様式第２号

平成２８年度　　地 域 貢 献 特 別 研 究 費 実 績 報 告 書

平成２９年３月３１日

<table>
<tr><td>申　請　者</td><td colspan="2">学科名</td><td>情報システム工学科</td><td>職 名</td><td>教授</td><td>氏 名</td><td>菊井　玄一郎</td></tr>
<tr><td>調査研究課題</td><td colspan="7">地域情報テキストに対するエンティティリンキングと情報可視化の研究</td></tr>
<tr><td rowspan="3">調査研究組織</td><td colspan="2">氏　　名</td><td colspan="2">所属・職</td><td>専門分野</td><td colspan="2">役割分担</td></tr>
<tr><td>代表</td><td>菊井　玄一郎</td><td colspan="2">情報システム工学科・教授</td><td>知識情報処理</td><td colspan="2">研究統括，計画立案，情報抽出，wikification手法の検討</td></tr>
<tr><td>分担者</td><td>福圓　琢真
宇田　一平</td><td colspan="2">大学院・システム工学専攻（前期），２年
同上</td><td>方言の形態素解析
地理情報システム，ウエブインタフェース</td><td colspan="2">地方議会会議録に対する形態素解析手法の検討
データ分析・解析結果のウエブインタフェースの検討</td></tr>
<tr><td>調査研究実績の概要
（地域貢献への反映を踏まえて記述のこと）</td><td colspan="7">本研究の最終的な目的は，地方議会の会議録の可視化や直感的な検索など，地域に関するテキスト（文章）の有効かつ効率的な活用に向けて，テキスト中の情報を分析してデータベースのように構造化する技術を開発することにある．　この目的を達成するため，本調査研究では地域に関するテキスト内の事物名を検出して実世界の事物と関係づける技術（エンティティリンキング技術），および，これら事物に対する言及内容を抽出する技術について検討した．前者では特に地域のランドマーク情報を同定する技術，後者では岡山弁の形態素解析技術の開発を進めた．　以下ではこれら２点の概要を説明する．

<u>１.テキスト中の地域関連エンティティ情報の抽出</u>
　総社市議会会議録を研究用データとして，この中で言及されている「ランドマーク」を同定し，地図上の位置に相当する「完全な住所」を得る処理のプロトタイプを開発した．
<u>１）処理の概要</u>
処理全体は２つのステップに分けられる．１つめは「ランドマーク表現（実体参照表現）の候補の検出」であり，２つめは「ランドマーク表現と実世界の事物との対応付け（認識）」である．以下では，まず，処理全体で必要とするランドマークデータベースの整備について述べ，次に各ステップについて説明する．
<u>２）ランドマークデータベースの構築</u>
まず，ウエブ上のオープンデータ（例：県内の小学校一覧）に対するパターン照合により，ランドマークと住所の対を収集した．ランドマークが複数のオープンデータに存在する場合はなるべく住所が完全なもの（長いもの）を選んだ．得られた住所に対してGoogleAPIのジオコーダーを適用して緯度経度を求めた．
<u>３）ランドマーク表現候補の検出</u>
日本語係り受け解析ソフトCaboChaに付属する固有表現抽出処理を用いてランドマーク文字列を追加してテキスト中のランドマークの候補を検出する．
<u>４）ランドマーク表現の実世界の事物との対応付け</u>
３）で得られた候補と２）のデータベースを照合して前者を後者のレコードに対応づける．データベースには正式名称が登録されているのに対して，テキストでは省略形などの変化形が出現して完全一致では同定漏れが発生する．そこでランドマークの種別（例：学校，病院）ごとに，省略形や変化形をルール化しこれを用いて照合を行う．
<u>５）評価結果</u>
精度評価の結果，79.4%の精度でランドマークが正しく同定できることが分かった．</td></tr>
</table>

次頁に続く

調査研究実績の概要

（地域貢献への反映を踏まえて記述のこと）

２．岡山弁の形態素解析技術の検討

近年，日本語解析処理技術は格段の性能向上を遂げたが，方言に対応するものはほとんどない．一方，地方議会会議録など岡山弁のテキストは存在し，さらに，今後，自動対話システムの普及が見込まれることから，計算機で方言を扱うことは必須になると考えられる．我々は，既に，標準語のモデルパラメータを語彙情報に基づいて自動変換することにより岡山弁の形態素解析が可能になることを示している．本調査研究では，ベースとなる標準語モデルの改良を試みるとともに，新たに作成した岡山弁データを用いて評価した．

１）解析手法の概要

解析モデルとしてクラスngramによるものを採用した．具体的には下記で与えられる$\log p(W,C)$を最大化する$W = w_1, w_2, \cdots, w_n$，$C = c_1, c_2, \cdots, c_n$を求める．ここで，$w_i$，$c_i$はそれぞれi番目の形態素表層形と形態素クラスを表す．

$$\log p(W,C) \cong \sum_i (\log p(w_i|c_i) + \log p(c_i|c_{i-1}))$$

なお，今回の調査研究により，形態素クラスとして用いる属性を前向きクラスと後ろ向きクラスで別にした（多重クラス）方が有効であることが分かったためそちらを採用した．

２）モデルパラメータの推定

まず，大規模な新聞コーパスから標準語用のパラメータを推定した．このコーパスは規模が大きくカバレージが広いという利点を持つが，岡山弁の付属語に対応する（標準語の）口語表現が少ない．そこで，「CSJ日本語話し言葉コーパス」のパラメータ混合すること(mixture model)により，精度向上を図った．次に標準語のパラメータを利用して岡山弁表現のパラメータを推定した．例を以下に示す．

$$\log P(\text{ゆーかす}|\text{連語}) = \log P(\text{言っ}|\text{動詞}) + \log P(\text{助詞}|\text{動詞}) + \log P(\text{て}|\text{助詞}) + \log P(\text{助詞}|\text{動詞}) + \log P(\text{聞か}|\text{動詞}) + \log P(\text{動詞}|\text{動詞}) + \log P(\text{せる}|\text{動詞})$$

３）評価結果

評価用の岡山弁データとして新聞投書欄の岡山弁コーナーなどの文章を参考に約3000語（形態素）の岡山弁データを作成し，正解解析結果を人手で付与した．この正解と自動解析結果を比較して評価した結果，次の表のように精度が大きく向上することが分かった

	適合率	再現率	F 値
既存手法	0.770	0.806	0.788
今回の手法	0.878	0.906	0.892

３．そのほかの活動等

・第17回．ことばの祭り・建部（主催：岡山弁協会，岡山市芸術祭参加）における講演(2016.12.4)

・岡山県立大学公開講座講師：計算機に岡山弁をしゃべらせる(2016.9.3)

・新聞取材１件，テレビ取材２件

成果資料目録

福圓琢真,菊井玄一郎,但馬康宏: ″テキストを入力とする岡山弁音声合成の試み″, 第18回音声言語シンポジウム，音声言語情報処理研究会, 2016-SLP-114(17),1-4 (2016-12-13) , 2188-8663 (2016)