

固有表現抽出を用いた観光コンテンツからの Linked Data 作成支援

似内 勇太†

奥野 拓†

† 公立はこだて未来大学

1 はじめに

近年、機械可読な形式で Web 上にデータを公開する LOD (Linked Open Data) と呼ばれる仕組みが注目されている。Web 上のデータを LOD 化することで、事物に関するデータや事物同士の関係を機械が処理しやすい形式で公開されるため、アプリ開発などデータの二次利用が期待できる。

本研究では Web 上のデータの LOD 化促進を目指して、非構造化データである文章から Linked Data 作成を支援するシステムを構築する。本手法では、観光コンテンツを対象に、観光情報サイトの文中からリソースの抽出とプロパティの推薦を行い、Linked Data の作成を支援する。

2 Linked Data 化支援ツールの現状

Linked Data は、リソースを表現するための枠組みである RDF を標準形式として Web 上の情報を提供する。Web 上のデータを RDF 形式で記述することにより、事物に関するデータや事物同士の関係を機械可読な形式で表現する。

現在、Linked Data 化する手間を省くために、Link Data や XLWrap など様々なサービスとツールが公開されている。これらは、CSV や XML のような構造化データから RDF データへの変換ができる。しかし、これまで非構造化データである文章から RDF データを作成する支援は行われていない。

3 文章からの Linked Data 化

3.1 文章からの Linked Data 化の利点

Web 上には人名、地名、住所、日付、出来事など様々なデータが公開されている。これらのデータの多くが

文章の形式で公開されている。文章に含まれているデータを Linked Data 化した場合、公開されるデータの範囲が広がり、データの有用性を高めることができる。

3.2 文章からの Linked Data 化の課題

文章を Linked Data 化するために RDF 形式にする場合、主に二つの課題がある。

一つ目は、文章中にリソースとなり得る多様な固有表現が含まれていることである。Web スクレイピング技術により、HTML タグ内の文章を自動で抽出することが出来る。しかし、HTML タグ内の文章に含まれる多様な固有表現を特定し抽出するのは難しい。そのため、RDF データを作成する際にユーザ自身が文章の内容を把握し、リソースを抽出する作業が発生してしまう。

二つ目は、Linked Data 化するために行うプロパティの選定作業が多くなることである。Web 上のデータを Linked Data 化する場合、既存のプロパティを用いることで、データの前処理やアプリケーション内で Linked Data を修正することなく、同じプロパティを使用して Linked Data と連携することができる。そのため、RDF 形式で記述する際に定義されているプロパティを調査した上で、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、文章は非構造化データのため、Web ページごとに出現するプロパティに規則性のあるものは少ない。そのため、Web ページの内容により、プロパティの出現にばらつきがあり、リソース間のプロパティを選定する作業が構造化データと比べて多くなる。

本研究では、自然言語処理技術の一つである固有表現抽出を用いて、上記の 2 つの課題を解決する。

4 関連研究

固有表現抽出を用いた文章からの Linked Data 化の関連研究として、田代らの 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 [1] がある。

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

† Yuta NITANAI and Taku OKUNO, Future University Hakodate

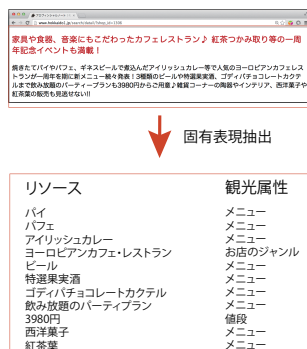


図 1: 観光情報サイトからの抽出例

田代らは、実世界の出来事を表す情報を「事象」とし、Nguyen らのブログからの行動属性を抽出する研究 [2] を参考に、事象間の意味を表現する主題や動作と動作の対象主などを事象属性と定義した。そして、固有表現抽出を用いて、Twitter とニュースサイトの記事から事象を抽出した。その後、抽出した事象と事象属性を元に作成した RDF データから事象ネットワークを作成し、事象間の関係を可視化した。田代らの研究では、ソーシャルメディアとマスメディアから事象が約 74.7% の精度で抽出された。

田代らが作成した RDF データは、プロパティの選定に関しては十分に考慮されておらず、抽出の際に定義した事象属性をプロパティに使用している。そのため、Linked Data としての公開には向いていない。

5 提案手法

本研究では、固有表現抽出を用いて観光情報サイトの文章中のリソースの特定とプロパティを選定する作業を支援する。固有表現抽出を行うために、教師データありの条件付き確率場 (Conditional Random Fields: 以下, CRF) を用いる。観光情報サイトから文章を収集し、収集した文章を教師データの形式へ変換し、機械学習を行い、図 1 のようにリソースと観光属性を抽出する。また、固有表現抽出によって抽出されたリソースに付属する属性情報から、リソースに付与できる既存のプロパティを推薦する。

6 実験と評価方法

本提案手法の評価を行うために、リソース抽出とプロパティ推薦の実験を行う。本研究では、CRF の導入のためにパッケージとして CRF++ を用いる。

今回の実験では、観光情報サイト内の観光地について書かれた文章から教師データを作成する。まず、観光情報サイトである Hokkaido1.jp に対して Web スクレイピングを行い、524 記事内に含まれる文章約 5000 文を収集した。収集した文章に対し、形態素解析エンジンである MeCab を用いて形態素に分割した。その後、収集した観光情報サイトの文章の内容から、リソースとなり得る店名やメニューなどを観光属性として定めた。形態素に分割した単語に対して、定めた観光属性を IOB2 タグの記述法を用いてチャンクタグを記述し、教師データを作成した。

既存のプロパティとして、広く一般的に用いられている schema.org, ダ布林コア・メタデータ・イニシアティブ・メタデータ用語語彙, フレンド・オブ・ア・フレンド語彙からプロパティを収集し、教師データを作成する際に付与した観光属性と結びつけることで推薦する。

リソース抽出とプロパティ推薦の結果について、抽出精度を評価する。観光情報サイトの文章からテストデータを用意し、固有表現抽出を行い、適合率、再現率、F 値を用いて精度の評価を行う。正解判定のために、テストデータの固有表現が既知語と未知語のどちらであるかによって、システムの出力が正しくできたかを考慮し評価する。

7 まとめ

本研究では、Web 上のデータの LOD 化促進を目的として、文章からの Linked Data 作成を支援を行う。ユーザが観光情報サイトの文章を Linked Data 化する際に、CRF を用いて固有表現抽出を行い、文章に含まれるリソースの発見の補助とプロパティの推薦を行うシステムを提案した。今後、提案した抽出システムを実装し、精度の評価と考察を行う。

参考文献

- [1] 田代和浩他: 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, 2013.
- [2] T.M. Nguen, T.Kawamura, Y. Tahara, A. Ohsuga: " Self-supervised capturing of users ' activities from weblogs ", Int. J. of Intelligent Information and Database Systems (IJIIDS), Vol.6, No. 1, pp. 61-76, 2012.