自 然 言 語 処 理 143−2
（２００１．５．３１）

固有表現に着目した双方向イベントトラッキングとその文書要約への応用

賀沢秀人　　平尾努
NTTコミュニケーション科学基礎研究所

概要

新聞記事などの時系列文書から特定のイベントに関連する続報記事を検索するイベントトラッキングと呼ばれる問題に関し、続報だけでなく既報記事も検索対象とした双方向イベントトラッキングという新しい問題設定を提案する。さらに、固有表現の重複に着目した新しいイベントトラッキング手法を提案、1ヵ月分の新聞記事に適用し既存手法に対する有効性を検証した。また、冗長度を考慮した文書選択基準であるMMRと提案手法によるトラッキング結果とを組み合わせ、要約に適した文書選択を試みた。その結果、MMRにより冗長性が排除される傾向が見られる一方、検索上位に誤りが増加するという問題点が明らかになった。

Bidirectional Event Tracking and Its Application to Multiple Text Summarization

Hideto KAZAWA and Tsutomu HIRAO
NTT Communication Science Laboratories

Abstract

In this report, we propose “bidirectional event tracking” task and describe primary results of our tracking method and its application to multiple text summarization. “Bidirectional event tracking” addresses the task of retrieving all document relating to a given document in the sense that the events described in these documents are directly linked regardless of its publication date. Our tracking method retrieves only the document which include named entity expression common with a given document. We tracked four events in one-month news articles, which amounts to 3,827 articles, and the results show that the requirement of common named entities improves tracking accuracy. We also reported our primary attempt to select non-redundant document from tracked document for summarization.

1　はじめに

近年、コンピュータやインターネットの普及により、膨大なテキスト情報が利用できるようになりつつある。このような状況で、これらの情報の中から求める情報をいかに捜し出すか、ということが大きな問題となっている。

この問題を解決することを目的とした技術に、情報検索がある。多くの場合、情報検索ではユーザーが求める内容を自然言語で入力する必要があるが、人間は、望んでいる内容を適切な検索要求として常に言語化できるわけではないため、求めるテキストをうまく捜し出せない場合も多い。

このような状況から、ユーザーが求める内容に近い内容を持つテキストをシステムに提示し、それを元に関連する文書を捜し出す、関連文書検索と呼ばれる技術が盛んに研究されてきた。関連文書検索では、ユーザーが検索要求を明示的に言語化する必要がないため、検索対象について漠然としたイメージしか持たない場合でも、検索を実行できるという長所がある。しかし、検索したいテキストに関する情報が明示

されない分、出力として何を正解とするかが比較的曖昧であり、系統的な研究を行う場合に問題となるという面もある。

例えば、2000年のアメリカの大統領選挙に関する文書が、関連文書検索の入力として与えられた場合、その出力として、2000年の選挙自体に直接関わる文書が適切なのか、それとも、今までに行われたアメリカの大統領選挙に関するものならば何でも良いのか、は、元となる文書からだけでは判断できない。

こうした問題点を踏まえ、イベントトラッキング[1]というタスクが、ここ数年の間に注目されてきている [1]。イベントトラッキングでは、検索対象を「元となる文書に関連する文書」という広い定義から、「元となる文書に書かれたイベントと直接関連したイベントについて書かれた続報」に、制限している。そのため従来の関連文書検索のタスクと比べ、正解の基準が比較的明確であるという長所がある。

我々は現在、続報に加え既報の検索も行う双方向イベントトラッキングについて研究を行っている。本報告では、我々の提案する固有表現の重複に着目したトラッキング手法を、双方向イベントトラッキングに対し適用した結果について報告する。また、双方向イベントトラッキングの応用として、複数文書要約の入力にトラッキング結果を用いる場合についても簡単に考察する。

本報告の構成は次の通りである。第2章では関連する従来研究について述べる。第3章では本報告で用いる双方向イベントトラッキング手法を説明する。第4章では新聞記事を用いた実験結果について述べる。第5章では複数文書要約との関連について考察する。

2 既存研究との違い

2.1 トラッキングの双方向性

米国においては、1998年からNIST[2]の主催により Topic Detection and Tracking(TDT) Workshop が開催されている [1]。TDTの参加者には共通の文書集合の中から、与えられたイベントについて続報記事を検索する、という課題が与えられる。(課題の詳細は [2] を参照) TDTの文書集合には、新聞記事の他、ラジオやテレビで流れるニュースを自動音声認識したもの (もしくは、それを人手により修正したもの) も含んでおり、英語以外に中国語のデータも扱っている点が特徴的である。

TDTでは検索対象は元文書の続報に限られている。[3] しかし、実際にイベントのトラッキングが必要になる場面では、過去に遡って既報を探す要求も多いと考えられる。また、一般にイベントは時間が経つにつれて他のイベントを派生する場合が多い。そのような場合に過去に遡って検索を行おうとすると、続報を探すときとは異なる技術的問題があることも十分考えられる。そこで、我々は続報だけでなく既報をも検索対象とする「双方向イベントトラッキング」を研究の目標とすることにした。

2.2 固有表現の扱い

(双方向) イベントトラッキングにおいては、そのイベント固有の主体や場所・時刻などの固有表現を正確に特定することが有効と考えられる。しかし、既存の研究では、固有表現は普通名詞と同様に扱われることが多い。

例えば、新谷らの研究では、JUMANによる形態素解析結果で、サ変名詞・人名・地名・普通名詞・未定義語とされた語を、文書の特徴ベクトルとして用いている [3]。しかし、各単語の重みづけは頻度・位置に基づいたもので、固有表現だけに特別の重みづけを行うということはしていない。

また、大竹らは [3] と同様の続報検索を、名詞連接を特徴ベクトルに加えて行っている [4]。名詞連接の中には、形態素解析辞書に登録されていない固有表現が含まれる可能性があるが、[4] ではそれらに対し特別の重みづけは行っていない。

本報告では、[5] の提案手法により抽出された固有表現に対して、特別の重みづけを行うトラッキング手法を提案する。

[1]後述の TDT では「トピック」トラッキングという呼び方がされているが、タスクの内容を考えると「イベント」トラッキングの方が適切と思われるので、ここではそのようにした。

[2]National Institute of Standards and Technology

[3]次節で紹介する [3, 4] の研究でも、続報のみが検索の対象となっている。

3 双方向イベントトラッキング手法

本研究の提案手法は、

1. 前処理として検索対象の全てのテキストを特徴ベクトルに変換する。
2. 元記事と全ての検索対象の対について、特徴ベクトルの関連度を計算する。

という二段階で実行される。以下、各段階について説明する。

特徴ベクトルへの変換

1. 記事本文を JUMAN(ver. 3.6)[6] で形態素解析する。
2. 記事本文から磯崎の方法 [5] で、固有表現を抽出する。
3. 固有表現の開始・終了位置と、形態素の開始･終了位置が一致する部分について、JUMAN の出力結果を抽出した固有表現で置き換える。
4. 置き換えた形態素解析結果から、名詞・動詞・固有表現を抜きだし、TF・IDF 法により重みづけを行う。具体的には、記事 x における、単語 m の重み $W(x,m)$ を次のように定義する。

$$W(x,m) = \frac{TF(x,m)}{1 + log(1/DF(m))}$$

ここで $TF(x,m)$ は記事 x における単語 m の出現回数、$DF(m)$ は記事集合の中で単語 m を含む記事の数である。

関連度の計算

記事 x と y の関連度 $R(x,y)$ の計算方法として、次の 3 種類を用いる。

1. COS

$$R_{COS}(x,y) = \frac{\sum_{m \in x and m \in y} W(x,m)W(y,m)}{\sqrt{\sum_{m \in x} W(x,m)^2}\sqrt{\sum_{m \in y} W(y,m)^2}}$$

これは、情報検索やテキスト分類で標準的に使われる cosine measure と同一の指標であり、提案手法との比較に用いる。

2. NE

$$R_{NE}(x,y) = \begin{cases} R_{COS}(x,y) \\ (x \text{ と } y \text{ に共通の固有表現がある}) \\ 0 \\ (\text{その他の場合}) \end{cases}$$

固有表現は基本的に表記揺れの心配がなく、またイベント固有の単語であることも多い。そのため一つ以上の固有表現を共有することを条件とする指標を考える。

3. AGENT

$$R_{AGENT}(x,y) = \begin{cases} R_{COS}(x,y) \\ (x \text{ と } y \text{ に共通の人名・組織名がある}) \\ 0 \\ (\text{その他の場合}) \end{cases}$$

固有表現の中でも人名・組織名は、そのイベントにおける主体を表している場合が多いと考えられる。そこで一つ以上の人名・組織名を共有することを条件とする指標を考える。

なお、全ての固有表現は固有表現抽出で得られたもののみ用い、JUMAN により固有表現とされた名詞 (人名・地名など) は普通名詞と同様に扱った。

また、最終的な出力にあたっては、適当な閾値を設けそれを上回る関連度を持つ記事を出力する。

4 実験

実験には CD-ROM 版毎日新聞の 1999 年 7 月一ヵ月分の記事のうち、掲載面が 1 面、 2 面、3 面、経済、国際、社会を選び、さらにそこから、著作権の関係で本文が収録されていない記事と、コラムや社説に類する記事を除いた、3827 記事を用いた。

検索するイベントとしては、次の 4 つのイベントを用い、上記記事中から関連する記事を、人手により全て捜し出した。

イベント **A**

ケネディ元大統領の長男が操縦する飛行機が行方不明となった事件。関連記事数は 15。

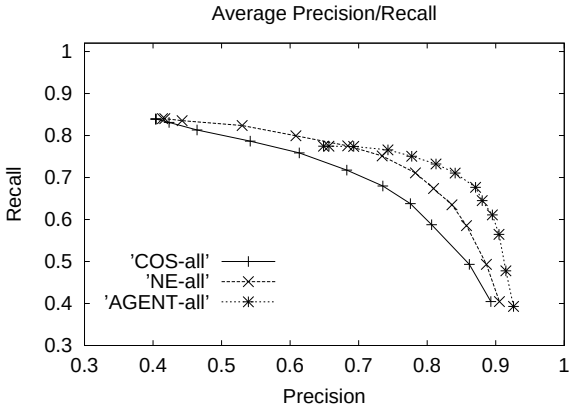

図 1: 関連度ごとの適合率／再現率

イベント **B**

全日空機がハイジャックされ、機長が刺殺された事件。関連記事数は 32。

イベント **C**

当時の台湾総統が中国との関係を「対等な国と国との関係」と発言した事件。関連記事数は 31。

イベント **D**

日本初の銀行免許取消し事件。関連記事数は 15。

実験は、上記各イベントに関する全ての記事を元記事として、それと関連する記事を前節で説明した 3 つの関連度を用いておこなった。したがって、各関連度を用いて 15+32+31+15 = 93 回の双方向イベントトラッキング実験をおこなったことになる。

図 1 は、3 つの関連度ごとに閾値を変化させた場合の適合率と再現率のグラフであり、COS、NE、AGENT の各曲線が R_{COS}、R_{NE}、R_{AGENT} の各適合率を使用した場合に対応する。

ここで、適合率・再現率は次のように定義される。

$$\text{適合率} = \frac{\text{トラッキング結果に含まれる正解数}}{\text{トラッキング結果の数}}$$

$$\text{再現率} = \frac{\text{トラッキング結果に含まれる正解数}}{\text{全関連記事数}}$$

図 1 から、R_{AGENT}、R_{NE}、R_{COS} の順に精度が高いことがわかる。これは、固有表現、特にイベントの主体や対象となりやすい人や組織の名称は、そのイベントを特徴づける語となりやすいことを示していると考えられる。

次に、各イベントごとに固有表現を重みづけした効果に違いがあるかを調べた。図 2、3、4、5、に、それぞれイベント A、B、C、D について、3 種類の関連度を用いてトラッキングを行った結果を示す。

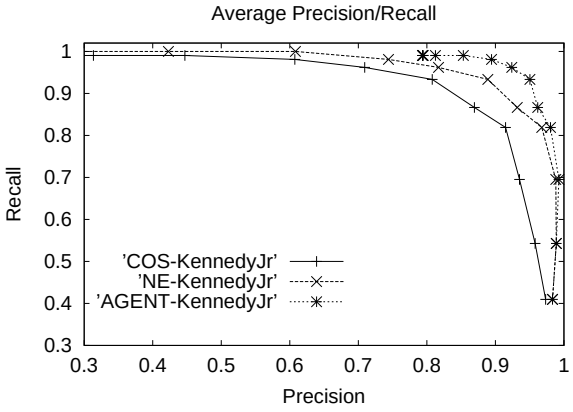

図 2: イベントごとの効果の違い (イベント A)

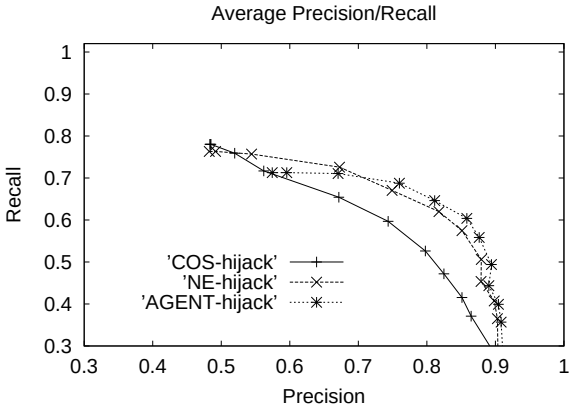

図 3: イベントごとの効果の違い (イベント B)

これらの図から、どの固有表現がトラッキングに好影響を及ぼすかは、イベントによりかなり異なることがわかる。イベント A、B に関しては、図 1 で見られた傾向と同様に、R_{AGENT}、R_{NE}、R_{COS} の順に精度が高くなっている。しかし、イベント C では固有表現全体に重みづけた R_{NE} を用いた場合に、大きな精度改善が見られている。

イベント C に関連する記事を調べてみると、ほとんどの場合「中国」「台湾」「中台」という

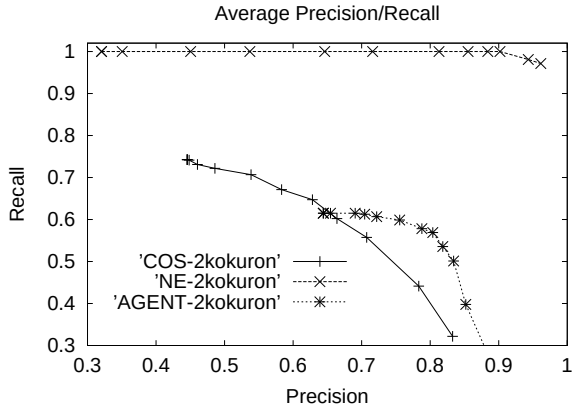

図 4: イベントごとの効果の違い (イベント C)

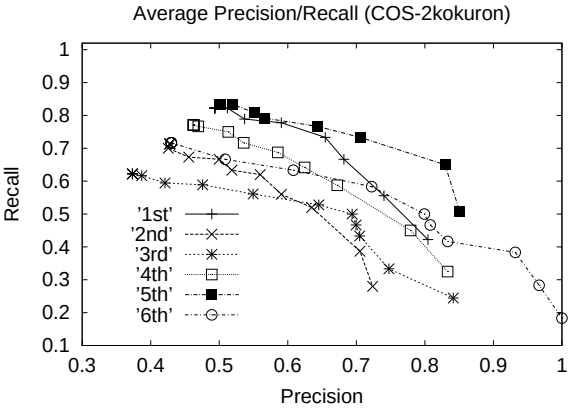

図 6: 元記事の時期の違いによる差

固有表現が現われており、R_{NE} を用いた場合、これらの語を含まない記事は拾わないため適合率が良くなっている。またイベント C においては、問題となった発言の主である「李登輝」という人名も重要な単語であるが時間が経過するにつれ、発言の主は書かれず「中台問題」といった形で、問題の名称だけが記述される場合が多かった。そのため R_{AGENT} を使った場合は R_{NE} を用いた場合ほど、精度は上がらなかったと考えられる。

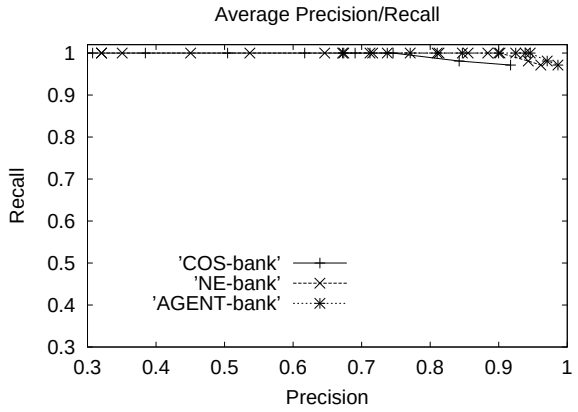

図 5: イベントごとの効果の違い (イベント D)

イベント D に関しては、関連度の違いにより大きな違いは見られなかった。

次に、トラッキングの元となる文書の時期により、トラッキング精度に系統的な違いが生じるかを調べた。図 6 にイベント C の関連記事を、第一報から三日毎に分け、それぞれの期間を元記事とした場合のトラッキング結果を示す。関連度としては R_{COS} を用いた。「1st」「2nd」… の順に最初の三日間、次の三日間…となっている。この結果からは、元記事の時期によりトラッキング精度に系統的な差は認められなかった。

5 トラッキング結果の要約

イベントトラッキングで得られるものは文書の集合であり、そのままではイベントの全体像を知ることが難しい場合が多い。このような場合、トラッキングで得られた複数文書を要約して提示することで、ユーザーのイベント理解を支援できると考えられる。本節では、現在、我々が進めているトラッキング結果の要約研究について簡単に報告する。

5.1 冗長性の排除

一般にトラッキングで得られる関連文書群は冗長性が高い場合が多い。例えば、あるイベントに関する新聞記事では常に新しい情報だけが提供されるわけではなく、以前の内容と重複があることが普通である。一方、要約という観点から見ると、このような重複のある入力は要約を冗長にする傾向がある。そこで、トラッキング結果を要約する際には、入力の冗長性をできるだけ排除することが大きな課題となると考えられる。

冗長性を考慮した文書の選択基準として MMR (Maximal Marginal Relevance) という指標が提案されている [7]。MMR は、検索要求との適合度から既に選択した文書との重複度

を引いたものであり、次の式で定義される.

$$MMR(Q, R, S) = \arg\max_{D_i \in R \setminus S}[\lambda Sim_1(D_i, Q) - (1-\lambda) max_{D_j \in S} Sim_2(D_i, D_j)]$$

ここで Q は検索要求、R は文書の全体集合、S は既に選択された文書、$Sim_1(D, Q)$ は文書 D の検索要求 Q に対する適合度、$Sim_2(D_1, D_2)$ は文書 D_1 と D_2 の重複度を表す指標である。

MMR の大きな順に文書を並べ直すことで、冗長性の低い (別の言い方をすれば、新規性の高い) 文書が上位に現われることが期待される。

5.2 MMR を用いた要約実験

MMR を用いて、イベントトラッキング結果から冗長性の低い要約が作成できるかを調べるために、次の実験を行った。

- 入力として、関連度として R_{COS} を用いた場合のトラッキング結果の上位 50 件を使用する。トラッキングを行う元記事は、イベント A〜D の最初の記事とした。
- 次の手順により要約を行う。
  1. トラッキングされた各文書について、先頭から n 文[4]を抽出し、その文書の抜粋とする。
  2. 作成した抜粋の集合から、MMR の大きさ順に上位 15 件 (トラッキング結果の 30%) の抜粋を選択する。ただし MMR の計算において、Q をトラッキングの元となった文書の抜粋、$Sim_1(D, Q)$、$Sim_2(D_1, D_2)$ として各抜粋中の重なり単語数を用い、$\lambda = 0.7$ とした.
  3. 選択された抜粋を時系列順に並べ、最終的な要約とする。

以下では紙面の都合上、MMR による抜粋の選択結果についてのみ考察する。

表 1 は、MMR により選択された上位 15 件と、トラッキングの出力上位 15 件とで、関連記事でないものが何件含まれていたかを示したものである。

表 1 から、MMR による選択で上位に誤った記事が多く現われたことが分かる。これは、

[4] n は対象文書の全文数の 1/10 とした。

イベント	MMR	トラッキング結果
A	6	3
B	5	1
C	5	1
D	4	3

表 1: 上位 15 記事中の誤り件数

MMR が単語の出現パターンが大きく異なる文書に対して大きな値をとるために、冗長性の低い文書だけでなく、関連文書でないものも選択してしまったことが原因であると考えられる.

しかし、各イベントに関して結果を調査したところ、イベント C に関しては、トラッキングの上位 15 件には、7 月 13 日〜26 日までの関連文書しか含まれなかったのに対し、MMR の上位 15 件には、7 月 13 日〜31 日までの範囲の関連文書が含まれることがわかった。このことは、MMR には、誤った文書を上位に挙げてしまうという欠点がある一方で、期待されたように冗長性を排除する効果もあることを示していると考えられる。

今後、MMR を利用してトラッキング結果の要約を行う際には、何らかの方法で上位に現われる誤った文書を検出する方法が必要と考えられる。

6 まとめ

関連度の計算において固有表現に重みをつけることにより、トラッキング精度を向上することがわかった。しかし、どの種類の固有表現が精度向上に寄与するかは、イベントの性質にも依存するため、今回の実験だけでは確定的なことはわからない。今後、より多くのイベント関連記事のデータを集め検討する必要がある。また、元記事の時期によりトラッキング精度に差があるかを調べた結果、今回は特に系統的な差異は認められなかった。最後に、MMR を利用してトラッキング結果の要約を試みた。そこでは、冗長性を排除するという MMR の特性が現われてはいるものの、上位に誤った文書が多く含まれるようになるという問題点も明らかとなった。

謝辞

研究に関し御助言頂いた、NTTコミュニケーション科学基礎研究所、知能情報研究部の片桐滋部長、同じく知識処理研究グループの前田英作リーダに感謝致します。

参考文献

[1] Topic Detection and Tracking Workshop, http://www.itl.nist.gov/iaui/894.01/tests/tdt/index.htm

[2] TDT2000 Evaluation Plan, http://www.itl.nist.gov/iaui/894.01/tests/tdt/tdt2000/evalplan.htm より入手可能

[3] 新谷研、角田達彦、大谷巧、長尾真: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、Vol.38 No.4 pp.855–862 (1997)

[4] 大竹清敬、増山繁、山本和英:名詞の連接情報を用いた関連文書検索手法、情報処理学会論文誌、Vol.40 No.5 pp.2460–2467 (1999)

[5] 磯崎秀樹:固有表現抽出のための可読性の高い規則の自動生成、情報処理学会研究報告 2000-NL-140 pp.69-76 (2000)

[6] 黒橋禎夫、長尾真:日本語形態素解析システム JUMAN version 3.61 (1999)

[7] J. Carbonell, J. Goldstein:The Use of MMR, Diversity-Based Reranking for Reordering Dosuments and Producing SUmmaries In Proc. of SIGIR'98 pp.335–336 (1998)