新旧地名・施設名対の抽出による文書の時空間マッピング

平山 拓実† 難波 英嗣† 竹澤 寿幸†

†広島市立大学大学院 情報科学研究科 〒731-3194 広島県広島市安佐南区大塚東 3-4-1

E-mail: †{hirayama, nanba, takezawa}@ls.info.hiroshima-cu.ac.jp

あらまし 我々は，様々な文書を地図上にマッピングするシステムを構築している．このシステムを用いることで，任意の場所に関する情報を容易に把握することができる．ここで，地名や施設名は年月とともに変わる可能性があるため，このシステムでは，古い地名表現を含んだ文書を地図上にマッピングできないという問題があった．そこで本研究では，新旧地名・組織名の対をテキストデータベースから抽出し，古い地名表現を含んだ文書のマッピングを実現する．提案手法を使って文書を時空間領域にマッピングすることにより，ある地点の歴史をさかのぼって調べたり，過去のある時点を言及した文書間の関係を調べたりすることが可能になった．

キーワード マッピング，可視化，情報抽出，地理情報，新旧地名

1. はじめに

本研究では，地名や施設名などの地名表現を含んだ文書を時空間領域にマッピングするシステムを構築する．本システムにより，大量の文書を読むことなく，任意の場所に関する情報を，時間をさかのぼって把握することが可能である，地域の歴史やある時間に何があったのかを容易に把握することができる．

一般的にマッピングシステムを実現するには，文書中の地名表現を抽出し，緯度経度データベースとのマッチングを行うことで緯度経度を付与(ジオコーディング)する作業が必要である．しかし，地名表現は年月によって移り変わっていくものである．例えば,「原爆ドーム」の場合,「広島県物産陳列館(1915 年-1921 年)」,「広島県立商品陳列所(1921 年-1933 年)」,「広島県産業奨励館(1933 年-1944 年)」と名称が年月により変化している．しかし，緯度経度データベースには，新しい地名表現しか登録されていないため，新地名表現と旧地名表現を対応させた新旧地名表現対応辞書が必要である．

我々の先行研究[1]では，新旧地名表現対応辞書を構築するために，Web ページから「[新地名表現]（旧[旧地名表現]」という限られたパターンを利用し，新旧地名表現対抽出を行った．しかし，新旧地名表現対を記述するパターンは，その他にも大量に存在する．そこで本研究では，上位下位関係の対を抽出する半教師あり学習アルゴリズムブートストラップ法[2]を利用することで，新旧地名表現対抽出を行うためのパターンを大量に収集する手法を提案する．この手法により収集されたパターンを利用することで，新旧地名表現対を網羅的に収集することが可能である．

さらに，収集した新旧地名表現対を利用することで，旧地名表現を含む文書のマッピング結果を閲覧するシステムを構築する．本システムは任意の場所における歴史をさかのぼって調べることを目的とするため，マッピング結果を年代で分割して表示する．これにより，ユーザが調べたい年代の情報を取得しやすくなる．また，同じ箇所に大量のピンが乱立する問題を緩和することができる．本研究は，文書が言及する年代を文書中の文から推定し，それを基に時空間上にマッピングすることを最終目標とする．

本論文の構成は以下の通りである．2 章では本システムの動作例を示し，3 章では関連研究について述べ，4 章では大量のテキストデータベースからの新旧地名表現対抽出について述べる．5 章では文書のマッピングについて詳しく述べる．6 章では評価実験について述べ，7 章で本論文をまとめる．

2. システム概要と動作例

本節では，構築するシステムの概要と動作例について説明する．本システムのマッピング対象文書は Wikipedia や旅行ブログ，Web ページなど様々な文書を対象としている．様々な文書をマッピングすることで，任意の場所における多種多様な情報を把握できるシステムとなっている．ただし，文書中の代表的な地名表現の推定はまだ行っていないため，現状では文書中の地名表現の全てに対してピンを立てている．つまり,一つの文書に対して複数のピンが立つことがある．

本システムは文書を年代ごとに表示する．例えば，図 1 の場合は 1900 年から 1930 年，図 2 は 1930 年から 1960 年の期間の「原爆ドーム」の結果を示している．このように表示することで，ユーザが調べたい年代の文書を探すことが容易になる．また，年代で提示することで同時期にどういったことが起きたかを調べる際にも役立つ表示となっている．さらに，分割する年代を自動決定できれば，一箇所に大量のピンが乱立する問題を緩和することも可能である．

図 1. 1900 年から 1930 年までのマッピング結果

図 2. 1930 年から 1960 年までのマッピング結果

次に，動作例について詳しく述べる．図 1，2 中のピンをクリックすることで，文書のタイトル，年代，地名表現を含む文が表示される．また，タイトルはクリックすることで，対象 Web ページへ移動することもでき，ユーザが興味のある文書を詳しく読むこともできる．年代は文の先頭に記述される．図 1 の場合，「[1914 年]」である．ただし，年代は文書の作成した年ではなく，文書で言及された出来事の年を意味する．図 1 においてマッピングされた文書には，1914 年に広島県物産陳列館が建築されたこと，1915 年に開館されたことに関する文書が存在した．それに対して，図 2 では，1933 年に広島県産業奨励館に改称したこと，1945 年に原爆が落ちたことに関する文書が存在した．このように，その年代において重要な出来事を視覚的に調べることが可能である．

3. 関連研究

本研究では，新聞記事や旅行ブログ，書籍といった様々な文書を地図上にマッピングしている．本研究と同様にマッピングシステムを構築する研究に郡ら[3]や鎌田ら[4]の研究が挙げられる．郡らは，複数の旅行ブログから代表的な行動経路とその行動のテーマを抽出し，地図上にマッピングをしている．鎌田らは，Twitter などのつぶやきからユーザの経路を抽出し，地図上に経路と投稿された写真を表示するアプリケーションの構築をしている．このように文書をマッピングする研究は多く，マッピングの対象とされる文書は多様である．本研究ではこれらの研究と異なり，Wikipedia や旅行ブログといった複数の種類の文書を同一の地図上にマッピングする．これにより，任意の場所における歴史や事件・事故などの様々な情報を取得できるシステムの構築が望める．

文書を地図上にマッピングするには，文書中から抽出した地名表現のジオコーディングが必要であり，旧地名表現の考慮が必要とされる．国分ら[5]は人手による自然言語処理用のシソーラスを構築する際，旧地名表現を出力しないために，旧地名表現の差別化を行った．しかし，シソーラスを運用するためには，増加し続ける旧地名表現を常に登録することが必要であると述べた．これに対し，本研究では莫大な時間とコストを必要としない手法を提案する．本手法は半教師あり学習アルゴリズムであるブートストラップ法[2]を用いることで，自動で新旧地名表現対抽出を行うことができる．

ブートストラップ法を用いた辞書構築は多く存在する．例えば，水口ら[6]は，Web ページからブートストラップ法を用いて，地名辞書や企業名辞書の構築を行っている．また，インスタンスを含む Web ページの検索に複数単語で組み合わせたクエリを用いることで実行時間の短縮を提案した．本研究は，地名表現の分類ではなく，新旧対応付けを行っている点で異なる．

本研究と同様に時間を考慮した情報抽出の研究がある．Ling ら[7]は，確率モデルを用いることで，イベントと年代の時間関係を抽出した．例えば，“Steve Jobs revealed the iPhone in 2007”といった文から Jobs が iPhone を発表したイベントは，2007 年開始時から終了時までの間に存在するという時間関係を抽出した．また，高久ら[8]は，単語の時系列頻度を用いた，教師あり学習により各米国大統領などの時間関係を抽出した．本研究では，より網羅的な抽出を行うため，頻度に大きく依存しない手法を用いる．

次に，本研究と同様に，任意の場所における過去のイベントをマッピングした研究について述べる．Jannik ら[9]は，電子化した書籍データから抽出した時間情報と地名表現を組み合わせてイベントを生成し，地図上に表示する手法を提案している．図 3 に示す動作例は重要度の高いイベントをマッピングしている．また，各イベントは時系列において前後のイベントと線で結ばれている．本研究では，重要な文書のみではなく，様々な文書を網羅的にマッピングするシステムを目指している．そのため，一箇所に複数の文書がマ

ッピングされ，Jannik らの提示方法では効果が薄いと考えられる．そこで，本システムではマッピング結果を年代で分割して表示する．

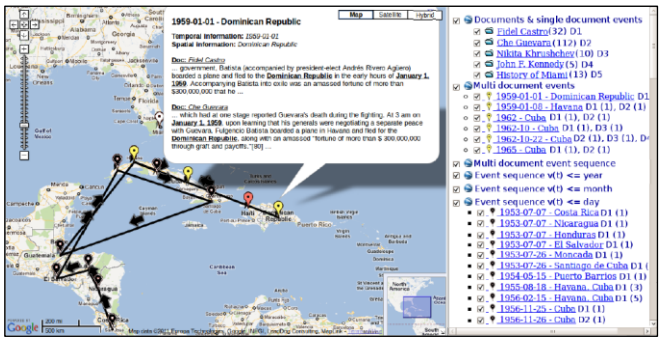

図 3. Jannik ら[9]のシステム動作例

4. 新旧地名表現対の抽出

4.1. 新旧地名表現対抽出手法の概要

新旧地名表現対の情報は，テキスト中で[新地名表現]<パターン>[旧地名表現]のように記述される．例えば，以下の例では，ポンペイ島はかつてポナペと呼ばれていたことがわかる．

> *毎年恒例ミクロネシアツアー。今年は太平洋の孤島、ポンペイ島（旧ポナペ）です。*

そこで，実際に新旧地名表現対を含む Web ページや新聞記事内の文を調べたところ「(旧」や「(当時は」など様々なパターンが存在した．先行研究[1]では「(旧」を手がかりにし，機械学習手法 CRF を用いることで新旧地名表現対を高い精度で抽出した．しかし，対象テキストデータベースを「(旧」を含む文に限定したため，新旧地名表現対の抽出件数が少ないことが考えられる．そのため，本研究では，Espresso アルゴリズムによるブートストラップ法を用いることで，より網羅的な新旧地名表現対抽出を行う．

4.2. ブートストラップ法

本研究で用いるブートストラップ法とは，シードインスタンスを基にし，新たなパターンやインスタンスを抽出する手法である．ブートストラップ法を図 4 を用いて説明する．例えば，シードインスタンスに新地名表現「ポンペイ島」と旧地名表現「ポナペ」を与えた場合，テキストデータベースから「(旧」などのパターン集合を抽出する．次に，抽出したパターン集合を用いて，新地名表現「さいたま市」と旧地名表現「浦和市」といった新たな新旧地名表現対を抽出する．このように，パターンとインスタンスを繰り返し抽出することで，網羅的な新旧地名表現対抽出が可能となる．

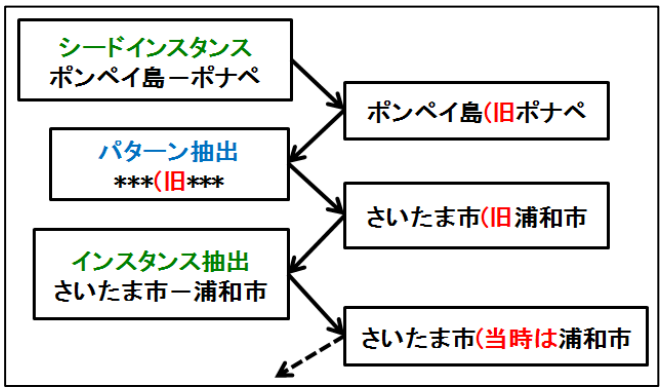

図 4. ブートストラップ法の概要

ブートストラップ法には，繰り返し抽出する間に誤ったパターンまたはインスタンスの抽出が行われた場合，精度が低くなる問題がある．しかし，この問題は Pantel ら[10]が提案した Espresso アルゴリズムを用いることで緩和できる．Espresso アルゴリズムでは，インスタンスとパターンの信頼スコアを相互再帰的に定義する．この信頼スコアを用いることで，信頼スコアの高いパターンと共起するインスタンスは信頼スコアが高く，信頼スコアの高いインスタンスと共起するパターンは信頼スコアが高くなる．これにより，誤った抽出を減少させ，信頼スコアの高いパターンとインスタンスを抽出することができる．パターンpとインスタンスiの信頼スコアはそれぞれ$r_\pi(p)$と$r_i(i)$で表し，以下の式を用いる．

$$r_\pi(p) = \frac{1}{|I|}\sum_{i \in I}\frac{pmi(i,p)}{\max pmi} r_i(i) \quad (1)$$

$$r_i(i) = \frac{1}{|P|}\sum_{p \in P}\frac{pmi(i,p)}{\max pmi} r_\pi(p) \quad (2)$$

P と I はパターンとインスタンスの集合を表し，$pmi(i,p)$は i と p の自己相互情報量を表している．$pmi(i,p)$は以下の式で求められる．

$$pmi(i,p) = \log_2 \frac{|i,p|}{|i,*||*,p|} \quad (3)$$

4.3. ブートストラップ法を利用した新旧地名表現対抽出手法

本研究では，図 4 のようにシードインスタンスに新旧地名表現対を利用することで，テキストデータベースから網羅的に新旧地名表現対抽出を行う．先行研究と異なり，パターンを「(旧」に限定しないため，より多くの新旧地名表現対を抽出できると考えられる．

本研究におけるパターンは，10 文字以内の文字列とする．また，抽出したパターン集合の内，「は,」のようにあまりにも頻度が多いパターンは実行時間が非常

に掛かる上，スコアの影響も低いので取り除くものとする．そして，次のインスタンス抽出に用いるパターン集合は信頼スコア上位 n 件のパターンを用いる．

本研究におけるインスタンスは，パターンの直前・直後の地名・施設名に関する固有表現とする．固有表現解析には日本語係り受け解析器 CaboCha の固有表現解析機能を用いる．これは IREX-NE で公開された定義[1]に基づいた固有表現に分類を行う機能である．本研究では，この機能を用いて，LOCATION または ORGANIZATION とされた語をインスタンスとして抽出する．以下に LOCATION と ORGANIZATION の定義を述べる．まず，LOCATION は，大陸や地域名，駅名，山といった固有の場所を指す名前である．次に，ORGANIZATION は，株式会社や学校，病院といったなんらかの目的を持った組織などの名前である．よって本研究では，LOCATION が地名であり，ORGANIZATION が施設名と対応すると考え，この二種類の固有表現を用いる．そして，次のパターン抽出に用いるインスタンス集合は信頼スコア上位 m 件とする．ただし 6 章の実験では，パターン抽出，インスタンス抽出の反復回数を 1 回としたため，インスタンスからのパターン抽出は行わない．そのため，m の値は定義しないものとする．パターン抽出とインスタンス抽出の繰り返しにより，抽出したインスタンスを用いて新旧地名表現対応辞書を作成する．

5. 文書マッピングシステム

本節では，本システムについて説明する．本システムの流れについて図 5 に示す．まず，文書に含まれる全ての地名・施設名に関する固有表現を地名表現として抽出する．次に，新旧地名表現対応辞書を参照して，抽出した地名表現を新地名表現に置き換える．そして，緯度経度データベースを用いて地名表現のジオコーディングを行い，地図上にマッピングを行う．

まず，文書中の地名表現の抽出について述べる．本研究では，地名表現の抽出に 4.3 節で述べた CaboCha を用いる．使用する固有表現もインスタンスと同様の LOCATION と ORGANIZATION である．

抽出した地名表現のジオコーディングには，地名・施設名の緯度経度データベースとのマッチングを行う．ただし，地名表現が新旧地名表現対応辞書の旧地名表現と一致した場合は対応する新地名表現の緯度経度でジオコーディングを行う．そして，ジオコーディングによって付与された緯度経度を基に，Google Maps[2]を用いて文書のマッピングを行う．

[1] http://nlp.cs.nyu.edu/irex/
[2] http://maps.google.co.jp

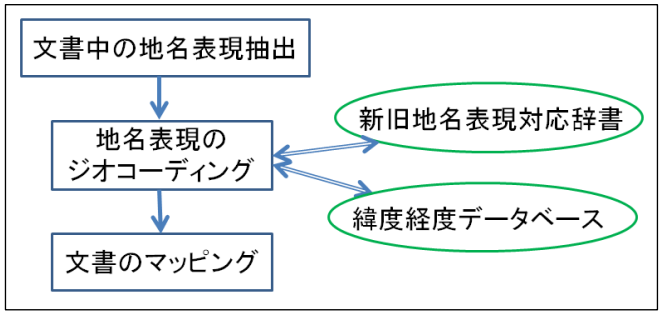

図 5. 文書マッピング概要

6. 新旧地名表現対抽出実験

本節では 4 章の新旧地名表現対抽出手法の評価実験について説明する．6.1 節では実験方法について述べ，6.2 節で実験結果，6.3 節で考察について述べる．

6.1. 実験方法

ブートストラップ法に用いるシードインスタンスには Wikipedia の「日本の廃止市町村一覧[3]」に記載された地名，JST 提供の企業名の新旧地名表現対，各 30 件を使用した．本研究では，施設名の収集が困難と考えたため，施設名の一部である企業名をシードインスタンスとした．そして，テキストデータベースには，NTCIR-5Web 検索タスク[4]に使用されたデータセットを用いた．先行研究の実験では，上記の 1.3TB あるテキストデータベースを用いた．しかし，ブートストラップ法による抽出は非常に時間が掛かるため，本実験では 88GB に減らしたデータをテキストデータベースとした．このテキストデータベースに対して，4 章のブートストラップ法を用い，パターンから抽出したインスタンスの信頼スコア上位 400 件を評価する．ただし，パターンとインスタンスの抽出反復回数は 1 回とする．また，インスタンス抽出に用いるパターン集合の件数は n=20 とする．評価尺度には，精度を用いる．

6.2. 実験結果

実験の結果，抽出したパターン数は 247 件，インスタンス数は 9,038 件であった．インスタンス信頼スコアの上位 400 件までの平均精度の推移を図 6 に示す．また，上位 400 件を 4 つの区間に分割し，1 区間ごとの平均精度を表 1 に示す．図 6 と表 1 を見ると，上位 250 件以降信頼スコアが下がるにつれ，精度も低下していることが分かる．また，上位 10 件の平均精度は 0.60 であった．これらのことから，信頼スコアが高ければ，精度も高いといった，ブートストラップ法を用いた上位下位関係抽出と同様の結果が得られた．そのため，新旧地名表現対抽出にブートストラップ法が有効であることが分かる．

[3] https://ja.wikipedia.org/wiki/日本の廃止市町村一覧
[4] http://www.lemurproject.org/clueweb09.php/

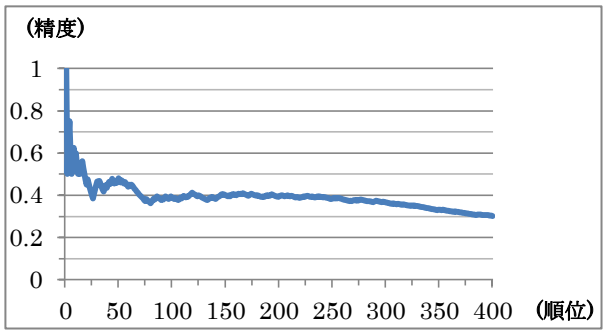

図 6. インスタンス信頼スコア上位 400 件の精度

表 1. 区間における精度

区間	精度
[1-100]	0.39
[101-200]	0.40
[201-300]	0.31
[300-400]	0.11

6.3. 考察

パターンの信頼スコア上位 10 件を表 2，正しく抽出されたインスタンスの例 10 件を表 3，誤って抽出されたインスタンスの例 5 件を表 4 に示す．

表 2. パターン信頼スコア上位 10 件

順位	パターン	順位	パターン
1	（旧	6	側と
2	の旧	7	三井
3	：旧	8	市場
4	(旧	9	（←
5	（旧	10	（当時

表 3. 正しく抽出されたインスタンス例

新地名表現	旧地名表現
ロシア	ソ連
ドイツ	西ドイツ
ジェネオン	パイオニア
JR	国鉄
ＪＲ	国鉄
魚沼市	堀之内町
マケドニア	ユーゴスラビア
日本学生支援機構	日本育英会
HP	コンパック
中国東北部	満州

表 4. 誤って抽出されたインスタンス例

新地名表現	旧地名表現
日本	日本軍
ITmedia	ZDNet
関東	東京
UFJ	三和
中央金庫	全国信用金庫連合会

抽出したパターンには「(旧」の他に，表 2 に示す「：旧」や「(←」，「(当時」，さらに，「、かつての」，「と合併する」といった様々な有効なパターンを抽出することができた．また，先行研究で抽出したインスタンス数 43,333 件と比べ，本実験では 1/10 以下のテキストデータベースから 9,038 件のインスタンスを抽出することができた．今後，反復回数を増やすことでより多くのインスタンスを抽出ができると考えられる．これらのことから，本研究の目的である，より多くの新旧地名表現対抽出に貢献できたと考えられる．

図 6 と表 1 を見ると，上位 1 件から 250 件まで精度がほぼ同値であることが分かる．これは，反復回数やインスタンス抽出に用いたパターン数(n=20)が問題であったと考えられる．そのため，反復回数を増やすことや，n の値を増やすことで，よりインスタンス間の信頼スコアの差を大きくする必要がある．

次に，抽出されたパターンを確認したところ，「が」や「と」など効果が期待できないパターンが存在した．しかし，テキストデータベースを確認したところ，インスタンスの直後に「を買収した」や「が合併した」など新旧地名表現対を決定付けるパターンを確認した．このことから，本手法の二項間パターン以外にもインスタンス直後のパターンを追加することで，より精度が向上すると考えられる．

表 3 を見ると，正しく抽出されたインスタンスの中に「JR－国鉄」と「ＪＲ―国鉄」がある．このように，共通の地名表現と対応付く地名表現を利用して，表記揺れを検出することも可能と考えられる．次に本手法で誤った抽出に表 4 を用いて考察する．まず，「ITmedia―ZDNet」は，ニュースサイト「ZDNet JAPAN」が「ITmedia」にリニューアルしたため，インスタンス信頼スコアが高くなったと考えられる．しかし，この対は地名表現ではないので誤りとした．「UFJ―三和」は「三和銀行」と「東海銀行」が合併し，「UFJ 銀行」に名称が変更した．しかし，テキストデータには「銀行」が省略して記述しており，緯度経度の特定が困難になる問題が起こる．例えば他に「UFJ グループ(UFJ)」やスーパーマーケットチェーン運営会社「三和」が一致する可能性がある．「中央金庫―全国信用金庫連合会」は本来「信金中央金庫―全国信用金庫連合会」であるが，固有表現解析において，「中央金庫」が ORGANIZATION となっていた．

「(旧」を含む文のみを抽出対象とした先行研究の精度 0.88 と比較すると，低下しているが，抽出の反復回数を増やすことや，機械学習を用いたインスタンス抽出により差を小さくできると考えられる．また，多くの有効なパターン抽出ができたことから，先行研究より多くの新旧地名表現対抽出が望める．

7. おわりに

本稿では，様々な文書を時空領域にマッピングするシステムを構築するため，新旧地名表現対を大量のテキストデータから抽出し，旧地名表現を含む文書をマッピングする手法を提案した．新旧地名表現対抽出にはブートストラップ法を用いることで実現した．実験結果より，上位 101 件目から上位 200 件目の区間における平均精度が 0.400 で他の区間より高い値であった．先行研究と比べて低い値であったが．有効なパターンの抽出ができたことから成果があったと考えられる．また，精度は機械学習を用いたインスタンス抽出を行うことで改善できると考えられる．今後の課題には，パターンの位置，インスタンス抽出における地名・施設名の正しい判定などが挙げられる．

謝辞

本研究のシードインスタンスの一部を提供してくださった国立研究開発法人科学技術振興機構(JST)に深く感謝致します．

参 考 文 献

[1] 平山拓実，難波英嗣，竹澤寿幸，“文書の時空間 3 次元地図へのマッピング”，電子情報通信学会技術研究報告 LOIS, Vol.115, No.110, pp.35-39, 2015.

[2] Yarowsky David, “Unsupervised Word Sense Disambiguation Rivaling Supervised Methods”, Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics (SCL’95), pp.189-196, 1995.

[3] 郡宏志，服部峻，手塚太郎，田島敬史，田中克己，“ブログからのビジターの代表的な行動経路とそのコンテキストの抽出”，情報処理学会研究報告データベース, Vol.2006, No.78, pp.35-42, 2006.

[4] 鎌田早織，坂本寛幸，井垣宏，中村匡秀，“マッシュアップ API を用いた異なるライフログサービスの連携”，電子情報通信学会技術研究報告 LOIS, Vol.109, No.450, pp.91-96, 2010.

[5] 国分芳宏，岡野弘行，“複数の観点で分類した自然言語処理用シソーラス”，自然言語処理，Vol.17, No.1, pp.247-263, 2010.

[6] 水口弘紀，河合英紀，土田正明，久寿居大，“Web 知識を利用したブートストラップによる辞書増殖手法”，電子情報通信学会，第 18 回データ工学ワークショップ論文集, E8-5, 2007.

[7] Xiao Ling and Daniel Weld, “Temporal Information Extraction”, Proceedings of the 24th AAAI, pp.1385-1390, 2010.

[8] 高久陽平，吉永直樹，鍛治伸裕，豊田正史，喜連川優，“時系列テキストを用いた恒久性と一意性に基づく関係の分類”，電子情報通信学会論文誌 D, Vol.96, No.3, pp.411-422, 2013.

[9] Jannik Strötgen and Michael Gertz, “Event-Centric Search and Exploration in Document Collections”, Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries, pp.223-232, 2012.

[10] Pantel Patrick and Pennacchiotti Marco, “Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations”, Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL, pp.113-120, 2006.