

小説データにおける情景描写を考慮した色抽出手法

佐野 泰成

近年、小説はネット通販や電子書籍として購入されることが多くなった。それと同時に、各購入サイトでは感想投稿や五つ星評価などの選書を支援する機能が利用されるようになった。しかし、感想投稿においては内容の信憑性が確かでないことやネタバレを含むなどの問題があり、五つ星評価においては情報量が少ないこと、評価のついていない小説の内容を判断することが難しいなどの欠点がある。このように、現時点ではインターネット上での小説の選書は不便な点が多く残っている。インターネット上で小説の印象を伝えるための研究もいくつか行われているが、解決すべき課題も多く改善の余地がある。

本研究では、インターネット上での小説の選書をよりわかりやすく、不便さのないものにするを目的に、小説の印象を表す色を抽出する手法を提案する。より具体的には、本手法の目的は (1) 個人の意見に左右されないこと、(2) 文字によるネタバレが起きないこと、(3) 描写対象を絞り具体的でわかりやすい印象を表現すること、(4) 複数種類の語に対して印象抽出することで信頼性を高めることの4つである。これを実現するため、本手法では、小説のテキストデータから情景描写部分を特定し、その部分の印象を最もよく表す色を抽出する。情景描写を対象とする理由は、小説の印象を色として表現するとき、多種多様な描写を総合して抽出するよりも、特定の描写に絞って抽出した印象色のほうが具体的で選書利用者にとって理解が容易であると考えられるためである。特に、情景描写は登場人物が五感を使って感じた描写であり、色から情景描写への想起もしやすいという特徴がある。本手法は、まず入力された小説のテキストデータに対し形態素解析を行った後、SVM (サポートベクターマシン) による学習データを元に情景描写を特定する。次に、単語と色彩を関連させた3種類の色彩データベースを用いて色変換対象語を特定し、色のスコアを色変換対象語の出現回数を考慮して算出する。最後に、2種類の抽出法 (抽出法A, 抽出法B) を使って、テキストデータを入力した小説の情景描写部分における印象を表す色を抽出する。ここで、抽出法Aは色のスコアが最上位の色を抽出色とし、抽出法Bは抽出された色を色相に基づいて13のグループに分け、スコア合計が最上位のグループからスコアが最も大きい色を抽出色とする。

評価実験として、本手法で抽出した色が読者の印象をどの程度反映しているかの検証を行った。その結果、抽出法Aでは読者の情景描写における印象を反映するには、精度にばらつきがあることが分かった。一方、抽出法Bでは安定して読者の情景描写における印象を反映することができることが分かった。

(指導教員 鈴木伸崇)