# LOD データセットのメタデータ作成支援

石川　琴巳

公共や民間が保有するデータのうち，誰もが容易に利用できるような形で公開されたデータのことを Open Data という．構造化されたデータを Web 上で相互にリンク付けして，それらを公開できる一連のしくみを提供する実践的方法のことを Linked Data といい，オープンデータに Linked Data の概念を適用したものを Linked Open Data(以下, LOD)と呼ぶ. LOD の表現方法の一つとして RDF がある. LOD データセットなどのオープンデータはインターネット上に複数存在する「データカタログサイト」に登録される.データ提供者はデータカタログサイトに所有するデータセットの登録や公開を行う他，データセットの登録情報の更新等の管理も行う．人手に頼るために，データセットの登録漏れやリンク切れなどの問題が発生した．上記問題を解決するために，クローラーによりデータセットの情報を自動収集することが進められている. 2018 年 9 月, Google Dataset Search ベータ版が公開された. Google Dataset Search は，必須プロパティを含むメタデータが付与されたデータセットを収集する．必須プロパティが付与されていないデータセットは自動収集されず，検索結果に表示されない．そのため本研究ではデータセットに対するメタデータ付与の支援を試みる．支援のためにメタデータの雛形の生成を行い，それを利用することで，データセット提供者らがメタデータを付与する作業の軽減を目指す．本研究ではメタデータの雛形を生成するための処理方式を確立することを研究目的とする. Google Dataset Search における必須プロパティとして，name と description がある. name は「データセットのわかりやすい名前」, description は「データセットの要約文」である. name や description に値を付与する際，どちらも共通してデータセットを表す名詞が含まれる必要がある．そこでデータセットからそのデータセットを表す名詞を抽出する.

名詞の抽出実験では形態素解析器として Mecab を使用する．Mecab は日本語に依存した形態素解析器であるため，実験対象のデータセットは日本語で書かれたものに限定した．実験の流れは以下の通りになる．抽出処理として，データセット中のトリプルのうち Object を対象に，条件に合致した名詞のみを抽出し頻出度を求める．そしてその中で頻出度上位のものを候補とする.

以上の処理評価のために，メタデータが付与済みのデータセットを複数用いて実験する.付与済みのメタデータを形態素解析し名詞を抽出して，本研究の処理結果と比較した．その結果，全名詞における再現率は平均して 36.6%，抽出した名詞における再現率は平均して 18.5%となった．再現率に関して，全名詞の再現率よりも抽出語の再現率の値が低い場合，出現頻度だけでデータセットを表す名詞を抽出することが難しいことがわかった．また再現率が他のデータセットと比較して低いデータセットに関して，データセットの中身を確認したところ，登録されたデータの内容が多岐に渡るものであった．出現頻度だけでデータセットを表す名詞を得られるデータセットもあれば，得られないデータセットもあるため，各データセットの性質に合わせて，名詞の抽出方法を対応させていくことが課題である.

（指導教員　阪口哲男）