科学研究費助成事業　　研究成果報告書

平成 ２６ 年　 ６ 月　 ５ 日現在

機関番号：６２６１５
研究種目：基盤研究(B)
研究期間：2010～2013
課題番号：２２３０００３１

研究課題名（和文）情報検索のためのテキスト間関係認識に関する研究

研究課題名（英文）Textual relation recognition for information retrieval

研究代表者
宮尾　祐介（Miyao, Yusuke）

国立情報学研究所・コンテンツ科学研究系・准教授

研究者番号：００３４３０９６

交付決定額（研究期間全体）：（直接経費）　　13,400,000円、（間接経費）　　4,020,000円

研究成果の概要（和文）：本研究では、情報科学論文の高度な検索というターゲットアプリケーションを想定し、検索クエリと検索結果との関係を分類して提示する手法の開発を行った。情報科学論文のアブストラクトを分析し、そこに一般的に現れるテキスト間関係として、手段、目的、結果、などを定義し、実際に100件のアブストラクトに対してこれらの関係を付与したデータを作成した。このデータを学習・評価データとして自動認識手法の開発を行い、実際に検索に利用できるレベルの精度を達成した。さらに、この手法を用いて情報科学論文3000件に自動的に関係を付与し、これを利用して検索結果の分類を自動的に行うプロトタイプシステムを開発した。

研究成果の概要（英文）：This research aims to develop a method for automatically classifying relationships between a search query and its search results, with an intelligent search engine of information science papers as a target application. By analyzing actual paper abstracts, we identified textual relations that appear in general in those papers, such as method, purpose, and result, and developed an annotated data set that consists of 100 paper abstracts. By using this data set for training/evaluation, we developed a method for automatically recognizing these relations, and achieved a practical level of accuracy. We applied this method to 3000 abstracts of information science papers, and develpoed a prototype search engine that automatically classifies search results based on these relationships.

研究分野：総合領域

科研費の分科・細目：情報学・メディア情報学・データベース

キーワード：情報検索　自然言語処理　テキスト間関係認識　含意関係認識

１．研究開始当初の背景

近年、構文解析や機械学習などの自然言語処理技術が飛躍的に発展したことにより、自然言語処理を応用した様々なアプリケーションが実用化されている。例えば、テキストから自動的に情報を獲得する情報抽出においては、構文解析や機械学習を適用することで精度が大幅に向上することが示されている。情報抽出は 1990 年代より研究が進められ、初期の頃はニュース記事から企業活動や人事異動を自動抽出する問題、その後は生命科学論文からタンパク質間相互作用を自動抽出する問題や、ユーザごとに必要な関係をウェブから取得するオンデマンド情報抽出など、様々な情報要求に適用されている。現在では 70%～90%以上の精度を達成しており、実用アプリケーションで広く利用されている。

一方、ウェブ検索に代表される情報検索においては、自然言語処理技術が有効活用されているとは言い難い。主な原因として、一般的な情報検索では検索クエリと「何らかの関係があるテキスト」を探すことを目的としており、多くの場合、文が表す意味を計算する必要がないということが考えられる。つまり、テキストがどのような意味を表しているかを計算せず、何に関して書かれているかを計算できればよいので、キーワードに基づく検索技術で十分である。

これまでの研究では、タンパク質間相互作用の自動抽出においては、自然言語処理技術を応用することで精度が飛躍的に向上することを示し、一定の成功を収めた。一方、論文検索においては、情報要求のタイプによって精度が大幅に向上する場合と、逆に大幅に悪化する場合があることを観察した。例えば、「特定のタンパク質による特定の反応に関する論文」を検索する場合は、文の意味に基づく検索手法が有効であった。一方、「あるタンパク質について書かれている論文」を広く集めたい場合には、キーワードに基づく検索手法が適しており、文の意味に基づく検索手法ではかえって精度が悪化した。すなわち、一般的な情報検索の問題設定においては様々な情報要求のタイプが混在しており、これらの情報要求について区別せずに研究・評価を行うと、キーワードに基づく検索手法より先の技術革新が望めない。

２．研究の目的

本研究では、検索クエリと検索結果との関係を分類して提示するという方針で、情報検索の高度化を目指す。上述のように現在一般的な情報検索では検索結果は「何らかの関係があるテキスト」であり、そこから実際に必要な情報を得る部分は個々のユーザに任されている。ここで、「何らかの関係」を細分類して提示することを考える。例えば、「あるタンパク質について書かれている論文」を検索した場合は、タンパク質が引き起こす生体反応について書かれている文献もあれば、そのタンパク質を病気の治療に利用することについての論文も存在する。このように、検索クエリと検索結果とのテキスト間関係を認識する問題を設定し、その解決手法について研究を行うこととした。

このようにテキスト間関係を認識するタスクを設定することにより、文が表す意味を利用した情報検索の高度化が期待できる。しかし、認識すべき関係の種類は、検索対象文書の性質や検索システムの利用目的によって異なる。そこで本研究では、論文検索をターゲットアプリケーションとして想定し、以下の３つの点について研究を行うことで、テキスト間関係を自動認識する技術を開発することとした。

(1) テキスト間関係の種類の定義：例えば論文のサーベイを考えると、検索すべき文献はいくつかの種類に分かれると考えられる。具体的には、（肯定的または否定的な）先行研究、応用研究、直接の関係はないが関連する研究、などに分けられる。さらに、対象分野にとって特別な意味がある関係も考える必要がある。したがって、論文検索というアプリケーションを設定したときにどのような関係を認識すべきか検討し、定義する。

(2) テキストの同値性・含意関係の認識：テキスト間関係を認識する前に、2 つのテキストの間に関係が存在するか否かの判定が必要となる。つまり、検索クエリと検索結果テキストのどの部分が対応しているのかを認識する必要がある。クエリがキーワードのみの場合は比較的簡単であるが、前述のように「特定のタンパク質による特定の反応」について検索する場合など、クエリがフレーズや文の場合は自明でない計算が必要である。この問題は近年の自然言語処理研究において含意関係認識という問題として研究が行われているが、未だ発展途上の研究テーマであるため本研究で高精度化を行う。

(3) テキスト間関係の認識：検索クエリと検索結果テキストとの対応を認識した後、クエリに対応する部分テキストが検索結果テキスト全体の中でどのように記述されているかを計算し、テキスト間の関係を認識する。

３．研究の方法

本研究では、研究目的で述べたように、(1) 論文検索において有用なテキスト間関係の定義、(2) テキストの同値性・含意関係を認識する技術の開発、(3) テキスト間関係を認識する技術の開発、の３点について研究を行った。また、最終的にこれらの技術を実装し、論文検索のプロトタイプシステムを構築す

る計画を立てた。以上の各点について、具体的な研究方法を以下に述べる。

(1) 論文検索システムを題材としたテキスト間関係の分析と定義

テキスト間関係認識に関する先行研究ではテキスト間に一般的に存在する論理的関係に着目しているが、本応募研究では情報検索の応用に特化したテキスト間関係を想定している。したがって、検索対象文書や検索システムの目的によって認識すべきテキスト間関係が異なる。そこで、本研究では論文検索をターゲットアプリケーションとして設定し、テキスト間関係認識の研究を行った。具体的には、論文検索結果の分類をシミュレートし、論文検索で必要となるテキスト間関係を整理した。さらに、論文の引用関係の種類についての既存研究の調査を行った。

この研究に基づき、論文検索システムにおいて有用なテキスト間関係の種類を決定した。この時、論文検索において一般的な関係（先行研究、応用研究、など）と、対象文書の特徴に依存する関係（例えば生命科学論文が対象であれば「生体反応に関するもの」や「病気治療に関するもの」など）とに区別することに留意した。

(2) テキストの同値性・含意関係の認識に関する研究の調査、実装

テキスト間関係を認識する前段階として、テキスト間の対応関係を認識する必要がある。この問題はテキスト間含意関係を認識する問題として近年の自然言語処理においてさかんに研究されているため、最近の研究成果について調査を行い本研究の目的に合致するものがあるか検討した。特に高精度が期待できる手法や本研究の目的（検索クエリと検索結果との対応関係を認識すること）に合致する手法については、既存システムの利用および既存手法の実装を検討した。さらに、含意関係認識手法を実装するために利用できるリソースとして同義表現認識や含意関係認識のコンペティションで利用されているデータについて調査を行い、本研究に利用可能かどうか検討を行った。既存のリソースが本研究の目的に不十分である場合には、既存リソースの拡張・改良を行うか、あるいは新たにリソースを作成することを検討した。

(3) テキスト間関係認識手法の開発、実装

本研究の主要目的であるテキスト間関係認識手法について研究を行った。高精度な関係認識を狙うために教師あり機械学習を用いた手法を優先的に考えるが、正解付きデータの作成が困難な場合や特徴量が多様すぎて教師あり学習になじまない場合は、ブートストラップ的手法やクラスタリング手法を検討した。テキスト間関係認識器の学習や評価に用いるための正解付きデータが必要であるが、これは本研究において新たに作成した。

(4)テキスト間関係認識を利用した論文検索システムの構築

テキスト間関係認識手法を実際に大規模な論文データベースに適用し、論文検索システムのプロトタイプシステムを作成した。検索結果表示のためのウェブインターフェースを作成し、自動認識したテキスト間関係をユーザが利用しやすい形式で提示する方法を検討した。

４．研究成果

(1) 論文検索システムを題材としたテキスト間関係の分析と定義

論文検索に関する既存研究では、手法とその効果を自動抽出し技術動向を俯瞰する手法など、情報抽出技術を応用することによって論文の内容を横断的に提示する手法が提案されている。特に、獲得する情報をより複雑化することで、様々な観点から横断的分析を提供するという研究の方向性が見られ、これらの手法を応用することで論文間の関係・差異を提示することができると考えられる。

そこで、情報科学論文のアブストラクト 30 件を対象としてアノテーション作業を行い、実際に論文中に現れるテキスト間関係の分析を行った。その結果、情報科学論文においては、生命科学分野におけるような確立された関係 (タンパク質間相互作用など)を予め想定することができないことが明らかとなった。これは、情報科学では現実世界のあらゆる問題が研究のターゲットとなり得ること、また物理的な現象ではなく研究のアイディアや手法、目的が論文の主眼になること、といったことが原因と考えられる。この分析結果に基づき、目的、結果、入力、出力、といった、情報科学論文に一般的に現れるテキスト間関係を形式化するアノテーションスキーマを構築した。分析に用いた 30 件のアブストラクトにおいては、このアノテーションスキーマにより、論文アブストラクトで表現されているテキスト間関係をほぼ網羅的に形式化できることが示された。

ここで確立したテキスト間関係アノテーションスキーマに基づき、テキスト間関係認識手法の学習・評価データとして、情報科学分野の学術論文のアブストラクト 100 件に対し、テキスト間関係のアノテーションを行った。このアノテーション作業の過程で表出したガイドラインの問題については随時検討を行い、ガイドラインの修正を行った。

最終的には、このアノテーションスキーマ・ガイドラインを用いて第 3 者によるアノテーション作業を行い、アノテーション作業の一貫性評価を行った。その結果、このガイドラインによって高品質なアノテーションを行うことができることが示された。このデ

ータは、現在リクエストベースで公開している。

(2) テキストの同値性・含意関係の認識に関する研究の調査、実装

まず、RTE チャレンジ（テキストペアを入力とし、それらの間に含意関係があるかどうかを判定するタスク）を中心として既存研究の調査を行った。既存研究では、２つのテキストの間の単語・句レベルの対応関係を構造的機械学習で学習するものが主流である。現在のタスク設定・データではこのような手法である程度の精度が達成できるものと思われる。また、依存構造木を書き換え規則で対象テキストに近づけていく手法や、入力テキストを論理表現に変換して自動証明器を適用する手法が提案されている。これらの手法は既存のデータにおいては前述の手法と大きな差はないが、入力テキスト間の距離が大きい場合には有効な手法であると期待される。

日本語においては標準的な評価データが存在しないため、含意関係認識の評価タスクを企画し、標準データセットの構築を行った。本データは、含意関係認識の研究のために公開されており、国内外の多くの研究機関が利用している。また、このデータを評価データとして利用し、形式論理による推論と統計的なパラフレーズ認識を組み合わせた含意関係認識システムの開発を行った。また、本システムの基盤技術として構文解析の高精度化および分野適応手法を開発した。

(3) テキスト間関係認識手法の開発、実装

研究項目(1)で開発したテキスト間関係アノテーションの学習・評価データを用いて、テキスト間関係認識手法の開発を行った。まず、予備実験として既存の情報抽出・関係認識手法をそのまま適用したが、一般的な情報抽出の精度より認識精度が非常に低いという結果が得られた。これは、本タスクが既存の情報抽出タスクよりも困難なものであることを示唆している。また、アノテーションスキーマが複雑なため学習データを大規模化しにくいという問題もあり、学習データが小さいために精度が低いという問題も考えられる。

そこで、外部リソース（ウェブから自動獲得した類義表現辞書）や半教師あり学習を利用する手法について実験を行い、認識精度が向上することを確認した。今のところ既存の情報抽出タスクよりも精度が低い状態ではあるが、論文検索システムで利用するには十分な精度が得られた。

(4)テキスト間関係認識を利用した論文検索システムの構築

上述の手法で自動認識したテキスト間関係を利用した論文検索システムのプロトタイプを開発した。単なるキーワード検索ではなく、そのキーワードが論文で果たす役割（手段、入力、結果、評価など）に基づいて検索結果を分類することができる。ここでは、研究項目(3)で開発した技術を応用し、情報科学分野の学術論文のアブストラクト約 3000 件に対してテキスト間関係を自動認識した結果を検索対象データとして用いている。

下図に本システムのスクリーンショットを示す。ここでは、「SVM」というクエリに対して、それを利用した研究（例：「SVM を用いた形態素解析手法」）、SVM を高度化するための手法（例：「SVM のための次元圧縮手法」）、というように、SVM がその論文中で果たす役割によって検索結果を分類することができる。

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 10 件）

(1) Yuka Tateisi, Yo Shidahara, Yusuke Miyao and Akiko Aizawa. 2014. Annotation of Computer Science Papers for Semantic Relation Extraction. Proceedings of LREC2014. 査読有

(2) Yuka Tateisi, Yo Shidahara, Yusuke Miyao and Akiko Aizawa. 2013. Relation Annotation for Understanding Research Papers. Proceedings of LAW-7. 査読有

(3) Yotaro Watanabe, Yusuke Miyao, Junta Mizuno, Tomohide Shibata, Hiroshi Kanayama, Cheng-Wei Lee, Chuan-Jie Lin, Shuming Shi, Teruko Mitamura, Noriko Kando, Hideki Shima, Kohichi Takeda. 2013. Overview of the Recognizing Inference in Text (RITE-2) at NTCIR-10. Proceedings of NTCIR-10. 査読無

(4) Alvin Grissom II, Yusuke Miyao. 2012. Annotating Factive Verbs. Proceedings of LREC 2012. 査読有

(5) Shun'ya Iwasawa, Hiroki Hanaoka, Takuya Matsuzaki, Yusuke Miyao, 2011. Jun'ichi Tsujii. A Collaborative Annotation between Human Annotators and a Statistical Parser. Proceedings of the Linguistic Annotation Workshop. pp.56-64. 査読有

(6) Tadayoshi Hara, Yuka Tateisi, Jin-Dong Kim, Yusuke Miyao. 2011. Parsing Natural Language Queries for Life Science Knowledge. Proceedings of BioNLP 2011. pp. 164-173. 査読有
(7) Hideki Shima, Hiroshi Kanayama, Cheng-Wei Lee, Chuan-Jie Lin, Teruko Mitamura, Yusuke Miyao, Shuming Shi and Koichi Takeda. 2011. Overview of NTCIR-9 RITE: Recognizing Inference in TExt. Proceedings of NTCIR-9. pp.291-301. 査読無
(8) Tadayoshi Hara, Yusuke Miyao and Jun'ichi Tsujii. 2010. Evaluating the Impact of Re-training a Lexical Disambiguation Model on Domain Adaptation of an HPSG Parser. Trends in Parsing Technology: Dependency Parsing, Domain Adaptation, and Deep Parsing. pp. 253-272. 査読有
(9) Yusuke Miyao, Alastair Butler, Kei Yoshimoto, and Jun'ichi Tsujii. 2010. A Modular Architecture for the Wide-Coverage Translation of Natural Language Texts into Predicate Logic Formulas. Proceedings of PACLIC 24. pp. 481-488. 査読有
(10) Makoto Miwa, Yusuke Miyao, Rune Saetre, and Jun'ichi Tsujii. 2010. Entity-Focused Sentence Simplification for Relation Extraction. Proceedings of COLING 2010. pp. 788-796. 査読有

〔学会発表〕(計 4 件)
(1) 建石由佳, 宮尾祐介, 相澤彰子. 2014年3月18日. 北海道大学. 情報科学論文のための意味関係検索システム. 言語処理学会第20回年次大会.
(2) 建石由佳, 仕田原容, 宮尾祐介, 相澤彰子. 2013年3月15日. 名古屋大学. 情報科学論文からの意味関係抽出に向けたタグ付けスキーマ. 言語処理学会第 19 回年次大会.
(3) 宮尾 祐介. 2012年11月22日. 横浜. 学術論文の高度な検索へ向けて－論文の意味内容を解析する試み. 図書館総合展.
(4) 内山清子, 相澤彰子, 高須敦宏, 難波英嗣, 宮尾 祐介. 2011年6月1日. 盛岡. オススメ論文検索システム: OSUSUME. 人工知能学会第25回全国大会.

〔図書〕(計 0 件)

〔産業財産権〕
○出願状況(計 0 件)

○取得状況(計 0 件)

〔その他〕
ホームページ等
http://kmcs.nii.ac.jp/mylab/

６．研究組織
(1)研究代表者
宮尾 祐介 (MIYAO、 Yusuke)
国立情報学研究所・コンテンツ科学研究系・准教授

研究者番号：00343096

(2)研究分担者

なし

(3)連携研究者

なし