3E3-5

ウェブ文書の構造を利用した場所名・住所ペアの獲得

Acquisition of pairs of landmarks and their addresses using structure of Web pages

佐藤貴大[*1] 岡崎直観[*1*2] 乾健太郎[*1]
Takahiro Sato Naoaki Okazaki Kentaro Inui

[*1]東北大学大学院情報科学研究科
Graduate School of Information Sciences. Tohoku University

[*2]科学技術振興機構 さきがけ
Japan Science and Technology Agency (JST)

When we try to use a mobile terminal to search for nearby shops or to analyze one's behavior on Microblogs, it becomes necessary to recognize representation of the location such as shops, school buildings, intersections, and so on and to specify where it is shown. In this paper, we propose the method that automatically expand language resources about the location by extracting pairs of landmark and address from Website to use a small amount of seed data, that includes pairs of known landmark and address.

1. はじめに

近年爆発的な普及を見せた Twitter に代表されるマイクロブログにおいて、人々は日常のあらゆる行動を手軽に日記のようにして投稿することが可能となった。マイクロブログの書き手の多くは一般の個人であり、投稿された内容からは、書き手の思想や行動、人々のつながりを抽出出来る。また、リアルタイム性の高い情報発信がなされるという特徴もある。このため、マイクロブログからの情報の抽出、マイニング、書き手の属性推定、個人に対する行動分析などの重要性が高まっている。

マイクロブログを扱う上で重要な意味をもつ情報の一つに、位置情報がある。投稿者の居住地域を特定することで、地元の人間の意見の信頼度を高めることや、情報配信の地域限定化などが可能となる。その他にも、特定の地域でのトレンド分析や評判分析などにも有用である。また、投稿された場所（あるいは、文中で示唆している場所）を特定することで、投稿者の行動を分析する助けにもなる。マイクロブログの中にはジオタグが付与されていているものも存在するがその割合は全体に対してごく少数であるため、投稿内容に基づいて位置情報を推定するの研究が多くなされている。

文中に含まれる語句を用いて位置情報を扱う場合、そこに示されるお店や学校、交差点のような場所を表す表現を認識し、その場所（住所）を特定することは非常に有益である。例えば「佐藤食堂で昼食をとり、東北ホテルに向かった」という文に対し、「佐藤食堂」と「東北ホテル」を認識することで投稿者の行動の軌跡が把握できる。しかしながら場所を表す表現は多岐にわたり、その表現に規則性のようなものを発見するのは容易ではないため、単純に正規表現によりマッチングを行うといったことは難しい。このとき、あらかじめ場所の表現の候補を保持していればマッチングを行うことが可能となる。

ウェブ上には人により場所名と住所の情報がまとめられた住所一覧ページが多数存在する。これらの多くは人手によりまとめられたもので、場所名と住所についての情報がページ内に規則的に配置しまとめられているため、この規則性を獲得できればページ内の場所名と住所のペアを高精度で獲得できる。

本研究では、この住所一覧ページに着目し、ページの中での場所名と住所の記載のパターンを獲得し、場所名・住所ペアの抽出を高精度で大量に行うことで、場所に関する言語資源を拡張していくことを目的とする。

連絡先: 佐藤貴大, 東北大学大学院情報科学研究科, 宮城県仙台市青葉区荒巻字青葉 6-3-09,022-795-7140,takahiro@ecei.tohoku.ac.jp

2. 関連研究

ウェブページには多くの情報が存在するが、ページ毎にその表記の方法が異なるためそこから情報を抽出しまとめるのは容易ではない。このため、ウェブページを対象とした数多くの情報抽出の試みがなされている。

Wang ら [6] は教師なしシステムとして同一ウェブサイトのページの類似性を用いて抽出を行う手法を提案した。まず同一のウェブサイトから得られたウェブページ同士で比較を行い、ウェブサイトで繰り返し出現するパターンを検出し、得られたパターンをもとに抽出を行う。例えば製品に関するページであれば製品名や価格がパターンとして得られることとなる。教師なしシステムではラベル付けなしでパターンを構築するため、ユーザは抽出後に得られた結果の中から必要な部分を選択する必要がある。

Freitag ら [3] は教師ありシステムとして、ウェブページ内のそれぞれの文字列に対して分類器を用いる手法を提案した。ページ内の文字列について人手によるラベル付けを行い、それぞれについて文字列長、文字種、文脈といった素性を与え分類器を学習する。分類器はラベルと素性に基づき、ページからユーザの必要とする文字列を抽出する。

ウェブページの中には人手によって情報がまとめられた一覧ページが存在する。Chang[4] らは一覧ページを対象とした情報抽出手法を提案している。一覧ページにまとめられた繰り返し構造のうちの一つをユーザに提示し、必要な部分を指定させる。ページ内の他の繰り返し構造からユーザに指定された部分を抽出することで必要な部分のみを取り出す。一覧ページ毎にユーザによる指示を必要とするため、大量のページを扱う抽出にはあまり適していない。

また、場所に関する情報に特化した研究として、村山らによる WEB 上の住所一覧ページをもとに場所に関する言語資源の拡張を行った研究が存在する。[5] この研究では、既知の場所名・住所・電話番号の三つ組を与えることでページの繰り返し構造を検出し、新規の三つ組の抽出を行っていた。

まず、住所一覧ページから DOM ツリーを作る。DOM ツリーとはウェブページを各要素をノードとする木構造で表現したものである。ウェブページの各要素にはタグが付けられてい

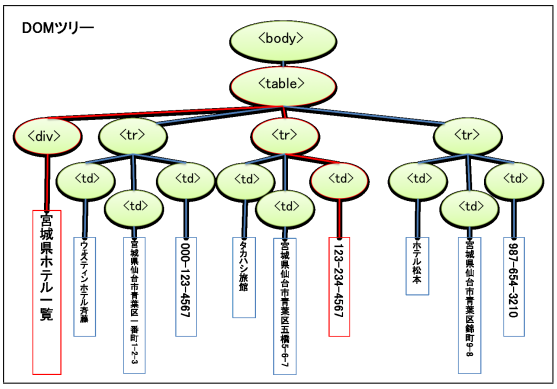

図 1: DOM ツリー

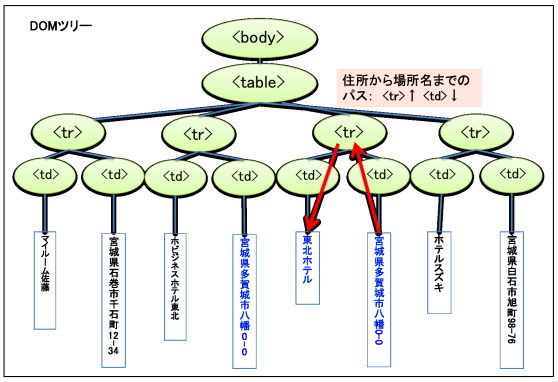

図 2: 住所一覧ページからのパスの構築

るため、DOM ツリーのノードはそれぞれタグを持っている。もとのウェブページの構造から、テキストを含むノードも存在する。DOM ツリーにおいて、あるノードから別のノードまでにたどったノードの並びを「パス」と表現する。例えば、図 1 において、「123-234-4567」を含むノードから「宮城県ホテル一覧」までのパスは

＜ td ＞↑ ＜ tr ＞↑ ＜ table ＞↑ ＜ div ＞↓

のようになる。住所一覧ページを DOM ツリーとして扱うことで、三つ組のページ内でのまとめられ方をとらえることが出来る。抽出は、DOM ツリーからのパスの構築と、発見したパスを用いての三つ組獲得にわけられる。

パスの構築では、既知の三つ組のシードデータとのマッチングを行う。まず、最も簡単に見つかると思われる電話番号のマッチングから行う。DOM ツリーの各要素とシードデータの電話番号とを比較し、次に、マッチした電話番号と組になっている場所名と住所のマッチングを行う。DOM ツリー内に三つ組が発見されたら、電話番号から場所名と住所へのパスを構築する。

三つ組の獲得では、まず、DOM ツリーから電話番号を正規表現を用いて検出する。検出された電話番号を基点として、構築されたパスをたどって新規の場所名と住所を発見する。パスをたどることに成功し、新たに三つ組が発見された場合、新規の三つ組として獲得する。これにより、ページ毎のユーザの指示を必要とせずに住所一覧ページからの大量の抽出を可能とした。

3. 提案手法

村山らの手法をもとに住所一覧ページからの場所名・住所ペアの抽出を行う。村山らの手法において抽出を行う際に問題となっていたのがパスの構築ミスによる誤った抽出と、パスを構築できなかったページから抽出を行えなかったことであった。このため、本研究では誤ったパスの除外と、パスが構築出来なかったページからの抽出に主眼をおいている。また、場所名・住所ペアのみを抽出の対象とすることで、電話番号の記載の無いページからの抽出も可能とした。本手法は大きく 4 つの工程からなる。

パスの構築 住所一覧ページから DOM ツリーを作る。DOM ツリーとシードデータとのマッチングをもとにして住所から場所名までのパスを構築する

パスへの制限 構築されたパスの中で誤ったものを除くため、使用するパスに制限を加える。

パスの共有 パスの構築が出来なかったページからの抽出を行うため、URL のドメインと階層が同じページ間でパスを共有する。

新規獲得 正規表現により新規住所を見つける。見つかった住所からパスをたどり、対応している場所名を発見し、新規ペアを獲得する。

以下にそれぞれの行程の詳細を示す。

3.1 パスの構築

住所一覧ページからのパスの構築を行う。まず、ウェブページの構造を扱うため、ページから DOM ツリーを作る。DOM ツリーの各ノードについて、シードデータとのマッチングを行う。全ノードに対してマッチングを行い、ページ内の全ノードについて既知の文字列かどうかを判定する。検出されたすべての既知の文字列に対して、シードデータをもとに正しい組み合せを調べる。既知の場所名と住所の正しいペアを特定したら、DOM ツリーをたどり、住所から場所名までのパスを構築する。

3.2 パスへの制限

住所一覧ページのパターンを上手く表現しないパスが構築されることがある。原因として以下のものがあげられる。

同じ建物への表記の揺れから２度取り上げられた場合
住所一覧ページを作る際に、同じ建物について異なる呼ばれ方がなされていたため、ページ内で重複が起きてしまったケースや、ページ編集者が意図的に２度あげたものなどが考えられる。例えば図３の例では「ビジネスホテル東北」と「東北ホテル」が同一の住所として記載されている。「東北ホテル」を既知の場所名としたときパスは二通り存在することとなり、このとき「ビジネスホテル東北」の住所を基点としたパスは誤りである。

同一の住所に複数の場所名が存在する場合
住所に存在する建物がビルである場合などがこれに該当する。例えば「東北書店」と「佐藤食堂」は共に東北ビルのビル内に存在しているため、共にその住所は「宮城県仙台市青葉区〇丁目××-×」となる。

以上のことから、高精度の場所名・住所ペアの抽出を行う上で、正しいパスの選択は必要である。本研究では、場所名・住

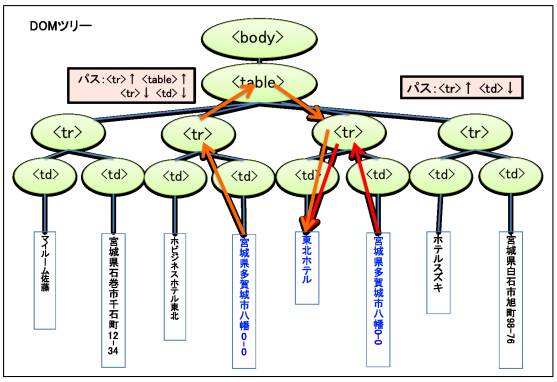

図 3: 誤ったパスの構築

所ペアの獲得に用いるパスに対する制限として構築されたそれぞれのパスについて、ページ内で最も高頻度で出現したパスを用いる頻度最大パスと、場所名から住所に至るまでにたどったノードの数を距離として、ページ内で最も距離の短いパスを用いる距離最短パスの２つの制限について比較を行った。２種類の制限の優劣は自明でないため、実験により比較を行う。

3.3 ページ間のパス共有

本手法では住所一覧ページからパスを構築する際に、既知の場所名・住所ペアとのマッチングを必要とするため、１つも既知のペアが存在していないページからは抽出が行えないという欠点がある。

パスを構築できなかったページからの抽出を行い新規獲得ペア数を増加させるために、本研究では URL のドメインと階層が同じページ間でのパスの共有を行う。これによりページ内で住所から場所名までをたどれるようになり、場所名・住所ペアの抽出が可能となる。

パスの構築の行程においてパスを構築できたページについて、URL のドメインと階層を確認する。同ドメイン同階層の住所一覧ページであれば住所から場所名までのパスが同じであるという仮定に基づき、コーパス内でドメインと階層が同じ住所一覧ページからのペア獲得を行う際、たとえそのページ内でパスを構築することが出来なくても、ドメインと階層が同じ他のページで構築されたパスを利用する。

3.4 パスに基づく新規ペア獲得

各住所一覧ページについて、ページ毎に構築または共有したパスをもとに新規の場所名・住所ペアの獲得を行う。

住所一覧ページから作られた DOM ツリーの各ノードについて、「県名、市町村区名、非記号文字列、番地名」による正規表現を用いて住所を検出する。検出された住所毎にパスをたどって対となる場所名を見つける。ツリーに対応するノードが存在せずパスをたどれない場合や、パスをたどった先のノードが文字列を持たない場合はその住所からの獲得を行わないものとする。

4. 実験

抽出に利用するパスとして適切なものを選択するために効果的な制限を実験により比較した。これにより抽出に利用するパスを制限することによる精度の向上と、その精度が実用に耐えうるものであるかを調査した。また、パスを構築できずに抽出が行えなかったページを活用するため、住所一覧ページ間でのパスの共有を行った。パスの共有による新規獲得ペア数の増加と、適合率の変動を調べた。

4.1 実験設定

実験にはウェブから収集された約 6700 万ページの日本語ページを持つ ClueWeb09 の資源のうち約 700 万ページから、２種類以上のルートからのパスが同じ住所を含む住所一覧ページ約２万ページを用いた。

シードデータとしては Yahoo!ロコから抽出された宮城県内の場所名・住所ペア約 10 万対[*1]を与えた。

4.2 評価尺度

評価は宮城県内のペアについて人手による判定を行い、適合率, 網羅率 (全体), 網羅率 (新規), 全新規獲得ペア数の 4 種類により行った。網羅率 (全体) はコーパスに含まれるペアのうち本手法で対象とすることの出来たものの割合、網羅率 (新規) は網羅率 (全体) から既知のペアを除いたものである。適合率、網羅率 (全体), 網羅率 (新規) の定義はそれぞれ以下の (1) 式,(2) 式,(3) 式に示す。

$$\text{適合率} = \frac{\text{宮城県内正解ペア数}}{\text{抽出された宮城県内の場所名・住所ペア数}} \quad (1)$$

$$\text{網羅率 (全体)} = \frac{\text{宮城県内正解ペア数}}{\text{コーパス内の全宮城県内住所件数}} \quad (2)$$

$$\text{網羅率 (新規)} = \frac{\text{宮城県内新規正解ペア数}}{\text{コーパス内の未知の宮城県内住所件数}} \quad (3)$$

4.3 実験結果

実験の結果を表 1 に示す。

表 1: パスの共有

パスへの制限・パスの共有	適合率	網羅率（全体）	網羅率（新規）	新規獲得ペア数（対）
制限：なし，　共有：なし	0.743	0.252	0.196	2,747
制限：頻度最大，共有：なし	0.916	0.251	0.195	2,435
制限：距離最短，共有：なし	0.997	0.247	0.191	1,963
制限：距離最短，共有：あり	0.997	0.320	0.269	20,936

パスの制限を行わなかったときと比較し制限を加えることで適合率が約 0.20 ほど上昇した。網羅率（新規）の低下がほとんど見られないことから、減少した新規獲得ペアの大部分は誤抽出によるものであったといえる。適合率は距離最短パスを選択したとき最も高く、0.997 であった。これは抽出の精度として実用に堪えられるものであるといえる。

また、ページ間のパスの共有により適合率を下げることなく網羅率 (全体) を約 0.07 上昇させ、新規獲得ペア数に関しては約 10 倍の増加が見られた。パスに対して制約を加えず、ページ間のパスを共有も行っていない結果と比較すると、適合率で約 0.25、網羅率 (全体) で約 0.07、新規獲得数で約 17,000 対もの上昇を確認できた。このことから同ドメイン同階層ページ間ではページの構造が類似し、抽出に利用するパスを共有出来ることがわかった。なお、本実験において評価に用いた宮城県内のペアに関して、パスの共有に起因する誤抽出は見られなかった。

4.4 分析

図 3 で示したように、誤ったパスは正しいパスよりもその出現回数は少なく、住所から場所名までの距離が遠い。このた

*1 コーパス内の WEB ページの収集年 2009 年に対しシードデータの収集年は 2012 年である

め、抽出に使用するパスに制限を加えることで誤ったパスを除去することが可能となり、適合率を上昇させることを可能とした。

頻度による制限より距離による制限のほうが適合率が高かった原因としてはページ内における正解パスの出現率の低さにあると考えられる。実験においてパスの構築に利用された既知の場所名・ 住所ペアの数 150 対に対しパスを構築できたページ数は 63 ページであった。すなわちパスの構築に利用された 1 ページ当たりの既知のペアの数は約 3 対程度にしかならなかったこととなる。このため、誤ったパスと正しいパスの出現回数が共に 1 回のページが出現してしまったため、誤ったパスを用いた抽出が発生してしまった。一方、距離による制限はパスを構築するための既知のペアの数には依存しないため、誤ったパスを除去することが出来た。

パスの共有をしなかった時の新規獲得ペア数は約 2,000 対であった。これは使用したコーパスの住所一覧ページが約 20,000 ページであったことから、不十分であるといえる。本手法では住所からパスをたどり場所名を発見することで抽出が行われるため、利用できるパスが存在しないページを抽出対象とすることは出来ない。しかしながら、今回パスの構築に利用された既知の場所名・住所ペアの数はわずか 150 対であった。この原因としては表記揺れによるものが考えられる。今回はシードデータとして yahoo!ロコの 1 サイトのみから収集された場所名・住所ペアを利用している。このため、yahoo!ロコのものと表記が異なる場合マッチングが行えない。このような場合、たとえシードデータに存在しているものと同一の場所名・住所であってもパスが構築できない。また、このほかにもコーパスが収集された 2009 年からシード データが収集された 2012 年までに発生した東日本大震災により、同一住所の建物が変わってしまい、上手くパスを構築できなかったことも考えられる。

ページ内でパスを構築できるページが少ないことから、ページ間のパスの共有を行うことで抽出の規模を拡大することができると考えられる。今回の実験では、パスの共有を行わなかった場合に比べて、パスを共有することで約 10 倍の新規獲得ペア数の増加が見られた。今回使用したコーパスを調査したところ、コーパス内には異なる 10 ページ以上を含むド メインが 271 種類存在していた。このうち今回パスの構築に成 功した 63 ページの住所一覧ページは 35 種類をカバーしてい た。パスの共有によりこの 35 種類のドメインに含まれるペー ジからの抽出が可能となったため、新規獲得数の大幅な増加が見られた。

パスの共有によって新たに抽出の対象とすることが出来たペー ジ数は約 5000 ページであった。このページ数はパスを構築で きた 63 ページに対し約 80 倍にもなる。しかし、今回の実験 で使用した全住所一覧ページ数が 2 万ページであったことか ら、実際にはコーパス内の約 4 分の 1 ほどしか活用できては いないこととなる。この問題に関しては、共有前のドメインの 種類が全コーパス内に 10 回ページ以上出現するドメインの内 約 0.13 にとどまっていることから、より多くのドメインから のパスの構築により、パスの共有後に抽出に利用できるページ の割合が上昇すると考えられる。

5. まとめ

本稿では獲得したパスに対する制限とページ間のパス共有による、ウェブ中の住所一覧ページからの場所名・住所ペアの高精度な大量抽出の手法を提案した。実験の結果本手法において、パスに対して加えるべき制約はパスの長さに基づく最短パス選択であり、さらにパスの共有により、高精度かつ大量の場所名・住所ペアの獲得が可能であった。実験により得られた精度は実用にたえられるものであり、また、新規獲得数も大きく向上したといえる。

しかしながら、4.4 節でも述べた通り、パスを構築できたページ数はコーパス全体のうちのごく少数で、URL のドメインと階層に基づくパスの共有を行ってもまだなお利用可能なパスを持たなかったことにより抽出が行えなかったページが多く存在する。

今後の課題として適切なシードデータの検討、表記揺れによりマッチングが行えなかった場所名への対応によりパスを構築出来るページ数を増加し、パスを共有できるドメインを増やすことでより大規模な抽出が期待される。

謝辞

本研究は、文部科学省科研費 (23240018)、文部科学省科研費 (23700159)、および JST 戦略的創造研究推進事業さきがけの一環として行われた。

参考文献

[1] Chia-Hui Chang, Kayed M., Girgis M.R., Shaalan K.(2006) A Survey of Web Information Extraction Systems. IEEE Transactions on Knowledge and Data Engineering, TKDE-0475-1104.R3

[2] Hammer, J., McHugh, J. and Garcia-Molina, Semistructured data: the TSIMMIS experience. In Proceedings of the 1st East-European Symposium on Advances in Databases and Information Systems (ADBIS), St. Petersburg, Rusia, pp. 1-8, 1997.

[3] Freitag, D., Information extraction from HTML: Application of a general learning approach. Proceedings of the Fifteenth Confer- ence on Artificial Intelligence (AAAI-98).

[4] Chang, C.-H. and Kuo, S.-C. OLERA: A semi-supervised ap- proach for Web data extraction with visual support. IEEE Intelli- gent Systems, 19(6):56-64, 2004.

[5] 村山 紀文, 南野 朋之, 奥村 学. 見たデータ付与のための住所録自動生成. 情報処理学会研究報告. 自然言語処理研究会報告 2004(73), 41-47, 2004-07-15.

[6] Wang, J. and Lochovsky, F. H., Data extraction and label assign- ment for Web databases, Proceedings of the Twelfth International Conference on World Wide Web (WWW), Budapest, Hungary, pp. 187-196, 2003.