クラウドソーシングで収集されたキーワードのオープンデータを利用した分類手法

麻和　昂生

福島県双葉町アーカイブプロジェクトは東日本大震災で被災し、原子力発電所事故を蒙った福島県双葉町の震災当時の避難所やそこに届けられた物品、地域の人々とそのコミュニティの記録を保存するとともに、未来の調査研究および利用活用の促進を目的としたプロジェクトである。収集された写真はデジタルアーカイブ化されており、これらの写真にクラウドソーシングを利用してユーザーに自由にキーワードを付与してもらっている。また、多言語アクセスの重要性が高まってきているという背景から、キーワード入力の際に言語も指定してもらっている。この活動には Crowd4U のマイクロタスクを用いている。現在、これらのキーワードは収集した段階にとどまっており、分類や組織化が為されていない。

本研究ではアーカイブ内のキーワードの中でも、特に物の写真に付与されたキーワードに対して、キーワード間に同義関係や類語関係、上位・下位関係といった関係性を構築する手法を研究する。

本研究ではまず、日本語形態素解析器 juman の代表表記という機能に着目した。これは異なる表記の単語であっても同一概念を示すものには同じ代表表記を与えることでそれらを同一のものとして扱うことができる機能である。アーカイブ内のキーワードを juman で解析し、代表表記を得られたキーワードとそうでないものとを比較・評価した。結果、望ましい代表表記が得られたキーワードは 668 件中 589 件と約 88.2%であった。また、同一の代表表記も多く得られたことから、画像同士を関連付けるためのタグとして活用する方法も考えられた。

キーワード間の関係性を構築するにあたって、日本語版 Wikipedia の構造化情報を問い合わせ可能なデータセットとした DBpedia Japanese を利用した。DBpedia Japanese では SPARQL という RDF 問い合わせ言語を用いることで必要なデータを入手することができる。本実験ではアーカイブ内のキーワードを抽出し、問い合わせに用いることで DBpedia Japanese に収録されたタイトル情報、リダイレクト先情報、カテゴリ情報を抽出した。結果、抽出したリダイレクト先情報によってキーワード間に同義関係を付与することができると考えられた。しかし、抽出できたリダイレクト情報は全 668 件のキーワードに対して 63 件とカバー率は約 9.4%にとどまった。また、抽出したカテゴリ情報を用いることで、同じカテゴリ情報を共有するキーワードの間に類語関係を付与できると考えられる。

今回の実験により、抽出できた情報を活用することでアーカイブ内のキーワード間に関係性を付与できる可能性が示された。一方で、単語や熟語ではなく句や文により表現されたキーワードには関係性を構築するための十分な情報が得られなかったため、これらのキーワードへのアプローチも考えていく必要がある。

（指導教員　阪口哲男）