様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　６　月　２４　日現在

機関番号：　１４３０１
研究種目：　挑戦的萌芽研究
研究期間：　2013～2015
課題番号：　２５５４０１４０

研究課題名（和文）Webコンテンツのメタデータ自動付与に基づくシンボルグラウンディング

研究課題名（英文）Symbol Grounding based on Metadata Recognition for Web Contents

研究代表者
河原　大輔（Kawahara, Daisuke）

京都大学・情報学研究科・准教授

研究者番号：１０４５０６９４

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：本研究課題では、Webコンテンツを対象として、テキスト中のメタデータ認識と、実世界の参照先の同定を行うシステムを研究開発した。メタデータとしては主にWebコンテンツの著者を対象とし、また地名表現についての参照先同定を対象とした。深い自然言語解析技術に加え、実世界情報を利用することによって、既存のシステムよりも高精度な解析が可能となった。

研究成果の概要（英文）：We developed a system for recognizing metadata of Web contents and real-world referents of named entities. This system identifies authors of Web contents as our target metadata and referents of location expressions in Web contents. By using deep natural language processing techniques and real-world information, we achieved more precise analysis than previous systems.

研究分野：　自然言語処理

キーワード：　シンボルグラウンディング　時空間情報　メタデータ　曖昧性解消　固有表現　Webコンテンツ　ジオコーディング

１．研究開始当初の背景

Web の社会への浸透により、人々はさまざまな情報を自分自身で取得し、また発信することができるようになった。そのような情報の中には、人々の経験が大量に含まれており、これを検索、集約、分析することによって、未来を予測したり、問題解決に利用するということが行われるようになった。このような経験は、一般的に、述語項構造(いわゆる5W1H)による記述として抽出されるが、言語の相対性、曖昧性、多様性により、集約・分析をするにあたって大きな問題をかかえている。すなわち、「昨年」のような相対的時間表現、「私」のような指示詞、「山田一郎」「総持寺」のような曖昧性のある固有名詞、さらには「金閣寺」「鹿苑寺」のような表現の多様性のために、これらを実世界の参照先として集約し、分析することが困難となっている。

２．研究の目的

本研究では、テキスト中の時間表現、場所表現、人・組織表現などの固有表現を認識し、それらに対して実世界の参照先との対応付けを行う。このタスクは固有表現に対するシンボルグラウンディングに相当する。シンボルグラウンディングは、人工知能研究分野において解決されていない最大の問題の一つであり、対象を上記の固有表現に限ったとしてもチャレンジングである。

子どもの言語獲得プロセスに見られるように、実世界の環境と言語理解は密接に関係している。実世界の環境なくして言語理解は実現できず、また逆に、言語理解なくしては実世界の環境を計算機処理することが難しい。そこで、本研究では、テキストの深い解析と実世界への対応付けを統合したモデルを開発する。

３．研究の方法

本研究では、まず、Web コンテンツを対象として、時間、場所、著者情報などからなるメタデータを付与するシステムを開発する。時間、場所などは固有表現認識によってある程度認識が可能であるが、各コンテンツの著者情報の認識はチャレンジングである。基本的には、著者情報の正解が付与されたデータを用いて教師あり学習を行うシステムを開発し、解析結果を分析しながら改良していく。特に、従来用いられていない言語の構造的特徴を考慮する。

次に、Web コンテンツのテキスト中における固有表現と実世界の参照先を対応付けるシステムの研究開発を行う。こちらも、まずは固有表現と実世界参照先が対応付けられたデータを構築し、それを用いて教師あり学習を行うシステムを開発する。データとしては、地名とその参照先(緯度・経度)が対応づけられた Twitter データを構築する。参照先の同定においては、空間的近接性と時間的一貫性を利用することを検討する。

４．研究成果

(1) Web コンテンツを対象として、時間、場所、著者情報などからなるメタデータを自動付与するシステムを開発した。このシステムは、基本的には、Web ページ(HTML)からテキストを抽出し、形態素解析、構文解析、固有表現認識を適用する。これによって、時間、場所、人名、組織名などを認識している。次に、これらの情報および元の HTML の DOM 構造に基づき、著者情報の抽出を行う。

著者情報の抽出は、基本的には、各 Web ページのテキスト部分から固有表現を含む名詞句を著者候補として抽出する。このとき、よりよい著者候補を生成するために、組織名が Web ページの著者となりやすいことを手がかりとして考慮する。この処理としては、組織名の末尾になりやすい語句をあらかじめ収集しておき、これを著者候補生成に用いる。これらの語句を「組織名末尾リスト」を呼び、「大学」や「組合」など 87 語句を収集した。たとえば、「石巻地区広域行政事務組合消防本部」という名詞句が Web ページに存在する場合、「組合」が含まれている位置までの文字列、すなわち「石巻地区広域行政事務組合」を著者候補とする。また、「ソニーエンターテイメント」のように、組織名末尾リスト中の語句が含まれない場合は、これをそのまま著者候補とする。

次に、著者候補のそれぞれについて、品詞情報、構文情報、元の Web ページ中の著者候補の位置などを素性として抽出する。構文情報としては、著者候補の係り元および係り先の名詞(句)を抽出する。たとえば、「ようこそ富士見市公式サイトへ」という文において、「富士見市」が著者候補として抽出されていれば、係り元の「ようこそ」と係り先の「公式サイト」を素性として抽出する。元の Web ページ中の著者候補の位置としては、HTML の DOM 構造における位置を素性として抽出する。

最後に、抽出した素性を利用して、著者候補を Ranking SVM でランキングする。訓練・評価用データとしては、国立国会図書館が作成した 4,502 ページに対して著者情報が付与されたデータを用いた。このデータを用いて、10-fold 交差検定で評価を行った。その結果、出力結果の TOP-1 に正解の著者が含まれている割合が 89.0%、TOP-5 で 94.8%であり、高精度に著者を認識することができた。上記で述べた組織名末尾リストと構文情報を利用しない場合は、TOP-1 で 87.6%、TOP-5 で 90.8%であり、これらの情報が有効に働いたと考えられる。

著者情報の抽出結果を分析したところ、「UNHCR 日本・韓国地域事務所広報室」や「尾道市・御調町・向島町合併協議会」のような著者が抽出できていなかった。前者のような日本語と英語が混ざった著者文字列については、構文解析が誤ることも多く、うまく抽

出できないことが多かった。また後者のような構造が複雑な名詞句について、うまく抽出できないことが多かった。これらの改良は今後の課題である。

(2) Web コンテンツ中の固有表現と実世界の参照先を対応付けについて研究するにあたり、参照先の表示が明確に定義される地名を対象とした。本研究では、地名の参照先を緯度と経度とし、Web コンテンツとしては Twitter のツイートデータを対象とした。ツイートの例として「首里駅から県庁前駅 にきました!」のような文がある。この文における地名「県庁前駅」には曖昧性があり、いくつかの県において同名の駅が存在するが、この文においては沖縄県にある「県庁前駅」を指している。このようなテキストに出現する地名を地名表現(LEX; Location EXpression)と呼び、その指示先を地名エンティティ(LE; Location Entity)と呼ぶ。

まず、日本語 Wikipedia を利用して LEX と LE のデータベースを構築した。緯度・経度情報をもつ記事から LE と LEX を抽出した。たとえば、Wikipedia に「県庁前駅(沖縄県)」という記事があるので、これ自体を LE とし、括弧を除いた「県庁前駅」を LEX とする。このときに同記事に記載されている緯度 26.21 度、経度 127.67 度を抽出する。この処理を Wikipedia 全体に適用した結果、「県庁前駅」という LEX については 7 つの LE が得られ、全体で 17,724 個の LEX について 18,256 個の LE が得られた。

ツイートデータとしては、2011 年 7 月 15 日から 2012 年 7 月 31 日にかけて収集した緯度・経度情報付きツイートを対象とした。まず、このツイート集合から曖昧性のある LEX を含むツイートを抽出した。次に、それぞれのツイートについて、ツイートに付与されている緯度・経度と、対象 LEX に対する LE 候補それぞれの緯度・経度を比較し、もっとも距離が近い LE 候補を LE として採用した。もっとも近い LE 候補でも 10km 以上離れているツイートは対象外とした。この手順によって 18 万ツイートを獲得し、これを学習・評価用のデータとした。このうち、10 ツイート以上得られた LEX は 354 個あった。なお、距離の計算は、緯度 1 度を 111.319km、経度 1 度を 91.187km として行った。

タスクは、ツイートに含まれる曖昧性のある LEX に対して、その LE を識別することである。この識別を行うために、ツイート中の曖昧性のある LEX について素性を抽出し、SVM (2 次多項式カーネル)を用いて学習を行った。SVM は LEX ごとに学習し、多値分類であるため one-versus-the-rest 法を用いた。基本的な素性として、ツイートに含まれる単語(語彙的素性)および LE の頻度(マジョリティ素性)を用いた。これらに加え、空間的近接性と時間的一貫性を素性として採用した。たとえば、「首里駅から県庁前駅 にきました!」というツイートにおいて、「首里駅」は曖昧性のない LEX である。「首里駅」から「県庁前駅」のそれぞれの LE に対する距離を計算し、これを空間的近接性の素性としている。時間的一貫性は、対象ツイートの直前のいくつかのツイートからも語彙的素性、マジョリティ素性および空間的近接性素性を抽出することによって考慮する。

実験では、上述のツイートデータから 10 ツイート以上存在する曖昧性のある LEX を抽出し、70,184 ツイート(354 個の LEX)を対象とした。ツイートの単語分割には、日本語形態素解析システム JUMAN を用いた。結果は表 1 に示す。表において、MB はマジョリティベースライン、B はベースライン(語彙的素性+マジョリティ素性)、SP は空間的近接性素性、TC は時間的一貫性素性を表す。ここでは、各 LEX のツイート数ごとに精度を計算している。「10〜」は全ツイート 70,184 件に対する結果である。この結果より、空間的近接性および時間的一貫性の素性はおおむね有効に働いていると考えることができる。

本研究で対象としなかった問題として、「ツイートに含まれる LEX の場所にユーザーが実際にいるかどうか」という問題が挙げられる。この問題の解決も、位置特化型のサービスなどで重要であり、今後取り組んで行きたいと考えている。

ツイート数	手法	正解数	精度
10〜100	MB	4,171	0.8528
	B	4,485	0.9170
	+SP	4,515	0.9231
	+TC	4,491	0.9182
	+SP+TC	**4,520**	**0.9241**
100〜1,000	MB	22,477	0.8726
	B	24,725	0.9599
	+SP	**24,752**	**0.9609**
	+TC	24,708	0.9592
	+SP+TC	24,737	0.9604
1,000〜	MB	36,896	0.9332
	B	39,041	0.9875
	+SP	**39,054**	**0.9878**
	+TC	39,036	0.9874
	+SP+TC	**39,054**	**0.9878**
10〜	MB	63,544	0.9054
	B	68,251	0.9725
	+SP	**68,321**	**0.9735**
	+TC	68,235	0.9722
	+SP+TC	68,311	0.9733

表 1 地名表現曖昧性解消の精度

〔謝辞〕
本研究を進める上で貴重なデータを提供いただいた国立国会図書館関西館に感謝いたします。

５．主な発表論文等
(研究代表者、研究分担者及び連携研究者に

は下線）

〔雑誌論文〕（計０件）

〔学会発表〕（計３件）

① Daisuke Kawahara and Martha Palmer. Single Classifier Approach for Verb Sense Disambiguation based on Generalized Features. In Proceedings of the 9th International Conference on Language Resources and Evaluation, pp.4210-4213, Reykjavik, Iceland, 2014.5.29.

② 粟村誉，荒牧英治，河原大輔，柴田知秀，黒橋禎夫．ソーシャルメディアにおける空間的近接性と時間的一貫性を考慮した地名の曖昧性解消．情報処理学会 第217回自然言語処理研究会，オホーツク・文化交流センター（北海道網走市），2014.7.4.

③ Takashi Awamura, Eiji Aramaki, Daisuke Kawahara, Tomohide Shibata, Sadao Kurohashi. Location Name Disambiguation Exploiting Spatial Proximity and Temporal Consistency. In Proceedings of the 3rd International Workshop on Natural Language Processing for Social Media, pp.1-9, Denver, USA, 2015.6.5.

〔図書〕（計０件）

〔産業財産権〕
○出願状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
なし

６．研究組織
(1)研究代表者
河原 大輔（KAWAHARA, Daisuke）
京都大学・大学院情報学研究科・准教授
研究者番号： 10450694

(2)研究分担者
なし

(3)連携研究者
なし