索結果を 3 つまで提出することができる．1 参加者の 1 つの検索結果は 1 run と呼ばれる．また，検索結果を提出する際には，トピック中の title, description, narrative のどの部分を用いたのかを明示しなければならない．

  TREC では，それぞれのトピックについて，各 run によるランキングの上位 100 個の文書を取り出し，それらの文書がトピックと関連があるか否かを人手により判定し，各 run の再現率，適合率などを調べている．これは，いわゆるプーリングと呼ばれる手法であり，参加者のテキスト検索の手法が多様であることを前提にしている．しかしながら，全ての run から取り出された文書の 1 トピック当りの平均のべ数が 7,805 であるのに対し，平均異り数は 1,611 であった．参加者のテキスト検索の手法が多様であればあるほどプーリングされた文書の異り数とのべ数は近い値を取るはずである．しかし，TREC-7 における異り数とのべ数の比は 0.21 であり，この値は過去 7 回行われた TREC のうち最も低く，検索結果をプーリングにより評価するための前提が崩れつつあることを示唆している．

```
<top>
<num> Number: 396
<title> sick building syndrome
<desc> Description:
Identify documents that discuss sick building syndrome or building-related illness.
<narr> Narrative:
A relevant document would contain any data that refers to the sick building or building-related illness, including illnesses caused by asbestos, air conditioning, pollution controls. Work-related illnesses caused by the building, such as carpal tunnel syndrome, are not relevant.
</top>
```

図 1: トピックの例

表 1: トピックの長さ

	最小	最大	平均
title	1	3	2.5
description	5	34	14.3
narrative	14	92	40.8
all	31	114	57.6

- CLIR track

  CLIR(Cross Language Informatin Retrieval) track は，トピックの言語と検索対象となる文書集合の言語が異なる場合を想定したタスクである．TREC-7 で対象となった言語は英語，フランス語，ドイツ語，イタリア語の 4 つである．トピックは，それぞれの言語について 7 つずつ，合計 28 個与えられた．CLIR track は前回の TREC-6 から行われているが，今回はイタリア語が対象言語として新たに追加されたことと，トピックに関連のある文書を 4 つの言語の文書集合から同時に取り出すタスクが追加されたことが主な変更点である．

- filtering track

  filtering track のタスクは，routing, batch filtering, adaptive filtering の 3 種類に分けられる．routing タスクは，文書集合に含まれる全ての文書にランキングをつける．事前にトピックと訓練用の文書集合が与えられ，ランキングをつけるべき文書集合は評価時に初めて与えられるため，適合性フィードバックなどによりトピックに特化したシステムを事前に構築できる点が ad hoc task と異なる．batch filtering は，文書集合に含まれる全ての文書に対してトピックに関連があるか否かを判定し，ランキングは行わない．adaptive filteling は，batch filtering と同様に文書がトピックに関連があるか否かを判定するのだが，文書集合が時系列順に並べられ，前に判定した文書中に含まれるタームを次の文書がトピックと関連があるか否かの判定に用いるなど，文書の判定がすすむにつれて判定基準を調整するタスクである．

  filtering track では，トピックとして過去の TREC で用いられた 50 個のトピックが，訓練用及びテスト用の文書集合として AP 通信のニュース記事が用いられた．また評価基準として，従来用いられていた ASP(Average Set Precision) [1]の他に，式 (1),(2) の 2 つの評価基準が TREC-7 から採用された．

$$F1 = 3R^+ - 2N^+ \quad (1)$$

$$F3 = 4R^+ - N^+ \quad (2)$$

  式 (1),(2) において，R^+ はトピックと関連があるとして取り出された文書のうち本当にトピックと関連があった文書の数，N^+ はトピックと関連がなかった文書の数をそれぞれ表わす．これらの評価基準は，ASP に比べて，トピックと関連のない文書を関連のある文書として取り出さないこと，すなわち N^+ の数を少なくすることに重点を置いて設定された．

[1]再現率と適合率の積である．

- high precision track

  high precision track は，5 分間という限られた時間内で，トピックと関連がある文書を 15 個取り出すタスクである．トピックと検索対象となる文書集合は ad hoc task と同じものが用いられた．制限時間以内なら，検索者からの適合性フィードバックなど，いかなる手段を用いても構わない．すなわち，ユーザインタフェースなども評価の対象に含まれる．検索者は検索システムに精通していることが前提とされ，テキスト検索の精度の上限を調べることも本タスクの目的のひとつとして挙げられる．

- interactive track

  interactive track は，20 分間という限られた時間内で，トピックと関連がある文書をできるだけ多く取り出すタスクである．トピックとしては ad hoc task で用いられた 8 トピックが，検索対象となる文書集合としては Financial Times の文書が用いられた．また，トピックとしては title, description, narrative の他に instance(そのトピックと関連がある文書の例) が与えられた．検索システムのパフォーマンスが検索者によってどの程度影響されるかを調べるために，参加条件として，8 人の検索者が 2 つの検索システムを用いて 4 つのトピックに関して検索を行うことが要求される．これは，TREC-6 での議論において，最低 8 人による評価を行わなければ統計的に十分信頼できるデータが得られないという結論に達したことによる．

- query track

  query track は，トピックから検索システムの入力となる質問 (query) を生成するタスクであり，TREC-7 から新たに追加された．参加者は，50 個のトピックのそれぞれから，以下の 5 種類の質問を作成する．

  1. very short
     2,3 単語からなる質問
  2. sentence
     トピックのみから抽出した一文
  3. feedback
     トピックと関連のある文書のみから抽出した一文
  4. manual structured query
     トピックとそれに関連のある文書の両方から作成した TIPSTER DN2 フォーマットによる質問
  5. automatic structured query
     manual structured query と同様だが，人手を用いないで作成した質問

  作成した質問は，参加者自身の検索システムと他の参加者の検索システムの入力として用いられ，検索結果

の精度が評価される．文書集合としては AP 通信のニュース記事が用いられた．

- SDR track

  SDR(Spoken Document Retrieval) track は，音声データとして格納された 100 時間分の放送ニュース原稿を検索対象としたタスクである．音声認識結果としては以下の 4 種類があり，音声認識システムの精度と検索システムの検索精度の相関関係を調べることも目的のひとつとしている．

  1. reference
     人手による書き起こし．
  2. baseline1
     主催者が用意した音声認識ツールの認識結果．WER(Word Error Rate，単語当たりの誤認識率) は 35%．
  3. baseline2
     主催者が用意した音声認識ツールの認識結果．WER は 49%．
  4. recognizer
     参加者が開発した音声認識ツールの認識結果．

  23 個のトピックが用意され，ad hoc task と同じ評価基準が用いられた．また，TREC-6 からの変更点として，検索の対象となる音声データの録音時間が 2 倍 (50 時間から 100 時間) になったことと，トピックの数が 50 個から 23 個に減ったことが挙げられる．

- VLC track

  VLC track(Very Large Corpus) track は，100GB という巨大な文書集合に対してテキスト検索を行うタスクである．文書集合は html 形式で書かれた web 文書であり，英語以外の言語で書かれた文書や非 ascii 文字を含んだ文書も存在する．検索対象文書数が非常に膨大であるため，検索結果の再現率による評価は行わず，適合率のみによって評価する．また，検索に要する時間や，タームの索引付けなどの検索システム構築に要する時間も評価の対象となる．トピックは ad hoc task と同じ 50 トピックが使用された．また，検索対象文書集合の規模と検索精度の相関を調べるために，文書集合のサイズを 100GB 全て，10%，1%と変化させ，それぞれに対してテキスト検索を行った．

2.2 参加者数

TREC-6 及び TREC-7 の参加者数を表 2 に示す．メインの ad hoc task では参加者数が増えているが，ほとんどのサブタスクでは参加者数は減少した．

表 2: 参加者数

	TREC-6	TREC-7
ad hoc	31	42
CLIR	13	9
filtering	10	12
HP	5	4
interactive	9	8
query	0	2
SDR	13	10
VLC	7	6

表 3: ad hoc task(automatic) の結果

	Ave Prec	R-Prec
Okapi	0.3033	0.3390
AT & T	0.2961	0.3171
Univ. of Massachusetts	0.2815	0.3178

3 評価結果

本節では，TREC-7 における各タスクの評価結果について述べる．

- ad hoc task

  ad hoc task の automatic の成績が上位の参加者とその結果を表 3 に，manual の成績が上位の参加者とその結果を表 4 にそれぞれ示す．

  表 3, 4 において，“Ave Prec” は 11 点平均適合率，“R-Prec” はトピックと関連のある文書が全て取り出された時点での適合率を表わす．また，表中の数値は，ad hoc task の 50 個のトピックについての平均値である．成績が上位の参加者の正解率の差は 1%前後であり，飛び抜けて成績の良かった参加者はいなかった．

  automatic タスクでは，確率モデルをベースにした検索システム (BBN Technology，Twentry One など) やベクトル空間モデルをベースにした検索システム (AT & T，University of Massachusetts など) が多く，

表 4: ad hoc task(manual) の結果

	Ave Prec	R-Prec
CLARITECH Corp.	0.3702	0.4140
MI Tech Inc.†	0.3675	0.4392
Univ. of Waterloo	0.3587	0.4012

† Management Information Technology Inc.

検索精度を上げるための手法はいくつも提案されていたものの，既存のテキスト検索の手法と大きく異なるような画期的な手法の提案はなかった．一方，manual タスクにおいては，検索者による適合性フィードバックをどのように実現するかに重点が置かれた．初期の検索結果をクラスタリングして検索者に見せる手法 (CLARITECH Corp.) やブーリアンモデルによる質問を検索者に作成させる手法 (University of Waterloo) などが発表された．

- CLIR track

CLIR track の成績が上位の参加者とその結果を表 5 に示す．ad hoc task と同程度の正解率が得られていることが注目すべき点である．

表 5: CLIR track の結果

	Ave Prec	R-Prec
IBM	0.2942	0.3359
Twenty One	0.2846	0.3272
Eurospider	0.2767	0.3175

このタスクでは，英語で書かれたトピックに対し，4 つの言語の文書集合から関連のある文書を同時に取り出さなければならない．そのため，2 言語間のクロスリンガルテキスト検索を行い，その結果をマージするという方法を採用する参加者がほとんどであった．個々の言語に対するテキスト検索結果をマージする際に，質問言語と同じ言語の文書の割合を増やす手法 (University of Maryland) や検索対象となる文書集合の大きさの比に従ってマージする手法 (TextWise, Inc.) などが発表された．

- filtering track

filtering track の routing タスクの成績が上位の参加者とその結果を表 6 に示す．ad hoc task よりも高い正解率が得られているが，これは事前にトピックと訓練用文書集合が与えられ，適合性フィードバックなどを用いてトピックに特化した検索システムの調整が行われたためと考えられる．

表 6: filterling track (routing) の結果

	Ave Prec
NTT Data	0.514
AT & T	0.419
CUNY	0.356

- high precision track

high precision track の成績が上位の参加者とその結果を表 7 に示す．

表 7: high precision track の結果

	Pre(15)	R-Pre(15)
Cornell Univ.	0.5853	0.5967
Univ. of Waterloo	0.5693	0.5772
Australian National Univ.	0.5120	0.5205

表 7 において，“Pre(15)” は取り出した 15 個の文書のうちトピックと関連のある文書の割合である．また，“R-Pre(15)” は「Relative Precision at 15」と呼ばれ，式 (3) のように定義される．

$$\text{R-Pre(15)} = \frac{\text{Pre(15)}}{\text{考えられる最大の適合率}} \qquad (3)$$

例えば文書集合内にトピックと関連のある文書が 10 個しかなければ，式 (3) の分母，すなわち最大の適合率は 10/15 として与えられる．トピックと関連のある文書の数が 15 以上なら，R-Pre(15) は Pre(15) と一致する．すなわち，Pre(15) が 1 でなくても，R-Pre(15) が 1 であれば，最も理想的な検索結果が得られたということになる．

このタスクでは，5 分間の制限時間以内ならいかなる手段を使うことも許されているのだが，R-Pre(15) の値は最高でも 6 割程度である．このタスクの正解率はテキスト検索の上限の目安とも考えられるが，TREC のように検索対象となる文書集合が大きくなると，これ以上の高い精度でのテキスト検索は難しいようである．

- interactive track

interactive track に関しては，参加者の検索結果の詳細を載せたデータシートが配布されなかった．参考までに，University of North Carolina の結果を表 8 に示す．

表 8: interactive track の結果

	irisa v.s. iriss		irisp v.s. iriss	
MIR	0.281	0.314	0.350	0.340

表 8 において，“MIR” は「Mean Instance Recall」と呼ばれる評価尺度であり，取り出された文書の再現率を全てのトピック，全ての検索者について平均した値である．ここでは，ユーザとのインタラクションにより適合性フィードバックをかける以下の 3 つのシステムが評価の対象となっている．

iriss 検索された文書がトピックと関連があるか否かを検索者に判定させる．

irisa 検索者に質問タームとその重みを提示し，その重みを調整させる．

irisp 検索者に既存の質問タームまたは新しい質問タームの候補を提示し，質問タームとして削除または追加するべきかを判定させる．

irisp が TREC-7 で新たに提案されたシステムであるが，この 3 つのシステムの中では一番良い結果が得られていることがわかる．

- query track

query track の成績が上位の参加者とその結果を表 9 に示す[2]．

表 9: query track の結果

(Query set)		run by APL†	run by Cornell‡
2-3 words	by APL	0.0559	0.1051
	by Cornell	0.1055	0.2457
sentence	by APL	0.0477	0.1142
	by Cornell	0.1846	0.3367
manual	by APL	—	—
feedback	by Cornell	0.0917	0.2020
automatic	by APL	0.2577	0.3219
structure	by Cornell	0.2296	0.4586

† Johns Hopkins University ‡ Cornell University

※ 数値は全て Ave Prec である．

表 9 に示すように，各参加者は，作成した質問を自分の検索システムだけでなく，他の参加者の検索システムの入力としてテキスト検索を行い，その精度を評価している．しかしながら，このタスクは今回から新たに始められたためか，参加者が 2 団体しかいなかった．もっと多くの参加者が参加し，ひとつの質問を多くの検索システムで試すようにすれば，質問の品質をより厳密に評価できると考えられる．

- SDR track

SDR track の成績が上位の参加者とその結果を表 10 に示す．

ベースラインに比べて，参加者独自の音声認識システムを用いた場合 (recognizer) には，書き起こし文を用いた場合 (reference) と近い正解率が得られていることがわかる．また，SDR track のセッションでは，各参加者は検索システムそのものよりも音声認識システムの認識率をどのように向上させたかという点に重点を置いて発表していた．

[2]manual structured query による結果は今回は提出されなかったと思われる．

表 10: SDR track の結果

	Umass†	AT&T	Sheff‡
reference	0.5668	0.4992	0.4916
baseline1	0.5063	0.4700	0.4243
baseline2	0.4191	0.4065	0.3471
recognizer	0.5075	0.5120	0.4713

† Univ. of Massachusetts ‡ Univ. of Sheffield

※ 数値は全て Ave Prec である．

- VLC track

VLC track に関しては，参加者の検索結果の詳細を載せたデータシートが配布されなかった．参考までに，Univesity of Massachusetts の結果を表 11 に示す．

表 11: VLC track の結果

	VLC	10%	1%
適合率 (上位 20 位)	0.598	0.419	0.208
システム構築時間 (min)	2616	1569	160
検索時間 (min)	438	45	6

VLC track のセッションでは，検索時間や検索システム構築時間を短縮する手法に関する発表も数多く行われた．また，TREC-7 における VLC track の結果は，商用のサーチエンジンによる結果よりもはるかに良いことが報告された．

4 NEC・東工大グループの手法

本節では，TREC-7 の ad hoc task における NEC・東工大グループのテキスト検索手法の概要 [4] と，TREC-7 における評価結果について述べる．

我々が重点を置いたのは質問拡張 (query expansion) である．すなわち，トピックに含まれるタームのみだけでなく，それらと意味的に近い単語を質問タームに加えることにより，検索精度の向上を目指した．また，テキスト検索自体には特別な手法は導入せず，検索エンジンとして SMART[6] version 11.0 を使用した．

質問拡張を行うために，以下の 3 つのシソーラスを準備した．

1. WordNet[5]

Princeton 大学によって作成されたシソーラスである．WordNet では，単語間に存在する意味的な関係は記述されているが，意味的関係にある単語間の類似度は記述されていない．そこで，WordNet 上で関連があ

るとされる単語間の類似度は，後述する文書内共起情報に基づくシソーラスと述語・項関係に基づくシソーラスによって与えられる類似度の平均とし，それ以外の場合の類似度は 0 とした．

2. 文書内共起情報に基づくシソーラス

単語 A, B の類似度 $SIM_{ocr}(A,B)$ を式 (4) の Dice 係数と定義したシソーラスである．

$$SIM_{ocr}(A,B) = \frac{2 \times f(A,B)}{f(A) + f(B)} \tag{4}$$

式 (4) において，$f(A)$, $f(B)$ は単語 A, B のコーパス中における出現頻度を表わし，$f(A,B)$ は 2 つの単語が同じ文書中に出現する頻度を表わす．このシソーラス (単語間の類似度) を検索対象となる文書集合から学習した．

3. 述語・項関係に基づくシソーラス

述語・項関係にある単語対によって 2 単語間の類似度を定義したシソーラスである．まず，subject-verb，verb-object，adjective-noun といった述語・項関係にある単語間の類似度を式 (5),(6),(7) の Dice 係数と定義する．

$$C_{sub}(v_i,n_j) = \frac{2 \times f_{sub}(v_i,n_j)}{f(v_i) + f_{sub}(n_j)} \tag{5}$$

$$C_{obj}(v_i,n_j) = \frac{2 \times f_{obj}(v_i,n_j)}{f(v_i) + f_{obj}(n_j)} \tag{6}$$

$$C_{adj}(a_i,n_j) = \frac{2 \times f_{adj}(a_i,n_j)}{f(a_i) + f_{adj}(n_j)} \tag{7}$$

ここで，$f_{sub}(v_i,n_j)$ は subject-verb という関係で動詞 v_i と名詞 n_j が共起する頻度を，$f_{sub}(n_j)$ は名詞 n_i がある動詞の subject として出現する頻度を，$f(v_i)$ は動詞 v_i の出現頻度を表わす．式 (6),(7) 内の記号も同様の意味を表わす．

さらに，動詞 v_k の主語として現われる 2 つの名詞 n_i, n_j 間の類似度を式 (8)，動詞 v_k の目的語として現われる 2 つの名詞 n_i, n_j 間の類似度を式 (9)，形容詞 a_k の被修飾語として現われる 2 つの名詞 n_i, n_j 間の類似度を式 (10) のように定義する．

$$sim_{sub}(v_k,n_i,n_j) = min\{\, C_{sub}(v_k,n_i), C_{sub}(v_k,n_j) \,\} \tag{8}$$

$$sim_{obj}(v_k,n_i,n_j) = min\{\, C_{obj}(v_k,n_i), C_{obj}(v_k,n_j) \,\} \tag{9}$$

$$sim_{adj}(a_k,n_i,n_j) = min\{\, C_{adj}(a_k,n_i), C_{adj}(a_k,n_j) \,\} \tag{10}$$

最後に，名詞 n_i, n_j の類似度 $SIM_{prearg}(n_i,n_j)$ を，全ての動詞 v_k, 形容詞 a_k に対する式 (8),(9),(10) の類似度の平均値と定義する．

$$SIM_{prearg}(n_i,n_j) = \frac{\sum_{v_k \in V}\{sim_{sub}(v_k,n_i,n_j) + sim_{obj}(v_k,n_i,n_j)\} + \sum_{a_k \in A} sim_{adj}(a_k,n_i,n_j)}{2\,|V| + |A|} \tag{11}$$

式 (11) において，V と A はそれぞれコーパス中に現われる動詞，形容詞の集合を表わす．

このシソーラスを構築するために，まず検索対象文書集合に含まれる全ての文を Apple Pie パーザ [8] により構文解析し，subject-verb，verb-object，adjective-noun の関係にある 2 つ単語対を取り出した．そして，得られた共起データから，式 (11) の類似度を計算した．

質問拡張は，元の質問タームベクトル $\vec{q} = (q_1, \cdots, q_n)$ (q_i は 質問ターム t_i の重み) と類似度の高い名詞を新たな質問タームとして追加することにより行われる．$\vec{q}$ と新たな質問ターム t_j との類似度は式 (12) によって計算する．

$$simqt(q,t_j) = \sum_i q_i \times sim(t_i,t_j) \tag{12}$$

式 (12) の $sim(t_i,t_j)$ は，WordNet，文書内共起情報に基づくシソーラス，述語・項関係に基づくシソーラスによる t_i と t_j の類似度の平均である．そして，$simqt(q,t_j)$ の値が大きい上位 30 個のタームを新たに質問タームとして加える．このように，質問拡張を行う際に，既存のシソーラス (WordNet) とコーパスから自動構築されたシソーラスを組み合わせて計算した単語間の類似度を利用する点が本手法の特徴である．

TREC-7 では，我々は以下の 3 つの検索結果 (run) を提出した．

- nectitech (トピックのうち title のみを使用)
- nectitechdes (description のみを使用)
- nectitechall (title,description,narrative を使用)

各 run の評価結果を表 12 にまとめる．

表 12: NEC・東工大グループの評価結果

	Ave Prec	R-Prec
nectitech	0.1898 (—)	0.2403 (—)
nectitechdes	0.2584 (14 位)	0.2993 (9 位)
nectitechall	0.2565 (15 位)	0.2989 (10 位)

トピック中の title,description,narrative の全てを使った場合 (nectitechall) よりも description のみを用いた場合 (nectitechdes) の方が結果が良かったが，これは narrative に否定的な表現 (トピックに関連しない文書とは何かを説

明した文) が含まれていたためと推測される．また，表 12 中の ( ) 内は，automatic タスクに提出された 85 run 内における順位を示している．ad hoc task には表 12 に挙げた 2 つ以外の評価基準もあるが，全ての評価基準において，nectitechdesc, nectitechall の順位は上位 15 位以内であった．特に，nectitechall の評価基準「Rank of 1st Rel」[3] における順位は 4 位であった．

5 TREC-8 への展望

今回の会議では，次回のコンテスト TREC-8 に関する議論も行われた．その要点を以下にまとめる．

- ad hoc task について．

  TREC-8 では，新しい検索対象文書集合を用意する予定である．現在，文書を提供してくれる可能性のあるいくつかの団体と交渉を進めている．

- high precision track は TREC-8 からは廃止する．

- Q & A track というタスクを新たに追加する．

  これは，ad hoc task が単にトピックに関連のある文書をユーザに提示するだけであるのに対し，ユーザの知りたい事項をもっと直接的に答えるシステムを作り，その精度を競うタスクである．例えば,「中国の人口は何人ですか」という質問に対し，中国の人口について書かれた文書を提示するのではなく,「中国の人口は約 13 億人です」と答えるような検索システムを作り，その精度を評価する．

- VLC track の名称を web track に変更する．

  VLC track ではもともと web 文書を検索の対象にしていた．名称を web track に変更することにより，このタスクが www 検索技術を磨くためのタスクであるという立場を明確にした．

6 おわりに

本論文では，TREC-7 の各タスクの概要と評価結果について報告した．また，NEC・東工大グループのテキスト検索手法についても概説した．

ひとくちに情報検索といっても，その用途は多様であることから，様々な観点から情報検索システムの性能を評価することが望ましい．TREC では，多くの sub track が用意され，様々な評価基準から検索システムの性能を競い合っている点は評価できる．今回の TREC-7 の会議においても，メインの ad hoc task だけでなく，その他の sub track についてもかなりの発表時間が割り当てられ，活発な議論が行われた．ただ，sub track の参加者数が前回に比べて減少したことは少々残念なことである．

日本では，TREC のような大規模な情報検索のためのコンテストは行われていなかったが，1999 年には学術情報センターが開発したテストコレクション NTCIR[3] を用いたコンテストや，情報検索と固有名詞抽出を対象にしたコンテスト IREX[7] が開催される予定である．これらのコンテストが情報検索技術の進歩に大きく貢献することを期待する．

[3] 各トピックに関連のある文書のうち最も上位にランキングされた文書のランキングの平均値

参考文献

[1] *Text Retrieval Conference (TREC-7)*, 1998.

[2] 福本淳一, 関根聡, 江里口善生. MUC-7, Tipster 参加報告. 情報処理学会情報処理学会自然言語処理研究会, Vol. 98, No. 82, pp. 101–108, 1998.

[3] N. Kando et al. NTCIR: NACSIS test collection project. In *20th Annual Colloquium of BCS-IRSG*, 1997.

[4] Rila Mandara, Takenobu Tokunaga, and Hozumi Tanaka. The use of WordNet in information retrieval. In *Proceedings of the the COLING-ACL workshop on Usage of Wordnet in Natural Language Processing*, pp. 31–37, 1998.

[5] George A. Miller. Wordnet: An on-line lexical database. *International Journal of Lexicography*, Vol. 3, No. 4, 1990.

[6] G. Salton. *The SMART Retrieval System Experiments in Automatic Document Processing.* Prentice-Hall, 1971.

[7] 関根聡, 井佐原均. IREX: 情報検索、情報抽出コンテスト. 情報処理学会情報処理学会自然言語処理研究会, Vol. 98, No. 82, pp. 109–116, 1998.

[8] Satoshi Sekine and Ralph Grishman. A corpus-based probabilistic grammar with only two non-terminals. In *Proceedings of the International Workshop on Parsing Technologies*, 1995.