平成 23 年度研究チーム活動中間報告（第 2 回目）

「多言語 Wikipedia の差異情報抽出手法に関する研究」

№120　研究代表幹事　灘本明代（知能情報学部）

Wikipedia はインターネット上に存在する大規模な百科事典である．その特徴は，ユーザが記事内容を作成，編集することができる点や，2012 年末現在で 284 の多言語版が存在し，情報共有が多様化している点などが挙げられる．このように Wikipedia はユーザにより作成・編集されているため，記載されている情報が十分ではなく，百科事典として内容を詳細する情報量が不足している記事が多く存在する．

一つの特徴である多言語性に注目すると，Wikipedia の各言語版は各々独立で管理されている場合が多く，それぞれの言語版を閲覧しているユーザが記事の編集を行っている．このことによりある話題に対しての記事が言語版によってコンテンツの充実度合が異なり同じ記事であってもある言語版では情報が充実しているが，他の言語版では情報が不足している場合がある．特に文化について書かれている記事では顕著である．

例えば，日本の伝統的な建物である「平等院」の日本語版 Wikipedia の記事であれば目次項目も多く，コンテンツが詳細に記述されている．それに対しの英語版の記事は目次項目が少なく，コンテンツの量も少ない．なぜなら英語版の記事を作成しているユーザが主に英語圏の人であり，「平等院」のことについて英語圏で共有し公開するに値する十分な関連知識を持ったユーザが記事内容を作成，編集に関わっていないことが考えられる．そのために，国や地域特有の事象に関してはユーザが十分な知識を持っておらず，日本語版で提供されているものに匹敵する情報量を追記編集することができないことが多い．このことからユーザが母語で地域などの要因で特異性の高い事象に関しては，内容の充実していない記事を追記，修正することは困難である場合が存在する．

そこで我々は Wikipedia の多言語性に注目し，言語間の差分情報を用いて Wikipedia の記事の情報補完を行う手法を提案する．

具体的には情報が不足している記事と同じ話題である他言語版とを比較し，差分情報を抽出する．そして，その差分情報をユーザが閲覧している言語版の Wikipedia の記事に挿入することにより，不足している情報の補完を行う．この時 Wikipedia は誰もが編集することから，ある言語版の 1 つの記事が別の言語版では複数の記事にまたがる場合が存在する．その為に本研究では，比較対象の記事を Wikipedia のリンク構造を用いて抽出する．ここで，抽出した差分情報を本研究では補完情報と呼ぶ．

昨年までは，日本語版 Wikipedia をユーザの閲覧する記事とし，英語圏の文化に関する補完情報を英語版 Wikipedia から抽出した．これに対し，今年度は英語版 Wikipedia をユー

ザの閲覧する記事とし,日本の文化に関する補完情報を日本語版 Wikipedia から抽出した.
これにより，我々の提案する手法は，日英双方向に有効であることが実証できた.
さらなる課題として，昨年我々が提案した手法ではユーザが閲覧している記事に対し関係のない情報が補完情報として抽出される場合が存在することを確認した.
例えば，映画「となりのトトロ」の場合，我々の手法では比較対象記事として「狭山丘陵」が抽出されている．この狭山丘陵はこの映画の舞台とされている場所である．しかしながら，狭山丘陵の Wikipedia の記事は，ほとんどが狭山丘陵の地理的説明やその開発や生息する野生動物について書かれており，となりのトトロに関する情報は記事のごく一部である．この場合，昨年の我々の提案手法では狭山丘陵の地理的説明やその開発や生息する野生動物がとなりのトトロの補完情報として抽出されているが，これらの情報はとなりのトトロとは関係のない情報であり，となりのトトロの補完情報とは言い難い.
そこで今年度は新たに得られた比較対象記事の分類を行い，その種類ごとに比較対象の領域を決定する手法を提案した.
具体的には①比較記事と同じタイトルを持つ比較基準記事，②比較記事と包含関係になっている包含関係記事，③記事の一部分が比較基準記事と関係する部分一致記事の 3 種類の記事に分類した．そして，その種類毎に閲覧記事と比較する手法を提案した．これにより，精度が昨年の提案の 77%から 85%に向上することができた.
研究成果として，2013 年度に論文が海外 Journal に 1 本，査読付き国際会議にて 4 本，査読付き国内会議にて 1 本，国内研究会にて 1 本，合計 7 本採択され，発表した.
特に難関な国際会議である WISE2012 に採択されたように，本研究成果は国内外から高い評価を得ることができた.