2005－DBS－137（Ｉ）（24）
2005／7／13

固有表現を用いたニュース記事分類手法の提案

戸田浩之[†,††] 片岡良治[†] 北川博之[††,†††]

コンピュータおよびコンピュータネットワークの発展により，アクセス可能な情報の量が増大している．その中でもニュース記事は，最新ニュースの閲覧やアーカイブの検索等様々な方法にて利用され，最も利用されているコンテンツの一つである．なかでも，アーカイブの検索等における，ニュース記事の見方として同一のイベントについての推移を一覧したいという要求がある．つまり，同じイベントについて書かれている記事をグループ化する事が求められている．我々はこの問題に対して，ニュース記事中の固有表現に着目し，固有表現を用いた分類を行うことで特定のイベントに関連するニュース記事を集めることができるのではないかと考えた．本稿では，ニュース記事中での固有表現の分布の分析および固有表現を用いた分類実験を行い，固有表現を用いたイベント特化型の分類の可能性について報告する．

Clustering News Articles using Named Entities

Hiroyuki Toda,[†,††] Ryoji Kataoka[†] and Hiroyuki Kitagawa [††,†††]

Due to the growth of the Internet, the amount of information accessible to the public has almost exploded. Especially, news articles are intensively used for latest news watching, retrieving interesting information from news archives and so on. In news archive services, there is a demand to group news articles describing the same event. To address this problem, we use Named Entities in news articles to tell which events the articles describe. In this paper, we present the results of experiments to measure the appearance tendency of named entities in news articles and accuracy of clustering taking named entities into consideration, and discuss validity of the proposed approach.

1. はじめに

コンピュータおよびコンピュータネットワークの発展により，アクセス可能な情報の量は増大している．その中でもニュース記事は，最新ニュースの閲覧やアーカイブの検索等様々な方法で利用され，最も利用されているコンテンツの一つとなっている．

しかし，ニュース記事は，多くの情報源から，時々刻々と配信されており，最新ニュースだけに限っても，それらを網羅的にチェックすることは難しい．また，ニュースのアーカイブとなると，その量はさらに増大し，一般に利用される全文検索システム等で検索をしたとしても，しばしば膨大な数の検索結果に直面することになる．

これらの事から，インターネット等で配信されているニュース記事を利用する立場からの問題点として，以下のことが挙げられる．

- 多くの情報源が同一の内容に関するニュースを配信する為，同内容の文書が大量に散在してしまうこと
- 続報等文書の内容に明らかな関係性がありながら，形式上明確な繋がりが規定されていない

前者の問題は早くから取り組まれており，キーワードと時間の情報を元に文書間の類似度を測定することで高精度のニュース記事の統合が可能となっている．

一方，後者の問題は，部分的には前者の問題と同様な解決法が適用できる場合もある．しかし，同一のイベントに関する記事ではありながら，新たな真実の発見や，事件の進展により，必ずしもキーワード情報のつながりだけではその類似性が判断が厳密に出来なかったり，また，時間的にも前者の問題と比較すると考慮すべきタイムレンジが大きくなったりと，関連性のない記事を誤って統合する可能性が高まる．

† 日本電信電話株式会社 NTT サイバーソリューション研究所
NTT Cyber Solutions Laboratories, NTT Corporation
†† 筑波大学 システム情報工学研究科
Graduate School of Systems and Information Engineering, University of Tsukuba
††† 筑波大学 計算科学研究センター
Center for Computational Sciences, University of Tsukuba

一般のニュースサイトで最新ニュースを中心とした記事を提示する場合には，前者の対応のみで比較的質の高いサービスが提供出来る．google news☆ はその一例である．

しかし，記事アーカイブの検索等，比較的長期間に渡って蓄積された記事の検索や，特定の事件の経過を追いたいような場合には後者の問題を解決する必要がある．

以上を踏まえ，我々は，ニュース記事アーカイブの検索において，イベントに着目した検索結果の分類を行うことを目的とする．

我々はこの問題に対して，ニュース記事中には，実世界のインスタンスを指す固有表現が内容において重要な約割を担っていると考えられること，イベントが時間の経過によって進展したり，推移したとしても，代表的な固有表現は置き換わらず，イベントに対応する記事の中で一貫して出現するであろうと考えられることから，固有表現の情報を用いたニュース記事の分類について検討し，その結果について報告する．

以下，2 章では関連研究について示し，3 章では固有表現及び新聞記事中の固有表現に関する分析結果について示す．4 章で固有表現を利用した文書分類の実験および結果について示し，5 章でまとめる．

2. 関 連 研 究

ニュース記事の検索において，固有表現に着目した研究として，筆者らの [1] が挙げられる．ここでは固有表現はニュース記事を特定する情報として重要であると言う点に着目し，検索結果中の特徴的な固有表現を検索結果とともに提示する技術を提案している．これにより検索システムの利用者は，提示されている人物や組織などの固有表現を元に，検索結果を概観でき，また所望の固有表現を指定することでその固有表現を含む記事に容易にアクセスできる．

このようなキーワード (上記では固有表現) を基準に検索結果を分類する手法は様々提案されているが [2][3]，基本的には個々のキーワードを独立に評価しており，提示されたキーワードリスト中に同様なイベントや内容を示すキーワードが重複して出現すると言う問題がある．

典型的な例を図 1 に示す．これは，「殺人」というキーワードでニュース記事を検索し，検索結果に対して，文献 [1] に基づくシステムで，検索結果を分類するキーワードのリストを作成したものである．検索結果の記事中に登場する人物や組織，場所などが提示されている．しかし，このうち，四角で囲った 4 つのキーワード「新田美代子」「新田哲也」「大阪府堺市」「泉北署」は，同一のイベント (事件) に関連するキーワードであるが，まったく関係ないラベルと同様に独立して出現している．これは，個々のキーワードが独立に評価されるためである．理想的な形としてはこれらを統合して提示することであるが，ラベルを最初に特定する手法の場合は必ずしもうまく統合できない．

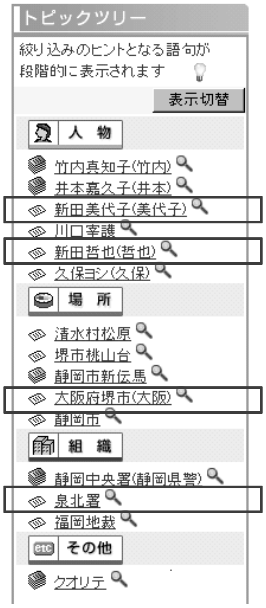

図 1 ラベル例
Fig. 1 Example of labels

一方，TDT(Topic Detection and Tracking) の分野では，ニュースイベントの新しさを評価する為に固有表現の利用が提案されている．基本的な TDT システムは，新しいニュースが到着すると，過去に蓄積された記事との類似性を判定し，予め設定された閾値を越えなければ，新しいニュースと判定するという仕組である．しかし，同じ分野の事件，例えばハイジャックに関する記事同士の場合，同じ様なキーワードを含む事が考えられ，別のハイジャック事件でも同じイベントとして認識され易いと言う問題がある．そこで，固有表現を利用することで，その違いを判定しようとしている．

Yang[12] らは，固有表現を利用することで，同じトピックに存在する，異なるイベントの識別性があがる事を示している．また，Kumaran[5] らは，LDC(Linguistic Data Consortium) のテストコレクションに付与されたカテゴリ毎に固有表現を用いたイ

☆ http://news.google.com/

表 1 2 × 2 分割表
Table 1 A two-by-two contingency table

	# of doc's on search topic T	# of doc's off search topic T
# of doc's containing w	A	B
# of doc's not containing w	C	D

表 2 χ^2 評価の結果
Table 2 result of χ^2 evaluation

トピック	人名	組織名	地名	固有物名	平均
関東の強盗事件の容疑者逮捕	7.64	2.81	1.96	1.95	4.08
東ティモール問題	3.80	5.33	12.8	0.03	4.94
携帯電話，簡易型携帯電話のサービス	1.53	6.68	0.20	2.69	3.99
便秘の原因と対策	3.63	1.04	0.71	3.51	1.28
外国人の参政権	5.91	1.26	0.84	2.95	1.52
ハイビジョンテレビ	3.02	2.21	0.24	5.00	1.20

ベントの識別が有益か否かの判定をしている．

3. ニュース記事中の固有表現

3.1 固 有 表 現

固有表現とは，もともと 1990 年代に行われた情報抽出の評価型ワークショップである MUC(Message Understanding Conference)[6] で生まれた概念である [7]．その定義としては,「新聞記事などの非構造のテキストから情報を抽出する為に，頻繁に重要になり，情報としての単位が明確な表現」と言うことになる．MUC において抽出の対象となったのは，人名，組織名，地名という固有名詞と，時間，日時，金額表現，割合表現の数値表現の 7 種である．日本では，1990 年代後半に行われた IREX(Information Retrieval and Extraction Exercise)[8] で「固有物名」という固有名詞のカテゴリが新しく作られ合計 8 種類を固有表現として扱い抽出する技術が研究されている．これらの固有表現抽出技術の精度は，新聞記事を対象とした場合には 90%を越える高精度で行うことが可能である．

3.2 新聞記事中の固有表現

ニュース記事中で固有表現が重要であると言うことは，直感的には理解できるが，必ずしも全てのトピックにおいて当てはまるとは限らない．そこで，本節では，IREX で利用された検索トピックとそれに対す正解文書の集合を用い，それぞれのタスクにおける固有表現の重要性について調査する．

IREX では，94，95 年の毎日新聞記事をコーパスとし，30 の検索トピックと，その検索トピックに対する正解を定義している．今回は，この 30 の検索トピックの正解文書集合それぞれの中での固有表現の重要性を評価した．

これらの新聞記事から形態素解析器 “茶筅” ☆を利用して，数を表す名詞と，非自立の名詞を除く名詞と未知語を記事中での語彙として取得した，また，同時に磯崎の手法 [9] による固有表現ツールを用いて固有表現抽出を行い，人名，組織名，地名，固有物名をそれぞれ取得した．☆☆

重要性の指標として，分布の偏りの程度を示す指標である χ^2 値を用いる．これは，分類を行う為の feature selection[4] や重要語の抽出 [10] 等で利用されている指標である．

各固有表現に対する χ^2 値は以下の式で算出される．

$$\chi^2(w,T) = \frac{(A+B+C+D)\times(AD-BC)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}$$

ここで，

- w は評価対象の固有表現，T は評価対象のトピックを示す．
- A, B, C, D は表 1 のそれぞれの場合の数を示す．

また，ここでは各トピックに対して，固有表現がどの程度偏って出現しているかを見るため，各固有表現の種類 (人名，組織名，地名，固有物名) 毎に平均 χ^2 値をとり，通常の形態素の平均 χ^2 値で正規化した値を表 2 に示す．この値は，以下の式で算出する．

$$\chi^2_{avg}(c,T) = \frac{\frac{1}{|N_c^T|}\times\sum_{w\in N_c^T}\chi^2(w)}{\frac{1}{|N_{morph}^T|}\times\sum_{w\in N_{morph}^T}\chi^2(w)}$$

ここで，

- c は固有表現の種類，N_c^T はトピック T の文書集合中で出現する，種類 c の固有表現の集合である．

☆ http://chasen.naist.jp/hiki/ChaSen/
☆☆ 今回は記事の同一性判定の為，固有表現を利用する事を考えているため，同一の意味を示す表現でも様々な表記が存在と考えられる数値表現は利用しなかった．

- N^T_{morph} はトピック T の文書集合中で出現する，形態素の集合である．

結果を表 2 に示す．ここでは，全種類の固有表現の平均で χ^2 値が高かった例を上に 3 つ，低かった例を下に 3 つ示している．

まず，個々の固有表現の種類毎に見ていくと，部分的に χ^2 値が高い部分が見られる．「関東の強盗事件」に関する話題では人名の値が高く，「東ティモール問題」では地名や組織名の値が高い．また，「便秘の原因と対策」では，人名☆の値が高い．これは，その分野で重要であると思われる固有表現の種類と一致している．

また，全種類の固有表現の平均値に着目すると，イベント性の強い話題では，χ^2 値が高い傾向にあり，逆にイベントよりも，一般的な知識を求めるような話題の場合には，あまり χ^2 値が高くない傾向にある．

以上より，定性的ではあるが，χ^2 値が高い固有表現の種類が話題の中で重要と考えられる種類とある程度一致していること，および固有表現の重要度とイベント性の強さにはある程度の相関があることが確認できた．

4. 提案手法と検証実験

我々は，固有表現がニュース中で重要なであること，またイベントとの関連性が強い事を元に，固有表現を用いて検索結果を分類する事でイベントと一致するような文書の分類ができるのではないかと考えた．

また，それに加えて，一般に文書の分類を行う場合に問題となる閾値の問題についても，固有表現に着目することで，なんらかの基準が得られるのではないかと考えた．つまり，イベントと固有表現との関連性が強ければ，閾値の設定が容易になるのではないかと考えた．

本章では，以上の考えを元に，固有表現を利用した分類がイベントを特化するような分類となっているか，また文書間の固有表現の分布の類似性を用いてイベントを特化する分類の閾値を決定することが可能であるかについて実験した結果について示す．

4.1 実験条件

以下に示す実験では，イベントに特化した文書を対象とするために，「汚職 or 贈賄 or 収賄」と「殺人」という検索キーワードを用いて，94，95 年の毎日新聞記事を検索した検索結果のそれぞれ上位 200 件を利用している．この検索には検索システム LISTA[11] を用いた．前者の文書セットの名称を [汚職]，後者を [殺人] とする．

表 3 正解セット
Table 3 Relevance Jadgment

	[汚職]	[殺人]
文書数	164	75
イベント数	42	17
平均形態素数	19.94	19.33
平均固有表現数	4.46	3.38

また，分類の精度を測定するために，それぞれ検索結果について，人手でイベント付けを行い，二つ以上の文書で言及されているイベントおよびそのイベントについて記述した文書を正解セットとして抽出した．この作業は，本稿の第一著者が，それぞれの検索結果の文書を読んでイベント名を付与し，全てが終了した時点で，イベント名の情報を元に，同一のイベントに関する文書を集めたものである．正解セットの詳細に付いては，表 3 に示す．

分類を行う為に，個々の文書毎に，以下に示すように形態素および固有表現からなる文書ベクトルを作成した．文書の解析法は 3 章に示した手法と同様である．

$$\vec{d} = (w_{1,d}, w_{2,d}, ..., w_{p,d}, x_{1,d}, x_{2,d}, ..., x_{q,d})$$

$$w_{i,d} = (1-\alpha)\times(log(1+n(t_i,d)))\times\log\left(\frac{|D|}{n(t_i,D)}\right)$$

$$x_{j,d} = \alpha \times (log(1+n(s_j,d))) \times \log\left(\frac{|D|}{n(s_j,D)}\right)$$

ここで，

- D はコレクション中の全文書集合
- $w_{i,d}$ は文書 d 中での形態素 t_i の重み，$x_{i,d}$ は文書 d 中での固有表現 s_i の重み
- p は一つのベクトルで利用した形態素の数，q は一つのベクトルで利用した固有表現の数
- $n(t_i,d)$ は文書 d での形態素 t_i の出現頻度，$n(s_i,d)$ は文書 d での固有表現 s_i の出現頻度
- $n(t_i,D)$ はコレクション D 中での形態素 t_i の出現頻度，$n(s_i,D)$ はコレクション D 中での固有表現 s_i の出現頻度
- α は，固有表現と形態素の重みを捜査する為のパラメータ

である．

ここで，p と q は，正解セット中の平均形態素数および平均固有名詞数より，$p = 20$ と $q = 5$ とし，tf-idf 値の上位から規定分をベクトルの feature とした．また，ベクトル間類似度には cosine similarity を用いた．

☆ ここで，χ^2 値の高い人物としては，健康に関するコラムの筆者等が並んでいる

また，以下の実験では，凝集法を用いて，分類を行った．クラスタ間の類似度にはクラスタのセントロイド間の類似度を利用するセントロイド法を利用している．分類精度の評価には表 3 に示した正解セットを利用し，FScore measure[13] を用いて評価した．FScore の算出式は以下に示す通りである．

$$FScore = \sum_{c \in C} \frac{|m_c|}{|m|} \max_{r \in R} \frac{2 \times |m_{r,c}|}{|m_r| + |m_c|}$$

ここで，

- C は正解セット中でのカテゴリ集合，c は正解セット中での 1 つのカテゴリ，m_c はカテゴリ c 中の文書の集合
- R は分類結果中でのカテゴリ集合，r は分類結果中での 1 つのカテゴリ，m_r は，カテゴリ r 中の文書の集合
- $m_{r,c}$ は $m_r \cap m_c$
- m は正解セット中に含まれる文書集合

を示す．

4.2 分類の精度について

まず，理想状態での分類手法の精度を評価するために，正解セットのみを用いて分類精度を評価した．最終的な分類数はそれぞれの正解セットのイベント数に合わせた．結果を図 2，3 に分類結果を示す．縦軸は分類精度を示す FScore，横軸はベクトル中での固有表現の重みを表す指標 α である．

まず，それぞれのグラフ中の黒い四角のプロットに注目する．[汚職] の結果では，形態素と固有表現の重み付けがほぼ等しくなる $\alpha = 0.5$ 以上で，0.85 程度の高い FScore を記録している．逆に形態素のみの場合は，0.25 程度とかなり低い．また，[殺人] の結果を見ると，[汚職] 程顕著ではないものの，固有表現の重みを増やすに連れて，分類の FScore が向上していることがわかる．

また，この結果を見る限りは，グラフの右端がほぼ FScore の最大値に近い値を持っている．これは，イベント応じたニュース記事の分類を行う為には，固有表現のみでかなり高い確度で分類が可能であると言うことを示している．

また，それぞれのグラフ中の白い四角のプロットは，3 章で述べた χ^2 値を用いて，各固有表現の種類毎に重み付けを変更した場合の結果である．つまり，文書コレクション中で重要と思われる固有表現の種類を優先して重み付けをしている．しかし，今回の実験ではあまり優位な結果は得られなかった．

この要因としては，今回の実験では，文書ベクトル中の feature をあらかじめ tf-idf に基づいて選択した後に重み付けを行ったために，実際のベクトルではあまり重みの値が効いていなかった為であると考えられる．今度，feature の選択時に χ^2 値を利用することも考えられる．

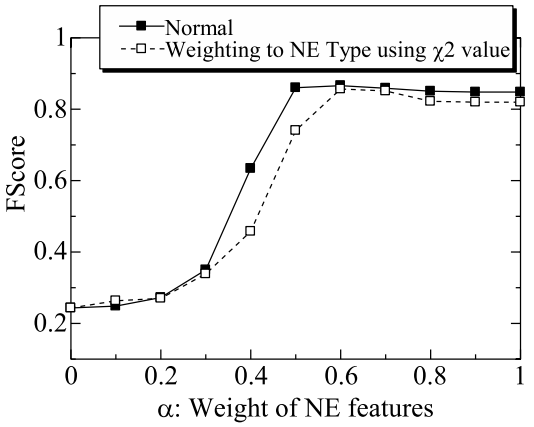

図 2 FScore とαの関係 (検索キーワード: 汚職 or 収賄 or 贈賄)
Fig. 2 Relationship between α and FScore(search keyword: 汚職 or 収賄 or 贈賄)

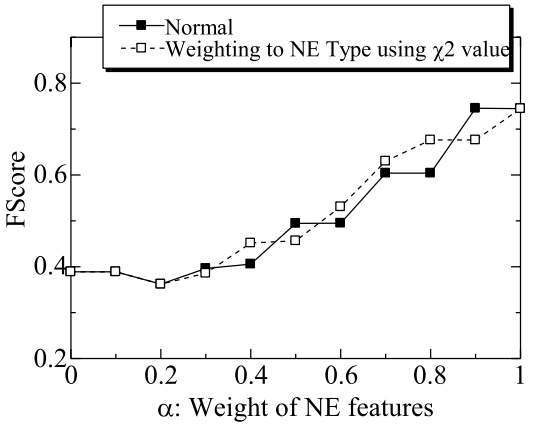

図 3 FScore とαの関係 (検索キーワード: 殺人)
Fig. 3 Relationship between α and FScore(search keyword: 殺人)

4.3 閾値設定について

次に，正解セットを含むそれぞれの検索結果集合 200 件づつを分類する実験を行った．結果を図 4, 5 に示す．縦軸に FScore および類似度，横軸はクラスタ数である．今回の実験では凝集法を用いた為，クラスタ 200 から 1 ずつ減る度にそのクラスタの FScore と，凝集された際の類似度をプロットしている．つまり，時間順で見ると，図 4, 5 は，左から右へと見る形となる．

まず [汚職] の FScore に着目すると，固有表現のみを用いた際の精度が形態素のみを用いた際の精度を大

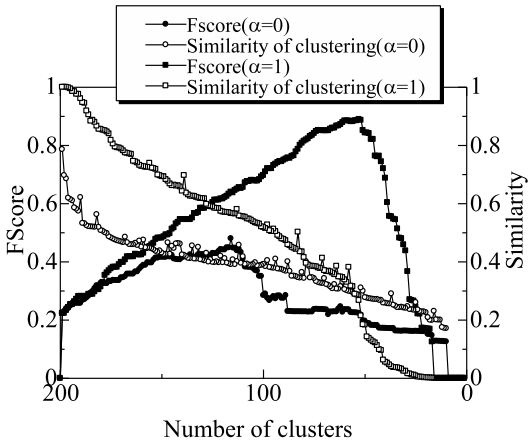

図 4 クラスタ数と FScore の関係 (検索キーワード: 汚職 or 収賄 or 贈賄)
Fig. 4 Relationship between α and FScore(search keyword: 汚職 or 収賄 or 贈賄)

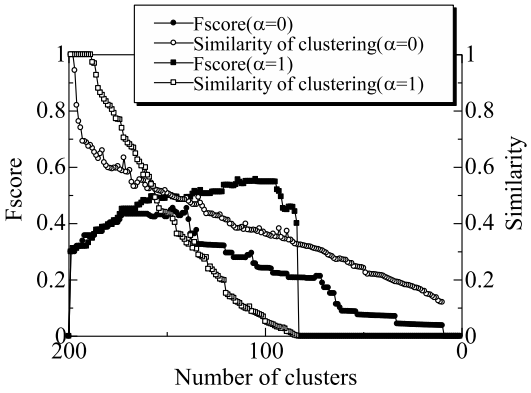

図 5 クラスタ数と FScore の関係 (検索キーワード: 殺人)
Fig. 5 Relationship between α and FScore(search keyword: 殺人)

きく上回っている事がわかる．また，分類時の精度についても，FScore が上昇している段階では，固有表現のみを用いた際の類似度が高くなっている．しかし，固有表現を利用した場合の類似度は，FScore が急激に低くなる部分で，類似度も同様に低下していることがわかる．これは，本章の最初に述べたように，イベントと固有表現の関連性が強いために，FScore を下げるようなクラスタの凝集，つまり，異なるイベント同士のクラスタを凝集が起こる部分で，類似度も同様に低下したのではないかと考えられる．

しかし，今回評価したもう一つの例である [殺人] では，この傾向を顕著に確認する事ができなかった．一つの原因としては，固有表現の feature が [汚職] の時ほど，有効でなかったということが考えられる．しかし，これがそもそも固有表現が有益でないのか，単に feature の選択法が間違っていたのかは，現段階では議論できない．

5. ま と め

以上，本稿では，ニュース記事のアーカイブ検索の際に，検索結果をイベントに基づく形で分類する手法について提案を行い，評価を行った．

まず，ニュース記事中で χ^2 値を用いて固有表現の種類を評価したところ，χ^2 値と重要度に相関があることが定性的に評価できた．また，全体的に固有表現に高い χ^2 値が得られる場合にはイベントが存在する話題であると言うことが定性的に評価できた．今後は，χ^2 値を用いた feature の選択等も検討する予定である．

また，分類結果の評価から，固有表現を用いた分類は，従来の形態素のみを用いた分類と比較して高精度の分類が可能であるとの知見を得た．さらに，固有表現を用いた分類のプロセスを観察することで，固有表現の分布に基づいて，分類終了条件の判定ができる可能性を見出した．今後は，TDT のテストコレクションなどを用いて，網羅的な評価と手法の改善を行う予定である．

参 考 文 献

1) Toda, H. and Kataoka, R.: “A Clustering Method for News Articles Retrieval System.” *Poster Proceedings of WWW 2005, pp988–989* (2005).
2) Zeng, H. J., He, Q. C., Chen, Z., Ma, W. Y. and Ma, J.: “Learning to Cluster Web Search Results.” *Proceedings of SIGIR 2004, pp.210–217* (2004).
3) Ferragina, P. and Gulli, A.: “The Anatomy of a Hierarchical Clustering Engine for Web-page, News and Book Snippets” *Proceedings of ICDM 2004,* (2004).
4) Yang, Y., Zhang, J., Carbonell, J. and Jin, C.: “Topic-conditioned Novelty Detection” *Proceedings of SIGKDD 2002,* (2002).
5) Kumaran, G., and Allan, J.: “Text Classification and Named Entities for New Event Detection” *Proceedings of SIGIR 2004,* (2004).
6) Grishman, R. and Sundheim B.: “Message Understanding Conference - 6: A Brief History.” *Proceedings of COLING 1996, pp.466–471* (1996).
7) 関根聡: “固有表現から専門用語”, **言語処理学会**

第 *10* 回年次大会 *(NLP2004)* **「固有表現と専門用語」ワークショップ** (2004).
8) 関根聡, 井佐原均: “IREX プロジェクト概要” *IREX* **ワークショップ予稿集**, *pp.1–5* (1999).
9) Isozaki, H. and Kazawa, H.: “Efficient Support Vector Classifiers for Named Entity Recognition.” *Proceedings of COLING 2002, pp390–396* (2002).
10) 松尾豊, 石塚満: “語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム” **人工知能学会論文誌**, *Vol. 17, No. 3, pp217–223* (2002).
11) Hayashi, Y., Tomita, J. and Kikui, G.: “Searching text-rich XML documents” *ACM SIGIR 2000 Workshop on XML and Information Retrieval, pp.27–35* (2000).
12) Yang, Y., Pierce and T., Carbonell, J.: “A study on Restrospective and On-Line Event Detection” *Proceedings of SIGIR 1998, pp28–36* (1998).
13) Zhao, Y. and Karypis, G.: “Evaluation of Hierarchical Clustering Algorithms for Document Datasets” *Proceedings of CIKM 2002, pp515–524* (2002).