論文 / 著書情報
Article / Book Information

題目(和文)	オントロジを利用したテキスト計量分析フレームワークの構築
Title(English)	
著者(和文)	川島隆徳
Author(English)	Takanori Kawashima
出典(和文)	学位:博士（工学）, 学位授与機関:東京工業大学, 報告番号:甲第9889号, 授与年月日:2015年3月26日, 学位の種別:課程博士, 審査員:猪原　健弘,桑子　敏雄,赤間　啓之,山元　啓史,戦　暁梅
Citation(English)	Degree:, Conferring organization: Tokyo Institute of Technology, Report number:甲第9889号, Conferred date:2015/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（論文博士）

論　文　要　旨 （和文2000字程度）

報告番号	乙 第　　号	氏　名	川島　隆徳

（ 要　旨 ）

本研究では，単語カテゴリの集合であるオントロジを利用するテキスト計量分析フレームワークを構築し，またフレームワークを実践するためツールを開発した．

テキスト計量分析は，自然言語処理や統計，ネットワーク解析等の手法を用いてテキストを科学的に分析する方法だが，その方法論自体についての研究は充分になされてこなかった．そこで，本研究では既存の研究を踏まえ，単語を単位とする計量分析においては，オントロジが有効であると仮定した．ここで言うオントロジの定義は，同義語・類似語の集合（カテゴリ）に対してその内容を示すラベルが付与された機械可読なデータである．3つのコーパスを対象として3種類のオントロジ導入方法を検討し，それらを用いた分析を行うことでオントロジの有効性を検討した．

1つめのケーススタディであるデジタルゲームの批評コーパスでは，単語の共起情報を用いて機械的にオントロジを生成するアルゴリズムを開発した．2段階に分けたクラスタリングを行うことで，より精度の高いオントロジを生成することが可能となった．また，このオントロジに含まれるカテゴリ単位での分析を行うことで，デジタルゲームの批評においてビジュアル的な要素等に関する記述が減少し，ゲームの遊び方等に関する記述が増加していることを明らかにした．

2つめのケーススタディである河川文化に関する講演会記録のコーパスは，河川を中心とした多彩な内容を含むため共起情報を用いたオントロジ自動抽出は機能しないことが判った．そこで，河川の専門家の手を借りた半手動のオントロジ構築を行い，専門家の知識をもってコーパスを分析する手法を開発した．構築されたオントロジのカテゴリの単位で計量を行い，統計的検定を行うことで，特定の講演において有意に出現するカテゴリ（話題）を抽出した．さらに，そのカテゴリ同士の組（共起）を数え，カテゴリをノード，共起をエッジとしたネットワークを構築することで，曖昧で多様な概念である「河川文化」の全体像を計量的に示すことに成功した．またこのネットワークを分析することで，河川に関するローカルな問題とグローバルな問題は，「川」や「山」などの土地と，「建築」などの工学を介してつながるということが示された．

3つめのケーススタディでは，文学，映画，演劇，デジタルゲームの4ジャンルに関する批評コーパスを扱った．このコーパスでは，ジャンル間の感性の違いを明らかにするため，他の2つのケーススタディでは扱わなかった形容語を対象として分析を行った．形容語についても共起を用いたオントロジ自動抽出は機能せず，また形容語自体は専門的な知識でカテゴライズできる語でも無いため，汎用的なシソーラスである「分類語彙表」の分類をオントロジとして利用した．ジャンル毎にカテゴリの出現を計量し，統計的に比較することで，それぞれのジャンルにおける形容語の特徴的な出現を明らかにした．具体的には，文芸は多用される形容語というものが無く，映画は直感的・感覚的な語の利用が多く，演劇は比喩の感覚を利用した評価が多いという特徴があった．そしてデジタルゲームは他のどのジャンルとも大きく異なり，ゲームを攻略するという意識の下での有利不利の評価や，画面上に現れる詳細な内容

についての評価が多いことが示された.

以上のケーススタディから，テキストの特性と目的によって，適切なオントロジの導入方法が異なることが明らかとなった．また，単語単位ではなく，オントロジのカテゴリを単位とした分析を行うことで，対象テキストの大局的な特徴を抽出可能であることが明らかとなり，オントロジの有効性が示された.

そこで，オントロジの導入を前提としたテキスト計量分析のフレームワークOSQTA(Ontology-based Semantic Quantified Text Analysis)を構築した．このフレームワークは，目的の設定，データの収集から実際の分析までをカバーする．特に，利用するオントロジの選択と分析手法の選択に重点を置き，ケーススタディから得られた知見を元にどのような選択を行うべきかを記述した．その他にも，これまでの研究から得られた経験的な規則等を盛り込んだ．OSQTAはテキスト計量分析の全てを網羅したフレームワーク-では無いが，本研究の3つのケーススタディだけで無く，複数の既存研究がこのフレームワークの範疇として説明可能であり，汎用的なフレームワークと言える.

さらに，フレームワークを実践するためのText Seerというソフトウェアを開発した．Text Seerを利用することで，OSQTAの基礎的な処理を自動的に行うことが可能となった.

OSQTA及びText Seerを利用することで，今後のテキスト計量分析研究を容易に行うことが可能となり，また分析プロセスを標準化することで精度を向上させることが可能となった点が本研究の最大の貢献である.

（論文博士）

論　文　要　旨（英文）　（300語程度）

（Summary）

報告番号	乙 第　　号	氏　名	川島　隆徳

（要　旨）

To establish a methodology for the quantitative text analysis, this research has conducted three analyses on three different corpuses. The focus of analysis is to introduce different kind of ontology, which are multiple sets of word categories in this research. In the first analysis, a new algorithm to generate ontology automatically is used for critiques of digital games. The quantities of category in the generated ontology indicate that critiques contains more references to market situation of games and existing games than existing research expected. Lectures about river culture were target of the second analysis. Because river culture is vast and ambiguous concept, these lectures contains variety of topics. Automatic generation of ontology didn't work for this, so semi-manual construction of ontology by specialist was conducted. With that ontology, co-occurrence of categories in each lectures were extracted, and it formed a network of related topics in whole lectures. This network captures the entire image of river culture which is useful to understand the river culture even for specialists. The purpose of third analysis was obtaining difference of adjective words usage in critiques of four different genres (literature, movie, play, digital game). In this analysis, existing general thesaurus was used as ontology. The quantities of each category showed that each genre has preference to specific kind of adjectives. And digital game was completely different from other media in that aspect. Through three analyses, it is indicated that type of suitable ontology (automatic, manual, existing) varies by characteristics of texts and purposes of analysis. Also, it turned out that the analysis in units of ontology categories is suitable to extract global features of texts compared to in units of word. Based on these result, a framework of quantitative text analysis and analyze software ("Text Seer") was developed.