情 報 学 基 礎 51－15
自 然 言 語 処 理 127－15
（1998. 9. 18）

IREX: 情報検索、情報抽出コンテスト

関根 聡† 井佐原 均‡
sekine@cs.nyu.edu isahara@crl.go.jp
† ニューヨーク大学 コンピュータサイエンス学科
‡ 郵政省通信総合研究所

９９年４月に行なう予定の情報検索、情報抽出のコンテスト（Information Retrieval and Extraction Exercise = IREX）の目的、内容の紹介を行なう。第一回目の今回はＩＲとＮＥの二つの課題を予定している。情報検索（ＩＲ）の課題では数百の記事を二年分の新聞記事から検索する事を目的とする。一方の課題は、固有表現抽出（ＮＥ）と呼び、組織名、人名、地名、固有物名などの固有名詞的表現や時間表現、数値表現を、文章から抽出する事を目的とする。この課題では、対象文章をあるトピックに限定した物と、自由トピックの物と二種類の物を用意する予定である。

IREX: Information Retrieval Extraction Exercise

Satoshi Sekine† Hitoshi Isahara‡
†Computer Science Department, New York University
‡Communications Research Laboratory, Ministry of Posts and Telecommunications

This paper introduces a contest scheduled to be held on April 1999. This contest focuses on Information Retrieval and Information Extraction, which is similar to TREC and MUC, the contests held in US. For the first round, there will be two tasks. One is a simple IR task, to retrieve several hundreds of documents from 2 year newspaper articles. The other is "Named Entity" task, in which proper names, like organization, person, location or artifact, or time and number expression should be extracted from documents. There are two sets of document, one is general documents and the other is topic dependent set.

1 はじめに

この論文では、９９年４月に行なう予定の情報検索、情報抽出のコンテスト（ＩＲＥＸ）の目的、内容の紹介を行なう。

情報化が完全に浸透した現在、個人が扱わなければいけないコンピュータ上の情報量は、その能力の限界をはるかに越える程になっている。特に、新聞記事やホームページなどのテキストの広大な海から自分に有用な情報を見つける事は至難の技である。このような状況に鑑み、情報検索や情報抽出といった技術が欧米、日本などを初め世界中で研究、開発されている。特に米国で行なわれているＭＵＣやＴＲＥＣといったコンテストは、この分野の技術を推進するために重要な役割を担ってきたと認識されている。日本でも、非常に優れた研究が行なわれてきたが、主に個々の研究所や大学等で行なわれてきたため、共通のプラットフォームでのシステム間の評価や問題点や議論の共有という事が難しい状況であった。そこで、このような問題点を認識している研究者、開発者を募り、共通のプラットフォームでのそれぞれのシステムの評価を行なう場を提供したいと考えている。ＩＲＥＸにおいては、単なる評価だけではなく、そこから派生する以下のような種々の効果を期待している。

- 問題点の共有と、それに基づくこの分野の飛躍的な進歩、発展
- 情報検索、情報抽出の研究の認知、宣伝、推進
- 研究者の裾野の拡大
- 膨大なデータの蓄積、テキストデータベースの拡大
- 長期的なプロジェクトへの緒

上記のような問題点を認識している研究者、開発者ならば、企業、研究機関、大学、個人を問わず参加資格がある。参加料は原則的に無料であるが、対象データは毎日新聞(９４，９５年版：有料)を利用する。

次章では、まず、これまで主に米国で行なわれてきたコンテストについて、その功罪を筆者なりに簡単にまとめた。ＩＲＥＸはこれらの考察や反省の上に立ったコンテストになるように設計していく。

2 コンテストの功罪

この章では、評価やコンテストを基にした方法論のメリット、デメリットをまとめ、デメリットを克服するような工夫についても述べたい。まず、これまで行われてきた、または行なわれる予定のコンテストの例を列挙する。内容や詳細については、それぞれの論文あるいはホームページを参照されたい。

コンテストの例

- **ATIS**: (1988-1995) 特定目的の音声認識 [6]
- **CSR**: (1986-1998) 大規模音声認識 [7] [5]
- **DARPA-MT**: (1994,1996) 機械翻訳 [14] [15]
- **DiaLeague**: (1995-1997) 自然言語対話システム [1]
- **MUC**: (1987-1998) 情報抽出 [2] [3]
- **Parseval**: (1991) 構文解析 (非公開)
- **SENSEVAL**: (1998) 語義の曖昧性解消 [4]

- **SUMMAC**: (1998) 要約 [9]
- **TDT**: (1998) トピック検知、追跡 [10] [13]
- **TREC**: (1992-1998) 情報検索 [8]

メリット

- サブゴールの設定
  自然言語処理の扱う対象は広くて深いため、それを一気に解決する事は不可能に近い。したがって、最終目的までにサブゴールを設定し、それぞれのゴールまでの技術の向上を計り、最終的なシステム全体の精度の向上を計るという手法は、研究者へのよい意味での動機付けになる。
- 技術向上の目安
  絶対的な基準を導入しての評価は、時を越えたシステム間の評価を可能とし、技術の向上の目安となる。
- アイデア、情報の交流
  同じタスク、同じ評価基準での評価は、それに参加する研究者の問題意識の共有化を促し、アイデアや情報の交流を促進する。
- データの蓄積
  評価の課題となったデータは、その後のシステム開発にとっての貴重なデータとなり有用である。また、複数の団体が参加する評価により、単独でデータの作成を行なうのに比較して、大規模なデータがより早く、より容易に集まる。
- システム、技術の比較
  システムや技術を比較する事によって、役に立つものの発見が容易にでき、技術の進歩に継がる。

デメリット

- 取組が狭い範囲に限られる
  自然言語に関する課題は広く、コンテストで設定できる課題は限られる。そのため、多くの努力が狭い範囲に限られてしまう可能性がある。これに対しては、課題を基礎的でかつ広い適応範囲を持つ物に設定する、または、複数の課題を設定する等の努力が必要である。
- 要素技術の評価ができない
  課題は大概、システムの出力結果に対して設定されるため、その中で利用されている個々の要素技術の評価ができない可能性がある。これに対しては、システム開発者の自己申告による要素技術の評価などを行なう事によって、改善の可能性がある。
- スコア重視、技術軽視
  開発者がスコアを重視しすぎる事によって、本当の技術開発がさまたげられる可能性がある。これに対しては、評価の基準をきちんと設定し、アドホックなテクニックによりスコアの向上ができないようにする必要がある。ただし、スコアを向上する為のテクニックを探すためには課題に対する考察が必要であり、そのような中から新しい技術が生れる可能性もある。また、実際のMUC，TREC，SUMMACを見ているとスコア偏重の様子は少ないように感じる。

工夫

- 基本的な課題の設定
  例えば、ＭＵＣでは情報抽出の前段階として、固有名認定、基礎的情報の抽出、照応解析、のような基本的な課題を作り、一般性を高めるような努力が行なわれている。

- サンプルデータを利用したシステム比較
  これも、ＭＵＣで用いられている手法であるが、技術報告の際に、指定されたデータに対してのシステムの動作の状況を報告するように推奨している。これによって、まったく同じ環境で、各システムがどういった理由でどのように動いたかといった細かな比較ができる事になる。

- 要素技術の比較
  ＭＵＣでのサブタスクの設定もこの範疇に入るが、９４年の音声の評価で、評価が終った後にアンケートが回り、各参加団体が自発的に、各要素技術の寄与度を報告するという事が行なわれた。ほぼ、すべての参加者が協力的であり、この結果得られた 10 参加者と 30 程度の要素技術のマトリックスは非常に興味深いものになっている。

- 比較実験
  ＴＲＥＣ，ＳＵＭＭＡＣでは、複数の比較実験を行ない、違った環境での評価を実現している。例えば、ＴＲＥＣ－７では、基本的な課題の他に７つの比較実験の課題が設定されている。

- 多種の評価基準
  今回のコンテストでもそうであるが、情報検索、情報抽出での評価は一般的に再現率、適合率を基にして行なわれている。しかし、これだけでは十分でないのは明らかであり、課題に合せて色々な評価基準が提案され、実際に使用されている。目的に合せた評価基準の設定が重要である。

3 課題

以下の２種類の課題を予定している。参加者はどちらかひとつ、または両方の課題に参加できる。

- 固有表現抽出 (ＮＥ)
  情報検索、情報抽出の基礎技術として、新聞記事からの組織名、人名 、地名等の自動抽出を行なう。以下に示すように、SGML タグを対象場所に振り、提供ツールを使って得られたオフセット情報を提出してもらう事で情報のやりとりを行なう。対象記事はトピックを限定した記事 (トピックは評価の２週間前に発表) と、自由トピックの記事、それぞれ 50 記事づつで行なう事を予定している。

  固有表現の抽出では以下の８種類の固有表現の抽出を行なう。

	開始位置タグ	終了位置タグ
* 固有名詞的表現		
* 組織名、政府組織名	<ORGANIZATION>	</ORGANIZATION>
* 人名	<PERSON>	</PERSON>
* 地名	<LOCATION>	</LOCATION>
* 固有物名	<ARTIFACT>	</ARTIFACT>
* 時間表現		
* 日付表現	<DATE>	</DATE>
* 時間表現	<TIME>	</TIME>
* 数値表現		

* 金額表現	<MONEY>	</MONEY>
* 割合表現	<PERCENT>	</PERCENT>

それぞれの固有表現文字列の開始、終了位置に、システムは重複や入れ子のない唯一のタグのペアを振る。もし、表現が重なっている場合は、原則的に長い単位の表現を抽出する。

- 情報検索（ＩＲ）

  検索課題として指定された内容の記事を２年分の新聞記事の中から自動的に検索するという課題である。質問の総数は30個を予定している。各参加者は、検索課題に関係する記事のＩＤを確信度の高い順に300記事（予定）まで提出する。評価は、再現率と適合率で行なう。

  具体的には、検索課題に示された内容の新聞記事を毎日新聞の９４，９５年版ＣＤ－ＲＯＭに含まれる記事から検索する。検索対象記事は提供ツール（mai2sgml.pl）によって変換されたデータに対して行なう。オリジナルのＣＤ－ＲＯＭに含まれているキーワード等の情報は使用してはいけない。検索課題にある情報のどの部分をどのように使用するかは参加者の自由意志である。

２つの課題とも、基礎的な技術を評価する事を目的にしている。例えば、情報検索や情報抽出においては、ユーザーインターフェースやユーザーの意図の導出が重要であるが、今回行なうような基礎的な技術の評価や、それに伴なうデータの作成、蓄積はより高度な技術の発展にも役に立つと確信している。また、将来においては、そのような高度な課題も含めていける事を希望している。

4　日程

1998年6月30日	ＮＥ定義、ＩＲ検索課題の叩き台公開
1998年7月31日	第一次参加申し込み締切
1998年9月16日	ＮＥ定義、ＩＲ検索課題の議論終了
1998年10,11月頃	予備試験（任意参加、結果非公開）
1999年2月28日	最終参加申し込み締切 （参加申し込みの方法については、以下に示すＩＲＥＸのホームページを参照されたい。）

＝＝本試験＝＝

1999年4月5日	ＩＲ検索課題配布
1999年4月12日	ＩＲ検索結果提出（日本時間23時59分まで）
1999年4月13日	ＮＥ評価対象データ配布
1999年4月16日	ＮＥ抽出結果提出（日本時間23時59分まで）
1999年9月	ワークショップ（予定）

5　その他の情報

この章では、ＩＲＥＸに関する詳細、現在の方針を示す。

- データ公開

  作成したデータは基本的にＩＲＥＸ参加、不参加に関わらず、一般に公開する予定にしている。ただし、コンテス

トへの参加を促すためにも参加者と非参加者の間でなんらかの差をつけたいと思っている。情報検索に関するデータについては、ＩＲＥＸ参加者に対しては、どのシステムがどの回答を出したかという情報を付与する事を予定している。

- ＢＭＩＲとの関係

  立案段階からＢＭＩＲの関係者には色々相談させてもらったり、情報をいただいている。基本的にＢＭＩＲ自身をそのまま使う事は(答が知られているので)難しいが、例えば、質問文を参考にさせてもらったり、そのデータをシステム開発用のデータとして利用する事はできると考えている。実際、ＩＲの検索課題のフォーマットはＢＭＩＲのフォーマットを踏襲する方向で検討されており、以下に記する予備実験では質問文をそのまま使わせてもらう事も検討されている。

- ＭＵＣ／ＴＲＥＣとの関係

  アメリカ政府等からの資金提供といった話は、可能だという話をいただいているが、今回は遠慮した。また、ＴＲＥＣの一部としてやらないかという打診を受けたが日程上の問題などのため、今回は少くとも一緒にはできない。ただ、Tipsterは９８年の秋に終了するため、次回というのはなさそうである。なお、ＭＵＣの議長であるグリッシュマンには、顧問になってもらった。

- ＮＡＣＳＩＳコレクションとの関係

  現在学術情報センターの研究グループで，学術論文抄録を用いた情報検索システム評価用テストコレクション[11][12]を構築しており，その一環としてテストコレクションのデータを用いたコンペティションが企画されている。ＩＲＥＸとしては、このコンペティションに対し全面的な協力関係を持ち進めていきたいと思っている。実際に、ワークショップ(成果報告会)は共同で開催する事を計画中である。

- 正解作成

  予備試験および、トレーニングデータの正解作成は、ボランティアベースで行なう。参加希望者はsekine@cs.nyu.eduまで。本試験のＮＥの正解は限られた数人で作成する。本試験のＩＲの正解作りは、基本的に学生のアルバイトが行なう。すべての解答は２人のアルバイトが関わり、同一の答を出した場合は、それをそのまま解答とし、異なった場合には、その２人が話し会う事によって解答を作るという方法を取る。どうしても合意に達っしなかった場合には関根などが判断をするという計画である。

- 予備試験

  予備試験は、各参加者に課題やデータ交換の方法に慣れてもらうと共に、運営側としても正解判定の練習、運営の方法の検討などの為に行なう。ＩＲの場合には、検索課題数は極端に少なくする予定である。ＮＥについては、40記事くらいで予備試験を行う予定である。

  共に作成したデータはできあがり次第、ＩＲＥＸ参加者には公開するので本試験のトレーニングデータとして使用して構わない。(ＮＥについては直後、ＩＲについても一ヶ月くらいで公開したいと考えている)ＩＲＥＸ不参加の人に対しては、ＩＲＥＸ本試験のデータと共に公開したいと考えてはいるが、その方法がどのような形になるかはまだ決めていない。

  結果非公開とあるが、当事者には結果をお知らせする。ＮＥについては現在スコアラーを作成しており、それも公開するので、正解データがあれば、自分で評価できるようになる筈である。

- 将来の方向、ＩＲＥＸの成功の基準

  今回のコンテストは将来の大きなコンテストの為の緒になればというスタンスである。

  各参加者にとっては、色々な成功の基準があると思われるが、ＩＲＥＸというプロジェクトにとっては、我々は、ＩＲＥＸ－２が今回のＩＲＥＸ以上の規模で実現できるような道筋を付けられたら成功だと考えている。つまり、魅力的な内容、議論、データ等を作りあげる事が目標である。この為には多くの方の参加、協力が不可欠である。

6 運営

以下に運営上の情報、その他を載せる。より詳細な情報については、以下に示すホームページを参照されたい。

- **主催：** ＩＲＥＸ実行委員会
- **ホームページ：** http://cs.nyu.edu/cs/projects/proteus/irex
- **メイリングリストアドレス：** irex@karc.crl.go.jp
- **実行委員長：** 関根聡 (ＮＹＵ)、井佐原均 (通信総研)
- **顧問：** 長尾眞 (京大)、田中穂積 (東工大)、Ｒ．グリッシュマン (ＮＹＵ)、石川徹也 (図情大)
- **実行委員：** 徳永健伸 (東工大)、黒橋禎夫 (京大)、奥村学 (北陸先端大)、野畑周 (東大)、北研二 (徳島大)、乾健太郎 (九工大)、峯恒憲 (九大)、中川裕志 (横国大)、藤井敦 (図情大)、若尾孝博 (ＴＡＯ)、神門典子 (学情)、橋田浩一 (電総研)、隅田英一郎 (ＡＴＲ)、村田真樹、内元清貴 (通信総研)、野口直彦 (松下)、奥村明俊、福島俊一 (ＮＥＣ)、小川泰嗣 (リコー)、酒井哲也 (東芝)、福本淳一 (沖)、木谷強、江里口善生 (ＮＴＴデータ)、中渡瀬秀一 (ＮＴＴ)、豊浦潤 (三菱)、落谷亮 (富士通)、荻野紫穂 (ＩＢＭ) (順不同、敬称略)

参考文献

[1] "DiaLeague" http://www.etl.go.jp/etl/nl/dialeague/ (1997)

[2] "MUC-6" http://cs.nyu.edu/cs/faculty/grishman/muc6.html (1995)

[3] "MUC-7" http://www.muc.saic.com/ (1998)

[4] "SENSEVAL" http://www.itri.bton.ac.uk/events/senseval/ (1998)

[5] "Speech Recognition Workshop Proceedings" http://www.itl.nist.gov/div894/894.01/proc/index.htm (1998)

[6] Proceedings of the Spoken Language Systems Technology Workshop *Morgan Kaufmann publishers* (1995)

[7] Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop *Morgan Kaufmann publishers* (1998)

[8] "Text Retrieval Conference" http://trec.nist.gov/ (1998)

[9] "Tipster Text Program" http://www.tipster.org/ (1998)

[10] "1998 Topic Detection and Tracking Project" http://www.itl.nist.gov/div894/894.01/tdt98/tdt98.htm (1998)

[11] Kageura, K.; Koyama, T.; Yoshioka, M.; Takasu, A.; Nozue, T.; Nozue, T.; Tsuji, K. "NACSIS Corpus Project for IR and Terminological Research", *Natural Language Processing Pacific Rim Symposium*, pp493–496, (1997)

[12] Kando, N.; Koyama, T.; Oyama, K.; Kageura, K.; Yoshioka, M.; Nozue, T.; Matsumura, A.; Kuriyama, K. "NTCIR : NACSIS Test Collection Project"[Poster] *The 20th Annual Colloquium of the British Computer Society Information Retrieval Specialist Group on Information Retrieval Research*, p.25–27 (1998)

[13] Charles L.Wayne: "Topic Detection and Tracking: A Case Study in Corpus Creation & Evaluation Methodologies" *Proceedings of the First International Conference on Language Resources & Evaluation* (1998)

[14] White, J and T. O'Connell: "The ARPA MT evaluation methodologies: evolution, lessons and future approaches" *Proceedings of the 1994 Conference Association for Machine Translation in the Americas* (1994)

[15] White, J and T. O'Connell: "Adaptation of the DARPA machine translation evaluation paradigm to end-to-end systems" *Proceedings of the 1996 Conference Association for Machine Translation in the Americas* (1996)