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　　６　月　２５　日現在

機関番号：　６２６１５
研究種目：　基盤研究(B)（一般）
研究期間：　2015～2017
課題番号：　１５Ｈ０２７５４

研究課題名（和文）文書閲覧・執筆支援のための遍在的テキストリンケージ

研究課題名（英文）A study on text linkage for document browsing and writing assistance

研究代表者
相澤　彰子（Aizawa, Akiko）

国立情報学研究所・コンテンツ科学研究系・教授

研究者番号：９０２２２４４７

交付決定額（研究期間全体）：（直接経費）　　12,400,000円

研究成果の概要（和文）：本研究では、ユーザの文書理解や文書執筆を支援するための言語処理技術の研究に取り組んだ。まず、テキストから専門用語を抽出して、表記揺れや曖昧性を考慮しながらWikipediaなどの外部の知識源に対応づける専門用語リンキング手法を開発しサーバを実装した。また、文書の意味構造解析や質問応答に関する要素技術の研究に取り組んだ。さらに、学術論文等の文書構造を解析してコーパスを作成して、専門用語リンキングを利用した言語横断検索や文書推薦を実現した。

研究成果の概要（英文）：In this research, we studied text processing techniques to assist users in understanding and writing documents. First, we developed a technique for technical term linking where technical terms are extracted from text and then linked automatically to corresponding external knowledge base entries, such as Wikipedia articles, considering variations of the notations and semantic ambiguity. Next, we worked on methods for document structure analysis and question answering. Then, we analyzed the structure of scientific papers and constructed a text corpus. Finally, we developed a system for cross-lingual retrieval and paper recommendation and showed the usefulness of the proposed technical term linking.

研究分野：　情報学／知能情報学

キーワード：　言語横断情報推薦　エンティティリンキング　専門用語翻訳　意味の分散表現　英語論文執筆支援

## １．研究開始当初の背景

研究者の情報アクセスを支援するために、情報検索の分野では従来から、ユーザの要求に合致する文書を提示するための検索技術の研究が進められてきた。また、電子図書館の分野では、著者や引用ネットワーク等に基づく情報推薦の研究が盛んに行われている。さらに科学計量学においては、最先端のマイニング技術を駆使したリサーチフロント分析や技術マップの自動生成などの試みが進行中である。

これらの既存技術は、膨大な情報の中から、研究に役立つ資源（論文など）を効率よく捜して入手するためのものといえる。一方で、利用者が入手した情報をいかに効率よく読み解き活用するか、という切り口での検討は必ずしも十分ではなかった。

## ２．研究の目的

本研究では、特定の学術分野に特化した内容を含むなど、利用者が理解に支援を必要とするテキストを想定して、ユーザの読み書きを支援するために必要となる要素技術の検討に取り組んだ。具体的には、テキストの任意の箇所に、関連する他のテキストや文書を対応づけて推薦するための遍在的な意味インデクシング法の実現に向けて、以下の３つの研究課題を設定した。

(1) 与えられたテキストから分野に特化した専門用語を自動抽出して、外部知識サーバに対応づけるための専門用語リンキングシステムの構築

(2) 文書構造の解析と、文書を構成する語や文に対する意味処理手法の研究

(3) 学術論文に解析手法を適用するためのツールや資源の整備と、デモシステムを通した有効性の検証

## ３．研究の方法

研究期間内では、特に情報学分野を中心とする学術論文を対象として、専門用語リンキングを利用した言語横断検索や文書推薦システムの実現を目指した。

### (1) 専門用語リンキングサーバ

まず、専門用語を対象とした辞書データベースの構築に取り組み、表記揺れ、対訳、文書中の用例へのリンク、あいまい性解消のため文脈、コーパス中での頻度などの情報を収集して、網羅性が高い日英対訳専門用語データベースを構築した。

次に、これらの言語資源を活用して、テキストから専門用語を抽出して、表記揺れやあいまい性を考慮しながら、Wikipedia などの外部の知識源に対応づける専門用語リンキングサーバを実現した。

さらに、専門用語リンキングの評価用データセットを構築した。このためにまず、文単位での和英の対応がとれた抄録を抽出して、対訳関係にある専門用語を人手によりアノテーションした。また、人手で正解データを作成するためのウェブアプリケーションを実装して、専門用語の抽出、翻訳、Wikipedia 記事へのリンキングの評価・分析が可能なデータセットを作成した。

### (2) テキストの意味構造と検索技術

論文コーパスから語やセクション、文書の分散表現を獲得して、関連情報の検索や、語義あいまい性解消に用いた。さらに、係り受け解析の結果や談話構造を、検索のランキングに反映させるシステムを構築した。

また、意味構造解析と情報の検索・提示に関する要素技術として、外部知識データベースに登録された用語の意味関係を分散表現に反映させる手法の有効性を検証した。さらに、任意の文を簡潔に提示するための文圧縮手法を検討するとともに、計算機による文章理解と質問応答に関する体系的な分析を行った。

### (3) 実文書への適用とシステム実装

自然言語処理分野の国際会議を網羅する ACL_Anthology 上の論文を対象に、PDF 構造解析に基づき得られたテキストをクリーニングして、自然言語文コーパスを整備した。抽出した各自然言語文には、文書 ID やセクションラベルを付与して、異なるレベルの類似度を結合して検索できるデータベースを構築した。

## ４．研究成果

### (1) 専門用語リンキングサーバ

まず、本研究で構築した評価用データセットに基づく予備的な分析を行い、これに基づき、「表記揺れへの対応」、「語構造の解析と共参照関係の解析」、「トピックモデルに基づくあいまい性解消」、「未知の用語に対する翻訳」の各機能を提案・実装した。

表記揺れへの対応：　収集した専門的なキーワードについて、対訳関係等を利用して表記

揺れのルールを抽出した。これを用いて、論文のテキストから抽出した専門用語の表記を正規化する機能を実装した。

語構造の解析と共参照関係の解析：　既存のエンティティリンキングの多くは固有表現を対象としているためリンク先の項目について階層性を考慮する必要がないが、専門用語は概念体系の中に位置づけられるものであることから、リンキングにおいて抽象的な概念に対応する必要があることが判明した。そこで、収集した専門用語データベースを利用して、専門用語の語構造を統計的に解析する手法を確立し、共参照関係の抽出に活用した。

トピックモデルに基づくあいまい性解消：専門用語リンキングで必要となるあいまい性解消のための文脈情報をトピックモデルの形で獲得・蓄積した。トピックモデルを和英が対応づけられた論文抄録から作成することで、トピック空間上での言語横断的な類似度計算が可能になった。

未知の用語に対する翻訳：　著者キーワード等から専門用語の対訳コーパスを構築し、機械翻訳手法を適用することで、未知の専門用語に対しても適切な翻訳候補が得られるようにした。これによって、言語横断的な見出し語の対応付けや、それに基づく言語横断検索が可能になった。

構築した専門用語対訳辞書や翻訳ツールを利用することで、日英・英日など言語が異なる知識ベースの参照を実現し、言語横断的な情報推薦が可能であることを示した（図１）。


図 1　専門用語リンキングサーバ画面

(2) テキストの意味構造と検索技術

WordNet の同義関係をニューラルネットワークの構造に取り込んだ分散表現の獲得について既存手法を拡張し、上位下位関係など任意の関係を考慮する手法を実装して有効性を評価した。また、明示的および暗黙的な談話標識への意味ラベル付与の手法を検討し、談話構造自体の分散表現の獲得を試みた。

長文をわかりやすく提示するための文圧縮手法に取り組んだ（図２）。また、文章の「意味」を構成する要素の分析に取り組み、自然言語処理の文章理解タスクを対象として、読解スキルの分類やタスクの比較を行った。さらに、深層学習により得られる分散表現を利用して文の類似度を求める手法の開発に取り組み、論文中に出現する文の重要度計算に適用して有効性を評価した。

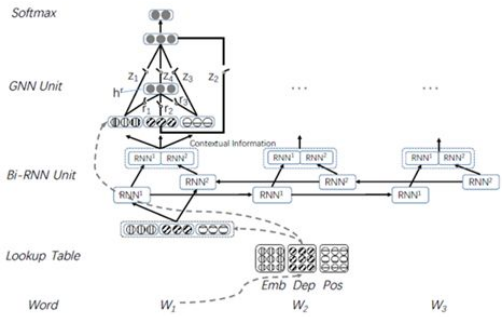


図 2 分散表現を用いた文圧縮

(3) 実文書への適用とシステム実装

与えられた任意のテキストに対して、専門用語を抽出し、対応する Wikipedia の記事に対応付ける仕組みを API として実現した。また、適用対象として情報学分野の論文に焦点をあてて、数百万件の抄録、数万件の全文テキストを用いたプロトタイプ版を構築して有効性を確認した。

以上のように本研究では、ユーザの文書理解や文書執筆を支援するための言語処理技術の研究に取り組んだ。まず、テキストから専門用語を抽出して、表記揺れやあいまい性を考慮しながらWikipedia などの外部の知識源に対応づける専門用語リンキング手法を開発しサーバを実装した。また、文書の意味構造解析や質問応答に関する要素技術の研究に取り組んだ。さらに、学術論文等の文書構造を解析してコーパスを作成して、専門用語リンキングを利用した言語横断検索や文書推薦を実現した。

## ５．主な発表論文等

〔学会発表〕（計　14　件）


① Akiko Aizawa: “Bridging the gap between PDF and natural language text.” Biomedical Linked Annotation Hackathon 4 (BLAH-4). (2018)

Database Center for Life Science, Tokyo, Japan. **招待講演**

②Saku Sugawara, Yusuke Kido, Hikaru Yokono, Akiko Aizawa: "Evaluation Metrics for Machine Reading Comprehension: Prerequisite Skills and Readability." The 55th annual meeting of the Association for Computational Linguistics (ACL 2017), pp.806-817. (2017) Vancouver, Canada.**査読有**

③Yang Zhao, Akiko Aizawa: "A Gated Neural Network for Sentence Compression using Linguistic Knowledge." The 22nd International Conference on Natural Language & Information Systems (NLDB 2017). (2017) Liège, Belgium **査読有**

④**岩月憲一, 相澤彰子: "英語論文の執筆を支援する定型表現集のカテゴリ構造の分析." 言語処理学会第23回年次大会** (NLP2017). (2017) **筑波大学(茨城)**

⑤Saku Sugawara, Hikaru Yokono, Akiko Aizawa: "Prerequisite Skills for Reading Comprehension: Multi- perspective Analysis of MCTest Datasets and Systems." 31st AAAI Conference on Artificial Intelligence (AAAI-17). (2017) San Francisco, California, USA **査読有**

⑥Thomas Perianin, Hajime Senuma, Akiko Aizawa: "Exploiting Synonymy and Hypernymy to Learn Efficient Meaning Representations." 18th International Conference on Asia- Pacific Digital Libraries (ICADL 2016). (2016) Tsukuba, Japan. **査読有**

⑦Saku Sugawara, Akiko Aizawa: "An Analysis of Prerequisite Skills for Reading Comprehension." Uphill Battles in Language Processing Scaling Early Achievements to Robust Methods, Workshop held in conjunction with EMNLP 2016. (2016) Austin, Texas, USA **査読有**

⑧Yusuke Kido, Akiko Aizawa: "Discourse Relation Sense Classification with Two-Step Classifiers". CoNLL 2016 (the SIGNLL Conference on Computational Natural Language Learning), Shared Task "Multilingual Shallow Discourse Parsing", co-located with ACL 2016. (2016) Berlin, Germany **査読有**

⑨**菅原朔, 横野光, 相澤彰子: "自然言語理解ユニットテストと意味表現の検討." 人工知能学会全国大会 (第30回). (2016) 北九州国際会議場 (北九州市)**

⑩**服部一浩, 横野光, 相澤彰子: "極小言語戦略による文テンプレート獲得." 言語処理学会第22回年次大会** (NLP2016) (2016) **東北大学(宮城)**

⑪Paul Willot, Kazuhiro Hattori, Akiko Aizawa: "Extracting Structure from Scientific Abstracts." 17th Asian Digital Library Conference (ICADL 2015). (2015) Seoul, Korea **査読有**

⑫Christopher Norman, Akiko Aizawa: "Technical Term and Keyphrase Extraction Using Measures of Neology." *Keyphrase*–Novel Computational Approaches to Keyphrase Extraction, Workshop in ACL-IJCNLP 2015. (2015) Beijing, China **査読有**

⑬**橋本捷人, 相澤彰子: "ベクトル空間モデルを用いた英文コロケーション誤り訂正." 第222回自然言語処理研究会**. (2015) **首都大学東京秋葉原サテライトキャンパス (東京)**

⑭**服部一浩, 横野光, 相澤彰子: "文書分類のためのフレーズパターンの生成." 2015年度人工知能学会全国大会 (第29回)**. (2015) **公立はこだて未来大学 (北海道)**

**6．研究組織**

**(1)研究代表者**

**相澤 彰子** (AIZAWA, Akiko)

**国立情報学研究所・コンテンツ科学研究系・教授**

**研究者番号：90222447**