平成23年度研究チーム活動中間報告（第1回目）

「多言語 Wikipedia の差異情報抽出手法に関する研究」

№120　研究幹事：灘本明代（知能情報学部）

Wikipedia はオンライン上の百科事典の一つであり，誰もがコンテンツを作成できる点や，一つの話題に対し複数の言語版で書かれている点などを特徴として挙げることができる．Wikipedia の各言語版はそれぞれ独立して管理されている場合が多く，各々の言語版を閲覧しているユーザが記事の作成，追記，修正を行っている．例えば，イギリスの伝統的なスポーツである「ローンボウルズ」の英語版の Wikipedia の記事には目次項目も多く，コンテンツが詳細に記述されているが，日本語版の記事は目次項目が少なく，コンテンツ量も少ない．なぜなら，日本語版の記事を作成しているユーザが主に日本人であり，「ローンボウルズ」のことについて十分な知識を持っていない点や，興味がイギリス人と比較して比較的少ない点が原因であると考えられる．一方，日本の伝統的な建物である「平等院」について日本語版と英語版の記事で比較すると，日本語版の記事には目次項目が多く，コンテンツが詳細に記述されているが，英語版の記事は目次項目とコンテンツが少ないといったように，「ローンボウルズ」と逆の現象が起こっている．このように，Wikipedia の記事の言語間で情報の量が異なり，自国の言語版だけでは得られる情報が不足する場合があることがわかる．

一般的にユーザは，母国語版の Wikipedia を閲覧する場合が多い．たとえ母国語以外の言語を読むことが可能であっても，母国語の記事を読んで理解することと比較し時間が掛かり困難な作業である．例えば，ある程度英語を読むことができる日本人ユーザが英語版の「Bowls」を全て読んで理解することは，日本語の記事を読んで理解することと比べて困難である．同様に，日本語をある程度読むことができる英語圏のユーザが日本語版の「平等院」を全て読むことは困難である．そこで我々はユーザが閲覧している母国語版の記事に対して，母国語版では不足している情報を他の言語版から取得し挿入することによって，よりユーザの理解度が上がると考えた．そこで本研究では，多言語版 Wikipedia 間の差異情報を抽出し，提示するシステムを提案する．

これによりユーザは，より有益な情報を容易に取得することが可能となる．

今年度は，想定しているユーザとして英語をある程度理解することができる日本人とし，ユーザの入力した事柄に関する日本語版と英語版の記事をそれぞれ抽出し，記事の内容を比較した上でその差異情報をユーザに提示する手法を提案した．この時，一つの日本語版記事が一つの英語版記事だけに対応するわけではなく，複数の英語版記事が対応する場合がある．例えば，「ローンボウルズ」の場合．日本語版はローンボウルズの世界大会の説明がローンボウルズという記事１つに書いてあるのに対し，英語版ではローンボウルズの記事だけでなくその世界大会の記事が存在し，複数ページまたがっている．そこで我々は，

比較対象となる複数の英語版記事を，記事間のリンクグラフと我々の提案する関連度を用いて抽出した．さらに，抽出された複数の英語版の記事と日本語版の記事を各々の記事の目次構造に基づいて比較し，その差異情報を提示する手法を提案した．研究成果として，国内研究会にて論文を 2 回発表，国内査読付き会議に 1 回発表した．