1G2-OS-08a-3

固有表現抽出を用いた歴史コンテンツのLOD化支援

Support for the Creation of Linked Open Data from Historical Contents Using Named Entity Recognition

似内　勇太 *1　　奥野　拓 *2
Yuta Nitanai　　Taku Okuno

*1公立はこだて未来大学大学院　　*2公立はこだて未来大学
Graduate School of Future University Hakodate　　Future University Hakodate

Recently, historical contents are published on the Web as well as books. In addition, they have been planned publishing in the format which is available for secondary use. Text of historical contents includes useful information such as period, events and the relevant person. This research adopts the Linked Opnaen Data and supports publishing of useful information in the text. This research proposes to extract resources and to recommend RDF properties using the named entity recognition when the publisher converts data in the text to the LOD. In the experiments named entities is extracted from the text of cultural heritage contents in order to evaluate extraction accuracy. The accuracy is evaluated by an F-measure using cross-validation. An F-measure of 71.8% is shown by the results of experiment.

1. 歴史コンテンツの現状

近年，市史や文化財などの歴史コンテンツは書籍での公開だけではなく，Web 上での発信も活発に行われている．例えば北海道では，北海道の文化財情報を発信している「北海道文化資源データベース」，道南の文化財や博物館の情報を発信している「道南ブロック博物館施設等連絡協議会ブログ」，函館の地域史資料を発信している「函館市史デジタル版」など様々な Web サイトが公開されている．さらに,「道南ブロック博物館施設等連絡協議会ブログ」では，歴史コンテンツのデータを二次利用可能な形式で公開することが計画されている．

現在，様々な種類のデータが Web API などの技術により，二次利用可能な形式で公開されている．しかし，それらのデータは，Web ページの表形式データを CSV 形式などの構造化データに変換したものが多く，文章に含まれる非構造化データはあまり公開されていない．歴史コンテンツの文章には年代や出来事，関連のある人物などの有益な情報が含まれており，それらの情報は二次利用可能な形式で公開する上では重要な情報である．本研究では，二次利用可能な形式として， Web 上にデータを公開する仕組みである LOD（Linked Open Data）を採用し，歴史コンテンツの文章に含まれるデータの LOD 化支援を提案する．

2. 文章中のデータを LOD 化する課題

LOD は標準形式の一つとして RDF（Resource Description Framework）を採用している．RDF は主語・述語・目的語の三つの要素でリソースの関係を表現する．主語となるリソースを URI で記述する．述語は主語と目的語の関係を表し，述語の対象となるリテラルや他のリソースを目的語として記述する．述語はプロパティと呼ばれ，RDFS（RDF Schema）を用いて別途定義し，URI で参照可能にする．Web 上のリソースを URI に変換し，RDF を記述することで LOD 化することができる．構造化データと比べ，文章に含まれるデータを LOD 化する上で主に二点の課題がある．

連絡先: 公立はこだて未来大学大学院 システム情報科学研究科，北海道函館市亀田中野町 116 番地 2，g2114025@fun.ac.jp

一つ目は，文章に含まれるリソースとなり得る様々な単語を人手で抽出する必要があることである．文章は非構造化データであるため，LOD 作成者が文章を読み，内容を把握した後にリソースを抽出する必要がある．そのため，リソースの抽出に多大なコストを要する．また，抽出したリソースの主語と目的語の対応付けや RDF の記述は人手で行う必要がある．

二つ目は，プロパティの選定作業である．LOD 化する際に，広く一般的に用いられているプロパティを利用することで，アプリ内でプロパティ同士の関連付けを行うことなく，同じプロパティを使用している LOD と連携することができる．そのため，既存のプロパティを調べた上で，既存のプロパティを使用するか，新規にプロパティを定義するかを決める必要がある．しかし，プロパティは様々な組織で公開されており，どのプロパティを用いるのが最適かを判断することが手間になる．

3. 固有表現抽出を用いた LOD 化支援

本研究では，固有表現抽出を用いたリソースの抽出とプロパティの推薦による LOD 化支援を提案する．今回，歴史コンテンツの一つである文化財コンテンツを対象とした．文化財コンテンツの形式は，一ページで一つの文化財を紹介し，ページタイトルが文化財名となっていることを想定する．

3.1 リソースの抽出

固有表現抽出を用いて LOD 化する関連研究として，ソーシャルメディアとマスメディアの文章に対して固有表現抽出を行い，Linked Data 化した田代らの研究がある [1]．田代らは，実世界の出来事を表す情報を「事象」とし，事象間の意味を表現する主題，動作と動作の対象主などを事象属性と定義した．その後，定義した事象属性をソーシャルメディアとマスメディアの文章から抽出した．本研究では田代らの手法を参考に文化財属性を定義し，固有表現抽出を行う．

文化財には,「求福山山車の人形その他付属品」や「旧笹浪家住宅主屋及び土蔵」など長い名称があり，機械学習を用いたとしても正確に抽出できない可能性がある．そこで，本研究では文章に含まれる文化財名の抽出をルールベースで行い，文化財名以外の固有表現は機械学習を用いて抽出する．機械学習の手法には，教師あり学習アルゴリズムである「条件付き確率場」

(Conditional Random Fields) を用いる．文化財名，地名，人名，戦争名，時代，年の六つを文化財属性と定義し、リソースとして抽出する．本研究では，文化財コンテンツの説明文から教師データを作成し，機械学習を用いて文章に含まれるリソースを抽出する抽出器を作成する．

基本的にページタイトルの文化財名を主語，抽出したリソースを目的語として，抽出結果をツリー構造で表示する．ユーザは主語と目的語の関係を自由に編集し，文化財以外の人物や地名などの情報も詳細に表現することができる．例えば,「川田 龍吉男爵（1856〜1951）退職後〜」という文があった場合，1856 と 1951 は文化財名が主語ではなく，川田龍吉男爵である．このような場合はツリー構造の主語と目的語の関係を編集し，川田龍吉男爵を主語，1856 と 1951 を川田龍吉男爵の目的語に変更することで川田龍吉男爵に関係がある年と表せる．

3.2 プロパティの推薦

本研究では，固有表現抽出の結果に付与された文化財属性の組み合わせをもとに，プロパティを推薦する．文章に地名が含まれていた場合，文化財が発見された場所や作られた場所などを表している可能性があるが，文化財属性を地名として抽出しただけでは，実際にどのような関連があるか分からない．そのため，本研究では文化財属性をより詳細に表現する補足情報を定義し，主語の文化財属性と目的語の文化財属性の組み合わせにより目的語の補足情報をユーザに推薦する．主語の文化財属性が「文化財名」で，目的語の文化財属性が「年」の場合，補足情報の「発見された年」と「作られた年」を推薦する．補足情報にプロパティを対応づけ，RDF 化する際に対応づけたプロパティを用いる．本研究では、広く一般的に用いられている schema.org，Dublin Core Metadata Initiative MetadataTerms，Friend of a Friend（FOAF）の三つの語彙と本研究で定義したプロパティを使用する．文化財属性の組み合わせとプロパティの対応関係を表 1 に示す．dc と dcterms は Dublin Core Metadata Initiative Metadata Terms，schema は schema.org，foaf は Friend of a Friend, cul は本研究で定義したプロパティを表す接頭辞である．

表 1: 文化財属性の組み合わせとプロパティの対応表

目的語の文化財属性	主語の文化財属性	補足情報	プロパティ
文化財名	文化財名	発見された場所	cul:discoverPlace
		所在地	schema:address
	-	なし	rdfs:seeAlso
地名	文化財名	発見された場所	cul:discoverPlace
		作られた場所	cul:createPlace
		所在地	schema:address
	地名	地名の旧名	cul:oldNamePlace
	人名	生まれた場所	cul:birthPlace
	戦争名	戦地	cul:warPlace
	-	なし	cul:address
人名	文化財名	製作者	dcterms:creator
		発見者	cul:discoverer
		設計者	cul:designer
		寄贈者	schema:contributor
		使用者	cul:user
	人名	師	cul:master
		弟子	cul:pupil
	-	なし	foaf:name
戦争名	-	なし	cul:war
時代	文化財名	発見された時代	cul:discoverEra
		作られた時代	cul:createEra
	なし	なし	cul:era
年	文化財名	発見された年	cul:discoverPlace
		作られた年	dcterms:date
	人名	生年	foaf:birthday
		没年	schema:deathDate
	戦争名	開戦年	cul:warStartDate
		終戦年	cul:warEndDate
	-	なし	dc:date

4. 抽出精度の評価実験

北海道文化資源データベースで公開されている文化財の説明文に含まれるリソースを抽出する実験を行った．実験では教師データを作成し，CRF++[2] を用いて機械学習を行った．

文化財名を正確に抽出するために，北海道文化資源データベースで公開されている文化財名を形態素解析エンジンである MeCab の辞書に追加した．北海道文化資源データベースの文化財，文化施設 (博物館・郷土資料館・文学館等)，歴史的建造物カテゴリから，無作為に抽出した 670 文の説明文から教師データとテストデータを作成した．説明文を MeCab を用いて形態素に分割し，各形態素に対して IOB2 形式で文化財属性を付加することで，教師データを作成した．

本研究では，5 分割交差検定を用いて適合率，再現率，F 値を算出し，抽出精度の評価を行った．

5. 実験結果・考察

実験の結果から適合率は 88.4%，再現率は 60.5%，F 値は 71.8%となった．年代や時代，戦争名など文章中に多く含まれる語は高い精度で抽出できた．しかし，人名は文章に含まれていても，名字だけが書かれている場合や「フランク・ロイド・ライト」など外国人の名前が書かれていることがあり，抽出に失敗していた．そのため，教師データの数を増やす他に，人名辞書を追加し対応する必要がある．今回，文化財名を正確に抽出するために文化財名を辞書に追加したが，文章中には略称や別称で記述されており，抽出に失敗している文章があった．今後，文化財名の略称と別称を辞書に追加することで対応する．

抽出精度の向上には限界がある．そのため，抽出に失敗した場合の修正作業の支援を行う必要がある．

6. まとめ

本研究では，歴史コンテンツの文章に含まれるデータの LOD 化促進のために，固有表現抽出を用いた LOD 化支援を提案した．文章に含まれるデータを LOD 化する際に課題となる，リソースの抽出とプロパティの選定作業に対して固有表現抽出を用いて支援を行う方法を提案した．文化財の説明文に対して抽出実験を行った結果，抽出精度は 71.8%となった．今後，抽出精度の向上と LOD 作成を支援するシステムの開発を行い，LOD 化支援の方法の有効性を評価する．

参考文献

[1] 田代和浩，王冕，越川兼地ほか：Linked Data を用いたソーシャルメディア×マスメディアの比較実験，人工知能学会全国大会，pp.1-4（2013）.

[2] 工藤拓：CRF++: Yet Another CRF toolkit，入手先〈http://crfpp.googlecode.com/svn/trunk/doc/index.html〉（参照 2015-03-23）.