Wikipedia の未執筆記事の執筆支援システムの開発

杉山 喬亮

Wikipedia はフリーの多言語インターネット百科事典である。日本語版 Wikipedia は約 135 万件の記事がウィキペディアン（wikipedian）と呼ばれる編集者の共同作業により執筆及び作成されている。英語版と比べて日本語版で約 520 万件以上の記事が未執筆となっている理由に、それぞれの言語版で執筆者数に差があることが挙げられる。一般的に執筆者数が増えるにつれて多種多様な記事が生成されることから、執筆者の不足により未執筆となっている記事が存在すると考えられる。

本研究では日本語版 Wikipedia の未執筆記事の執筆支援システムの開発を試みる。システムによる見出し文や構成の自動生成といった支援を通して執筆者の負担を減らし、Wikipedia の記事執筆の促進を図る。

執筆支援システムではカテゴリの推測、見出し文の生成、構成名の生成という手法を提案し、それぞれの手法を正解率によって評価した。

カテゴリの推測の結果、上位 10 件までに正解カテゴリを含む正解率が 0.81 となった。見出し文の生成では 52 件のデータセットを用いて「一致」「部分的に一致」「一致しない」の 3 段階の基準により評価した。その結果、「一致」と判定された見出し文が 24 件、「部分的に一致」と判定された見出し文が 8 件、「一致しない」と判定された見出し文が 20 件であった。構成名の生成の評価では生成された構成が意味的に妥当かカテゴリごとに判定し、正解率の平均を測定した。平均正解率は 0.778 という結果が得られた。

各提案手法の評価結果から、執筆支援システムに一定の汎用性を得ることができた。一方で、現在の提案手法は自然言語の自由記述に起因した類語や表記揺れに対処できておらず、正解率に影響が見られた。また、構成名の漏れが考慮されていなかった。これらの問題点に対し、提案手法に word2vec や類義語辞書の作成といった既存の手法を組み合わせることや、構成名の出現率のばらつきによって生成可否を決定する基準値を可変にすることで、システム全体の汎用性を高めることを今後の課題とする。

（指導教員　高久 雅生）