

# Web ニュース記事視覚化のための情報抽出 —地名と地名参照情報との関連付け—

石田 大和<sup>†</sup> 関 洋平<sup>†</sup> 青野 雅樹<sup>†</sup>

豊橋技術科学大学<sup>†</sup>

## 1 はじめに

近年、テキストから地名を抽出することにより、情報を地理的に把握する研究が行われている[2][3]。しかし、テキスト内に複数の地名が出現した場合に、これらの地名が表す情報を、視覚的にユーザに提示するシステムは実用化されていない。

本研究では、1つの文書中の、複数の地名が示す情報を区別するためのラベル付けを行うことで、地理的な情報を視覚的にユーザへ提示するシステムを提案する。提案の有効性を評価するために、まず、事件・事故のニュース記事を対象として、ラベル付けを行うシステムを実装した。本研究では、各地名は4つの地名参照情報（**発生場所**、**被害者住所**、**加害者住所**、**組織所在地**）のいずれかを持っているものと仮定し、これらをラベルとして、抽出した地名への自動付与を行った。

## 2 関連研究

ニュース記事から地名を含む事件情報を抽出する研究は数多くなされている。金山ら[1]は、事件について報道している記事から意味構造を抽出することにより、関連人物のプロフィールや、事件の動機などを要素とする犯罪スキーマを動的に構成した。

また、ニュース記事の視覚化を行うシステムとしては、株式会社クネヒトのハザードマップ[2]が知られている。これは、複数の配信元から提供される新聞記事から地名を抽出し、1つの事件に対して1つの関連する地名を地図上にプロットすることで視覚化を行っている。

本研究では、事件・事故を扱うニュース記事からの地名抽出を行い、各地名に地名参照情報ラベルを付与する。また、1つの記事中に出現する複数の地名をそれぞれのラベルと共に地図上にプロットすることで、ニュースの地理的な広がりユーザが容易に理解できることを目標としている。

## 3 地名抽出ツール LONE

本提案システムでは、記事中に出現する地名のうち、地図にプロットすべき地名のみを抽出するために、自作の地名抽出ツールである LONE[3]を使用した。これは、都道府県や市区町村などの住所階層を元に、詳細な表記の地名を優先して抽出するツールである。2006年12月6日に配信された事件・

事故を扱う 86 記事中、地図にプロットすべき 144 地名を正解とした実験で、LONE の性能は、精度：0.952、再現率：0.819、F-値：0.881 であり、高い精度で地名抽出が可能である。

## 4 地名参照情報の判定

本提案システムでは、LONE によって抽出された地名の集合を入力として、各地名の周辺に存在する手がかり語を用いたルールにより、4つの地名参照情報をラベルとして、それぞれの地名に付与する。地名参照情報判定ルールは、優先順位を付けた4つのルールで構成される。

判定のために、Yahoo! ニュース[4]から取得した16トピック130記事を訓練データとし、有効と考えられるルールを手で作成した。なお、判定の前処理として、ふりがなや人物の年齢表記を削除するため、括弧内が平仮名や数字のみの記述をあらかじめ削除した。以下に、4つの地名参照情報判定ルールを示す。

### 4.1 地名の直前の語による判定

このルールでは、**被害者住所**と**加害者住所**、および**組織所在地**の判定を行う。入力された記事中で、LONE 抽出地名が ( ) や == で囲まれている場合、その直前の手がかり語によって、地名参照情報を判定する。ここで用いられる手がかり語を表1に示す。

加害者住所を判定する手がかり語が、被害者住所のものより数が少なく限定されているのは、本来加害者ではない人を加害者とするのを避けるためである。なお、組織所在地を判定する手がかり語は定義せず、表1に示す手がかり語以外の語としている。これは、記事中での組織の表現が多様であり、特定の手がかり語のみで判定するのが困難なためである。

この判定ルールでラベルの付与が行われる例を、図1に示す。この例では、表1中の手がかり語としては定義されていない、「商事」が地名を含む括弧の直前に出現するため、組織所在地と判定される。

表1 人物住所の手がかり語 (一部)

地名参照情報	手がかり語
被害者住所	さん/ちゃん/くん/男性/女性/学生
加害者住所	容疑者/被告

会見を行った〇〇商事（豊橋市）は…  
⇒「豊橋市」は”組織所在地”

図1 組織所在地ラベル付与の例

Information Extraction and Visualization for Web News  
- Labeling of Location Name with Reference Information -  
† Hirokazu Ishida, Yohei Seki and Masaki Aono  
† Toyohashi University of Technology

## 4.2 文節を用いた判定

このルールでは、**被害者住所**、および**加害者住所**の判定を行う。あらかじめ LONE 抽出地名を含む 1 文を CaboCha によって文節ごとに分割し、文頭から次の手順で確認する。表 1 に示す手がかり語が、LONE 抽出地名から後ろ 2 文節以内に出現した場合、その地名を被害者住所、または加害者住所と判定する。この判定ルールでラベルの付与が行われる例を、図 2 に示す。

名古屋市の自営業、佐藤容疑者を…  
⇒「名古屋市」は“加害者住所”

図 2 加害者住所ラベル付与の例

## 4.3 場所を表す格助詞による判定

このルールでは、**発生場所**の判定を行う。LONE 抽出地名から文末までの間に、「で、」または「に、」が出現した場合、その地名を発生場所と判定する。この判定ルールでラベルの付与が行われる例を、図 3 に示す。

豊川市の路上で、帰宅途中の女性が…  
⇒「豊川市」は“発生場所”

図 3 発生場所ラベル付与の例 (1)

## 4.4 文末表現による判定

このルールでは、**発生場所**の判定を行う。LONE 抽出地名が出現する文中において、読点および句点の直前の 6 文字に、事件・事故の発生を示唆する手がかり語 (図 4, 全 18 語) が出現した場合、その地名を発生場所と判定する。この判定ルールでラベルの付与が行われる例を、図 5 に示す。

あった、わかった、負った、死亡、墜落、逮捕、発見、炎上、出火、見つ

図 4 発生場所の手がかり語 (一部)

一宮市の住宅から出火、焼け跡から…  
⇒「一宮市」は“発生場所”

図 5 発生場所ラベル付与の例 (2)

## 5 実験

提案システムの地名参照情報判定ルールの有効性を調べるために、評価実験を行った。実験対象の Web ニュース記事としては“Yahoo! ニュース”を用いた。

### 5.1 実験方法

実験では、LONE によって記事から抽出された地名集合の要素を対象としている。

LONE で抽出された地名に対して、人手で 4 つの地名参照情報を付与し、システムが付与したラベルとの比較を行った。テストデータは、Yahoo! ニュースで 2008 年 1 月 6 日と 7 日に配信された国内社会カテゴリの全 315 記事中、事件・事故を扱っている 160 記事とした。

## 5.2 実験結果

実験結果を、表 2 に示す。なお、LONE 抽出地名は LONE によって抽出された地名数、ラベル付与地名はシステムがラベルを付与した地名数、正解地名は付与したラベルが正解であった地名数とする。

表 2 実験結果

地名参照情報	LONE 抽出地名	ラベル付与地名	正解地名
発生場所	130	119	103
被害者住所	50	51	46
加害者住所	50	37	36
組織所在地	23	22	13
合計	253	229	198
	精度	再現率	F-値
発生場所	0.866	0.792	0.827
被害者住所	0.902	0.920	0.911
加害者住所	0.973	0.720	0.826
組織所在地	0.591	0.565	0.578
全地名	0.865	0.783	0.822

実験結果から、訓練データを用いて人手で作成した地名参照情報判定ルールが、他の一般的な事件・事故を扱うニュース記事にも適用可能であることが確認できた。高い精度が得られているため、本システムの目標である地図を用いたニュース記事の視覚化において、誤った情報の提示を避けることができる。一方、組織所在地ラベルの付与において、判定ルールで明示的な手がかり語を利用していないことによるとと思われる判定ミスが多く見受けられた。対策として、地名参照情報判定への、固有表現を利用したルールの追加が挙げられる。

## 6 おわりに

ニュース記事中の地理的な情報を、視覚的にユーザへ提示するシステムへの第一歩として、事件・事故を対象とした地名参照情報判定システムを提案した。実験の結果、今回作成した判定ルールが有効であることが確認できた。今後は、手がかり語の妥当性を統計的に調査することや、事件・事故以外のニュース記事を対象とした地名参照情報の設定、および判定ルールの作成を行うことを検討している。

### 参考文献

- [1] 金山淳一, 北條孝, 田村直良. “文章の構造解析による新聞記事からの事件情報抽出”, 情報処理学会研究報告 自然言語処理, Vol.2002, No.104 pp. 1-6, 2002.
- [2] 株式会社クネヒト, Knecht [クネヒト]-ハザードマップ [online], c2006-2008. [cited 2008-01-08]. Available from: <http://hazard.knecht.jp/>
- [3] 石田大和, 関洋平, 青野雅樹, “Web 上のニュース記事視覚化のための地名抽出手法”, 電子情報通信学会, 東海支部, 卒業研究発表会, 名城大学, 2007.
- [4] ヤフー株式会社, Yahoo! ニュース [online], c2008. [cited 2008-01-08]. Available from: <http://headlines.yahoo.co.jp/hl>