地理的抱合関係を用いた自動ニュースマップの実装

2004MT089 櫻井敦規

指導教員 河野浩之

1. はじめに

最近では, Web ページや新聞記事から必要な情報を得るためにテキストマイニングが導入されている. テキストマイニングとは, テキストから有益な知識や情報を抽出する手法である[1]. テキストマイニングは新聞記事の分類やスパムメールのフィルタリングなど多様な領域で適用されている. 地域のニュースから情報を得るには地名は重要な要素となる. そして, 地名との共起関係に着目して街の話題語抽出[2]に関する研究が行われている. この研究では, イベントと関係のない地名が対象に含まれるという問題点があった.

本研究では, 地域ニュースを対象に地名の地理的抱合関係を考慮した地名抽出を行い, 抽出精度の考察を行う. そして, 抽出した結果を地図上に記事を表示していく. これは, 地名など抽出した結果を地図上に表示することはあまり行われていないので, 視覚的に分かりやすくするために行う.

2. 地理的抱合関係を用いた地名抽出

安村ら[4]は, イベントが発生した時, Web にイベントに関する Blog エントリが作成されるものとして, Blog エントリの収集をしている. そして, Blog エントリからイベント情報を抽出している. イベント情報は, イベントの発生時間と発生場所としている. イベントを処理できる地理情報システムの構築をするため, Web からイベント情報を抽出する手法の提案を行っている.

日本の地名には同名の地名が複数存在する. 地名抽出で誤抽出に対処するために地名の地理的抱合関係を考慮して地名の登録を行っている. 地理的抱合関係とは, 「下位階層の地名に対応する領域が, 上位階層の地名に対応する領域に含まれる」という関係である.

3. 地名抽出とニュース記事の視覚化

ニュース記事のシステム構成図について, 図 1 を用いて説明する.

3.1. ニュース記事収集と前処理

本研究では, ニュース記事を用いる. まず, wget で毎日新聞の地域ニュースを収集することにする<図 1 の(1)>. wget は HTTP や ftp を用いて, Web ページからファイルをダウンロードすることができる. 収集したデータを1つの CSV 形式のファイルにする.

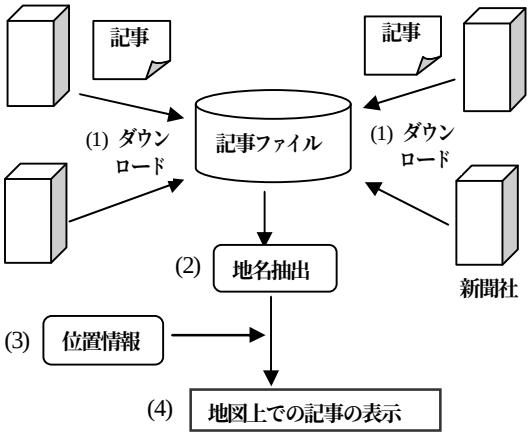

図 1 ニュース記事の視覚化システムの構成図

地名の抽出するための前処理として, 茶筌(version2.4.1)[*1] を用いてニュース記事に書かれている文書の形態素解析を行う. 解析された記述はそれぞれ見出し語ごとに分かち書きされる. 例えば, 「バングラデシュ:大型サイクロン被災者を支援、日進市内9カ所に救援金箱 ／愛知」という記述が与えられた場合, 以下のように分かち書きされる.
「バングラデシュ｜:｜大型｜サイクロン｜ ｜被災｜者｜｜支援｜、｜日進｜市内｜9｜カ所｜に｜救援｜金｜箱｜／｜愛知」

この形態素解析の結果から地名抽出を行っていく<図 1 の(2)>.

ニュース記事ごとに, 文書ベクトルを与える. ニュース記事に出てきた地名の回数を求め, 回数の一番大きいものを用いる. TF-IDF 法はテキストからキーワード抽出するために用いられるが, 求めているものと違ったものが得られる可能性もある. したがって, 地名の出現回数で文書ベクトルを与えていくことにする.

3.2. 地名データ作成

抽出した地名に緯度･経度の位置情報を付加するために地名データを作成しておく. 地名は地理的抱合関係を考慮して行う.

本研究では, 位置情報を CSV アドレスマッチングサービス[*2] を利用して収集する. このサービスは, 東京大学空間情報科学研究センターが提供している. CSV 形式で保存されている住所データを緯度･経度を付加したデータに変換を行っている. 緯度･経度の変換に利用されているデータは国土交通省の提供する「街区レベル位置参照情報[*3]」で

[*1] ChaSen's Wiki, http://chasen.naist.jp/hiki/ChaSen/
[*2] CSV geocording Service, http://newspat.csis.u-tokyo.ac.jp/cgi-bin/geocode.cgi?action=start

ある.

3.3. ニュース記事の視覚化

地名データと照らし合わせながら，抽出した地名に緯度･経度の位置情報を付加していく＜図 1 の(3)＞．地名を抽出して，そこにニュース記事を表示するようにする．本研究では Google Maps を用いて視覚化処理を行う＜図 1 の(4)＞．Google Maps は Web ページでも利用ができ，提供されるほとんどの機能を利用できる API (Application Program Interface) が公開されている[3]．本研究では以下の2つの機能を備えたシステムを実装できるようにする．

1. 抽出した地名にマーカーを作成する．
2. 地方ごとに，マーカーの色を変えて表示する．
3. マーカーをクリックすると情報ウィンドウを表示し，記事のタイトルと内容を表示する．

4. ニュースマップ作成と評価

4.1. ニュースマップ作成

Google Maps を用いて，地名抽出した結果を地図上で表示するニュースマップの作成を行った．プログラム行数は約430行であり，地図上で記事表示するプログラムの一部は図2のようになる．

```
function addNewsItem(place,lng,lat,num,,
                                area,title, content) {
    var marker = createmarker(lng,lat,area);
     map.addOverlay(marker);
GEvent.addListener(marker, "click", function() {
     marker.openInfoWindowHtml(html);
  });
```

図2 地図上での記事表示プログラム

addNewsItem() 内での処理について説明する．var marker = createmarker(); で地方ごとにマーカーの色を設定してマーカーを生成し，その結果を marker に返している．map.addOverlay() では，生成したマーカーを地図上にニュース記事を表示する．GEvent.addListener() で，マーカーをクリックした時に情報ウィンドウを開くようにイベントを設定し，抽出した地名，地名ナンバー，記事のタイトルと内容を表示する．図2のプログラムを実行すると，図3のようになる．

4.2. 地名抽出とニュースマップの評価

本研究では 1054 件の毎日新聞の地域ニュースを用意する．地域ニュースとなっている場所と地名抽出した地名と一致しているかで判定を行う．

この方法で判定を行うと 1054 件のうち885件が正しくなり，分類精度は 83.96%となった．これは地理的抱合関係を用いることにより，同名の地名に対する誤抽出に対処できたからだと考えられる．

しかし，地域によっては精度の低いところも見られた．原因として，記事の内容によって精度に影響を及ぼしていることがわかった．例えば，宮崎県の記事では内容が大分県のことについて詳細に記述されているために抽出自体はうまくいっていても，本研究の評価方法で判定した場合は抽出ができていないことになってしまった．

また，茶筌の性能も関係していることが分かった．山口県にある「周南市」を茶筌で形態素解析すると「｜周｜南｜市」と分かち書きした．その結果，「南」と誤抽出されてしまった．地名の判定方法を考え直したり，茶筌の辞書を使用したりすることで問題解決すると考えられる．

図3 ニュースマップ

5. まとめ

本研究では，毎日新聞の地域ニュースを収集して，システムでどの程度元のニュースのように再現されているかで性能評価を行った．そして，抽出した結果を視覚的に分かりやすくするために，Google Maps を用いて地図上で視覚化も行った．記述者の傾向によって精度は影響してくるが，それでも地理的抱合関係を用いた地名抽出の性能は高く，83.96%という結果が得られた．これは，記事の内容を重視したもので実践的なものといえる．

参考文献

[1] Andreas Hotho, Andreas Nürnberger, Gerhard Paaß, “A Brief Survey of Text Mining (GLDV-Journal) , ” fuer Computerlinguistik und Sprachtechnologie , pp.27-30, 2005.

[2] 倉島健, 手塚太郎, 田中克己, “Blog からの街の話題抽出手法の提案, ” DEWS2005 2C-i10, pp.3-5, 2005.

[3] Rich Gibson, Schuyler Erle, “Google Maps Hacks 地図検索サービス徹底活用テクニック, ” O’REILLY, 2006.

[4] 安村祥子, 池崎正和, 渡邉豊英, 牛尼剛聡, “blog マッピングを用いたイベント情報抽出,” DEWS2007 D8-3, pp.4-5, 2007.

*3 街区レベル位置参照情報ダウンロードサービス, http://nlftp.mlit.go.jp/isj/