様　式　C－１９、F－１９－１、Z－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ３０ 年　 ９ 月 １１ 日現在

機関番号：１４３０１
研究種目：若手研究(B)
研究期間：2016～2017
課題番号：１６Ｋ２１１２４

研究課題名（和文）情報抽出技術とLODを用いた地域研究論文の構造化と分析

研究課題名（英文）Information Extraction and Linked Open Data Construction of Area Studies Literature

研究代表者
亀田　尭宙（KAMEDA, Akihiro）

京都大学・東南アジア地域研究研究所・助教

研究者番号：１０７５１９９３

交付決定額（研究期間全体）：（直接経費）　　2,500,000円

研究成果の概要（和文）：1963年から現在に至るまで京都大学東南アジア地域研究研究所から発刊されている地域研究の学術雑誌『東南アジア研究』に掲載されている学術論文や書評について、コンピュータによるテキスト解析を行うためのコーパス化の処理を行った。また、そのコーパス化に用いたツールを公開した。できたコーパスに対し、統計的なテキスト分析や知識抽出を行い、一部Linked Data化も行った。その結果、論文中に現れる国家間の関係、書評で言及されている他の関連書物、文中で記述されている地名や人名を記述・把握することができるようになった。

研究成果の概要（英文）：From 1963 to the present, Center for Southeast Asian Studies, Kyoto University has been publishing academic journal "Tonan-ajia Kenkyu (Southeast Asian Studies)". We make it text based corpus for analysis with computational textual processing. Also, we released the tool used for helping corpus construction. We conducted statistical text analysis and knowledge extraction for the corpus we made, and made linked data. As a result, it became possible to describe and grasp the relation between countries appearing in the paper, other related books mentioned in the book review, place names and person names described in the articles.

研究分野：情報学

キーワード：コーパス作成　情報組織化

１．研究開始当初の背景

情報学分野では論文データのコーパス化(例: https://acl-arc.comp.nus.edu.sg/) やその長期的な分析 (https://www.aclweb.org/portal/content/acl-2012-workshop-rediscovering-50-years-discoveries ）といった取り組みが行われている一方、人文社会系の分野でその取り組みを行うには、電子的な論文の入手からハードルになることが多く、困難な状況にある一方で、人文社会系を中心に多様な手法が学際的に入り混じる地域研究の分野において、京都大学東南アジア地域研究研究所（旧：東南アジア研究所）は 1963 年から 50 年以上の長きにわたって雑誌『東南アジア研究』を発行し、分野の知識流通を支えるとともに、近年はそれを大学のリポジトリにて公開している（ https://repository.kulib.kyoto-u.ac.jp/dsp ace/handle/2433/53669)。そこで、この論文をコンピュータによる分析や柔軟な検索に用いることができるようにテキストコーパス化するとともに、実際に分析を行うことが有意義であると考えた。

２．研究の目的

PDF で提供されている雑誌『東南アジア研究』をテキストコーパス化し、そのデータ化の補助システムやフォーマットに関する開発や策定を行う。また、抽出された知識を RDF で表現して、分析を行い、その活用の可能性を探る。

３．研究の方法

テキストコーパス化を支援する基盤と仕様を作り、作業者を雇用し、作業者のフィードバックを得ながらコーパス化を行った。

その後、得られたコーパスに対して知識抽出のためのアルゴリズムを適用し、論文単位の分析や文単位の分析を行った。文単位のミクロな分析については研究代表者が過去に開発した手法（Akihiro Kameda, Kiyoko Uchiyama, Hideaki Takeda, Akiko Aizawa： Extraction of Semantic Relationships from Academic Papers using Syntactic Patterns, The Fifth International Conference on Information, Process, and Knowledge Management, 2013)をもとに、地域研究ドメインで必要になる工夫を加えながら行った。

４．研究成果

コーパス化補助のシステムと Markdown に基づいたテキストのフォーマットは https://github.com/cm3/dt で公開している。画面イメージは Fig.1 に示す通りである。

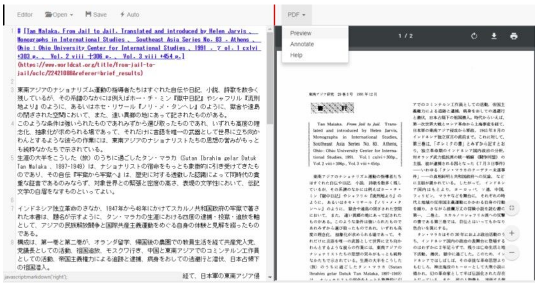

Fig.1 テキストコーパス化支援システム

元 PDF やフォーマットのインストラクションが右のパネルに配置され、左側でテキスト入力や訂正を支援する機能が提供されている。

PDF の一部は用いられた OCR が古かったためか埋め込まれたテキストデータに誤認識が多く、OCR のかけなおしを行った。また、頻度の高い誤認識の修正や、本コーパスでの仕様に則ったスペースやカンマの正規化を自動で行う機能をつけることで作業者の負担を減らした。結果、データ化については、『東南アジア研究』に掲載されている論文のうち、特殊記号の多い言語学の論文など一部を除き、書評も含めて 666 篇をデータ化することができた（うち 485 篇が論文）。

分析では、例えば国家間の関係が論文内での共起関係から見えてきた。旧宗主国との関係近隣諸国との関係が強く、また、日本の論文誌であることから、日本との関係に触れたものも多かった（Fig.2、除く日本）

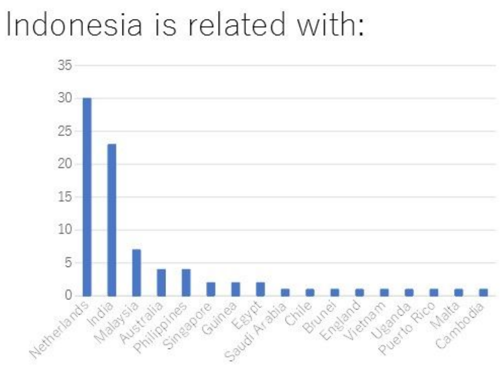

Fig.2　コーパスから見えた国家間関係

また、書評において言及されているほかの書物、文中の地名や人名などの情報に関して、研究所内他プロジェクトと共同で開発した
JSON Editor (Fig.3)を用いて、RDF 形式の情報を付与することで、概念を介した暗黙的な論文間の関係に関しても記述･分析することができた。

Fig.3 JSON Editor の画面

さらに、この RDF 形式のデータの活用として RDF データが指しているリソース同士の関係を元に、ナビゲーションを行うデモを作成した（Fig.4)。例えば、書評記事の中に、東南アジアのナショナリズム運動の指導者に獄中などで書かれた文学が列挙されていたとするとその著者の Wikipedia ページを閲覧している際に、ブラウザの拡張機能を呼び出すことで、そういった文学の著者と著書のリストがオーバーレイで表示され、他の指導者の Wikipedia ページや著書の Worldcat ページに飛ぶことができるようになっている。こういった機能によって、情報探索の終着点としての人間が読むことのみを想定されていた学術雑誌のコンテンツが、人々の日々の知識探索行動の途中で活用されることが期待される。

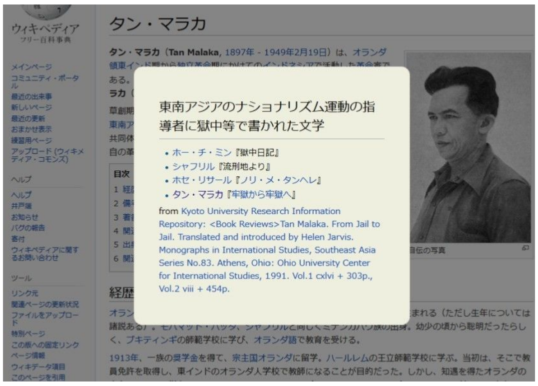

Fig.4 論文知識によるナビゲーション

5．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　0 件）

〔学会発表〕（計　5 件）

1. Akihiro Kameda:"Interactive Knowledge Extraction Tools for Area Studies", The International Workshop on Knowledge Extraction and Semantic Annotation (KESA 2017).
2. 亀田 尭宙:"地域研究における論文と史料からの用語文脈の抽出",第 113 回 CH 研究発表会 (2017)
3. Shoichiro Hara and Akihiro Kameda:"Platform for Humanities Open Data"International Symposium on Grids & Clouds 2017 (ISGC 2017).
4. Akihiro Kameda, Shoichiro Hara: "Constructing Linked Knowledge around Southeast Asian Studies", Digital Humanities 2017.
5. Akihiro Kameda: "Corpus Construction and Analysis of Tonanajia Kenkyu", Association for Asian Studies Annual

Conference 2018 (invited talk in the panel).

〔図書〕（計　0 件）

〔産業財産権〕

○出願状況（計　0 件）名
称：発明者：権利者：種
類：番号：出願年月日：
国内外の別：

○取得状況（計　0 件）名
称：発明者：権利者：種
類：番号：取得年月日：
国内外の別：

〔その他〕
ホームページ等
http://xinchao.cias.kyoto-u.ac.jp/projects/16K21124/
にプログラムやデータに関する情報を掲載している
（一部、雑誌論文の刊行後に追加の予定）

６．研究組織 (1)研究代表者
亀田 尭宙（KAMEDA Akihiro）
京都大学東南アジア地域研究研究所　助教
研究者番号：10751993
(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：

(4)研究協力者
（　　）