プロジェクト名：　メタ知識構造の言語的・統計的モデリング手法の研究

プロジェクトディレクター：　宮尾 祐介　准教授（国立情報学研究所）

〔1〕 研究プロジェクト

(1) 目的・目標

本研究では、因果関係、理由、目的といった、知識を構成するための普遍的関係を定式化し、これに基づき構造化された知識をテキストデータから自動抽出する手法を開発する。人間は、現実世界の現象を何らかの方法で認識・理解し、その結果を上述のような関係を用いて知識化するが、その知識は現在のところテキストデータという形でしか客観化されない。このような人間の知識が持つ構造、すなわちメタ知識構造に着目し、それを計算可能なモデルとして定式化すること、さらにはテキストデータから実際に構造化された知識を抽出する技術を確立することを目指す。このとき、メタ知識構造は必ずしもテキストの中で明示的に言語化されないため、言語的手がかりと統計的特徴の両面を利用する手法が必要となる。

(2) 必要性・重要性（緊急性）

データ中心科学では、データを解析・検索・可視化する技術に加えて、そこから知識を抽出するプロセスが本質的であるが､後者は今のところ研究者やデータアナリストなどの人間に依存せざるを得ない。抽出された知識は学術論文や特許文書などのテキストデータとして蓄積されるが、それ自体をデータとして利用する技術は今のところ無く､効率的な研究開発サイクルのボトルネックとなっている。つまり、解析すべき一次データは爆発的に増加しているものの、そこから生まれる知識の評価や活用のスピードは上がらないという状況に陥っている。本研究は、データ解析の結果得られた知識を効率的に評価・再利用するための基盤的支援技術を提供するものであり、データ中心科学を推進するためのボトルネックを解決するために必須なものと言える。

(3) 期待される成果等（学問的効果、社会的効果、改善効果等）

現在、社会活動の様々な成果がテキストデータという形で蓄積されている。例えば、学術分野では、多様な研究の成果が学術論文や特許として公開されている。しかし、これらの成果や波及効果を客観的かつ定量的に把握・活用する方法は今のところ無い。

本研究の応用先は、企業活動、ヘルスケア、研究開発、政策決定など様々な場面における知識の評価・活用やそれを用いた意思決定に渡る。無論、新しいアイディアを生み出す創造性や、想定外の問題を柔軟に解決する能力は本研究でカバーできるものではない。しかし、基本的な情報分析や複雑な情報下での知識の把握は本研究の技術によりサポートされるため、人間はより創造的な知的活動にフォーカスすることができる。

(4) 独創性・新規性等

データマイニングなど統計的分析を用いるフレームワークは、相関関係や共起関係を可視化するものであり、そこから因果関係、理由、目的などを認識するのは研究者やデータアナリストの役割である。このプロセスを自動化することは困難であるが、本研究は人間が抽出した知識を再利用可能な形でモデル化することを目的としており、今までのデータ解析フレームワークとは本質的に異なる。

自然言語処理では、理由を問う質問に答える質問応答技術や、大規模テキストから因果関係知識を自動獲得する手法が研究されている。これらの手法は明示的な言語的手がかり（例えば、理由を表すため

に“because”を使う）を利用しているが、実際のテキストで表現されるメタ知識構造は実は非明示的なものがほとんどである。本研究は、これらの関係を包括的にモデル化する点と、非明示的構造をもターゲットとする点に新規性がある。

生命科学など知識の構造がはっきりしている（例えば、タンパク質相互作用など）分野では、構造化された知識をテキストデータから自動抽出する研究が行われてきた。しかし、科学技術の多くの分野、特にコンピュータサイエンスなど工学的要素が大きい分野では、そのような知識構造を予め定めることができない。したがって、上述のようなメタ知識構造に着目する必要がある。

本研究の自動抽出技術は、機械学習としては半教師あり構造学習の範疇に入るが、扱う構造が複雑であること、大規模な学習データが構築できないこと、などを考えると、非常にチャレンジングな研究テーマである。

(5) これまでの取り組み内容の概要及び実績

プロジェクトディレクターは、自然言語の構文解析の研究に従事してきており、特に深い構造に基づく構文解析において顕著な業績を挙げている。また、構文解析を生命科学分野の関係抽出やテキストマイニングに応用する研究も行っており、これは技術的には本提案研究と深い関連があるものである。これらの研究成果は、多くのトップカンファレンスやジャーナルに採録され、また受賞対象ともなっている。最近では、テキスト間の意味的関係を自動認識するテキスト間含意関係の研究など、より深い意味構造を対象とする研究を進めており、これらは本研究の基盤となるものである。

(6) 国内外における関連分野の学術研究の動向

テキストやその他のデータから因果関係を自動抽出する手法はこれまで国内外で様々な研究がある。代表的なものは、統計的手法を用いて大規模データから因果関係を抽出する手法であり、データマイニングやテキストマイニングにおいて多くの研究がある。しかし、これらの手法は実際は因果関係ではなく相関関係あるいは共起関係を抽出するものである。また、手段や目的といった関係はこのような単純な共起関係では抽出することができないため、あまり研究が行われていない。

自然言語処理では、文と文との間の関係を解析する談話関係解析の研究が行われている。人手で談話関係が付与されたコーパス（Penn Discourse Treebank など）が開発されたため、これを学習・評価データとして統計的機械学習を用いる手法が主流である。談話関係の中には因果関係など本研究が対象としている関係も一部含まれており、オーバーラップする研究であると言える。しかし、現在の研究の主流は言語的手がかり（because など）が明示的に現れる場合であり、一般的なケースで関係認識を高精度で行うことは難しい。また、既存のリソースでは談話関係がアドホックに決められており、関係の種類について理論的妥当性は不明である。本研究では、談話関係全体ではなく、一般的に知識記述に用いられる関係に限定するが、理論の構築も含めて包括的に研究を行う点で独創的である。

生命科学分野ではテキストマイニングがさかんに研究されており、その中で論文中に記述されたイベントについての属性として事実か推量か、確実性はどれほどか、といった情報を自動認識する研究が行われている。これもテキスト中に記述された知識に関する一般的関係を対象としたものであるが、本研究で対象とするものとは本質的に異なるものである。

〔2〕 研究計画

(1) 全体計画

本研究は、テキストの中で明示的・非明示的に表現された因果関係、理由、目的といったメタ知識構造を自動認識し、それに基づき構造化された知識を自然言語テキストから自動抽出する手法について研

究を行う。当面は、メタ知識構造が明確な学術論文や特許文書などのテキストを対象とし、将来的にはより一般的なテキストデータを対象とすることを検討する。

このような手法を開発するにあたって、メタ知識構造の 2 つの性質に着目する。一つは言語的な性質で、ある関係（例えば理由）を示す明示的な言語表現（手がかり表現）を利用する。もう一つは統計的特徴で、ある概念（例えば理由になりやすい概念）や概念間関係（因果関係になりやすい 2 つの概念）の統計的・確率的分布が、テキストを横断して共通することを利用する。これら 2 つの性質はそれぞれ不十分で相補的であるため、最終的にはこれらを統合した自動抽出手法を開発する必要がある。

具体的には、以下の研究項目を推進する。

・メタ知識構造の定式化・理論化
・学習・評価データとしてアノテーションコーパスの構築
・言語的手がかりに基づく自動抽出手法の開発
・統計的特徴に基づく自動抽出手法の開発
・言語的・統計的手法を統合した自動抽出手法の開発

(2) 各年度の計画

平成 26 年度

・SDRT 理論に基づく因果関係のアノテーション、連体修飾節の分類に基づくメタ知識関係アノテーション、および学術論文の分析に基づくメタ知識構造アノテーションの 3 点について、統計モデルを学習・評価するためのアノテーションコーパスを作成する。特に、各アノテーションの間の関係を分析するために同じ文書データに対してアノテーションを行うこと、また一部のアノテーションについては英語を対象にデータを拡大する。
・上記のアノテーションコーパスの分析に基づき、各理論・分析の間の理論的関係について考察を行う。平成 25 年度の研究における新聞や学術論文の分析に基づき策定したアノテーションスキーマを基にして、メタ知識構造の形式表現を定式化する。現在までに因果関係・理由・時間関係の形式表現と、学術論文におけるフローの形式的表現の策定が進められており、これらを拡張することを検討する。特に、理論化において共通かできる点に付いてはできるだけ共通化することを検討する。これにより、メタ知識構造のモデルの妥当性・一般性について検証を行う。
・平成 25 年度の実験では、既存の関係抽出手法を単純に適用するだけでは高い精度が得られないため、メタ知識構造を高精度で自動認識するための手法について研究を行う。上記のアノテーションコーパスを学習データとして用いた教師あり学習手法によるアプローチを主に検討する。ここでは、既存研究で用いられる様々な言語的手がかりとともに、大規模単語クラスタリングなどで得られる外部知識や、下記の教師なし学習あるいは半教師あり学習によるモデルを組み入れることで精度を向上させることを目指す。
・上記の手法と平行して、構造的クラスタリングを行う手法を参考に、教師なし学習によるメタ知識構造の認識手法について研究を行う。

平成 27 年度

・これまでに策定したアノテーションスキーマを異なるドメインのテキストデータに適用し、アノテーションコーパスを作成する。これにより、メタ知識構造のモデルの妥当性・一般性について分析を行う。また、これまでに開発した自動認識手法のドメイン汎用性について検証を行う。
・アノテーションコーパスの分析に基づき、メタ知識構造に関する統一的理論の構築を行う。メタ知識構造には一定の規則性（推移律など）が認められるが、これを網羅的に説明しかつ定式化した理

論はこれまでに提案されていない。これまでのコーパス分析と関連研究の分析に基づき、統一理論の構築を進める。

・言語的手がかりと統計的特徴を統合して自動認識を行う手法について研究を行う。大規模論文データから自動獲得する意味クラスの情報を利用する手法や、アノテーションコーパスと大規模テキストデータを利用して半教師あり学習手法を適用する方法、概念間関係の構造を隠れ状態としてみなして統計的学習を行う手法などが考えられる。

・前年度に引き続き、教師なし学習によるメタ知識構造の認識手法について研究を行う。特に、学術論文におけるメタ知識構造のように複雑な構造を持つようなデータに対して、構造的特徴を捉えながら高精度でメタ知識関係を認識できる手法について検討を行う。

〔3〕 研究推進・実施体制

・研究代表者

〔国立情報学研究所〕　宮尾祐介

・共同研究者

〔国立情報学研究所〕　藤田　彬、建石由佳
〔統計数理研究所〕　持橋大地
〔お茶の水女子大学〕　戸次大介、金子貴美、田中リベカ
〔情報通信研究機構〕　飯田　龍

〔4〕 研究の進捗状況

平成 26 年度は、以下の 3 点について研究を進めた。

1 . 談話構造理論と形式意味論を統合した理論 SDRT をベースにし、談話関係に形式的意味定義を与え、それに基づき談話関係を再整理する研究を行った。第一バージョンとして談話関係のリストと、それをアノテーションするためのテスト・スキーマを作成し、それに基づき少量のデータに対して実際にアノテーション作業を行った。その結果、このアノテーションスキーマでは作業者間の判断が一致しないという結論に至った。特に、現在のスキーマではイベントの時間関係に基づき談話関係を判断するようになっているが、イベントの時間の判断にゆれが生じること、さらにイベントの時間を形式的に決定すると、直感に基づく談話関係とは異なる談話関係がアノテーションされてしまうこと、などの問題が明らかとなった。以上のことから、今年度は Penn Discourse Treebank などの既存の談話構造アノテーションスキーマを参照しながら、上記スキーマを改良し、再びアノテーション実験を行う予定である。

2 . 連体修飾節を介して従属節と主節が接続される表現に対し、談話関係をアノテーションする実験を行った。アノテーションスキーマを策定し、新聞記事を対象に 20000 事例のアノテーションを行った。さらに、本データを用いて談話関係の分布や、連体修飾節に出てくる名詞のパターン等の分析を行った。その結果、連体修飾節中の名詞の種類によって談話関係にある程度の傾向があるが決定的ではないことや、連体修飾節とその前節との関係が内の関係か外の関係かによって分類ができる可能性があること、などが明らかとなった。また、本データで簡単な自動認識実験を行ったところ、単純な手法でも F 値 0.55 程度の精度が出ることが分かった。今後、上記の分析を進め、連体修飾節中の名詞や動詞、前節・後節のテンス・アスペクト等による分類が可能かどうか、検討を進める。

3．情報科学分野の学術論文に対する意味アノテーションについて、ACM および ACL の英語論文計300 論文についてアノテーション作業を行った。また、英語論文のアノテーションにあたり、新たに関係の分類を細分化し、さらに用語に対する分類にトップレベルオントロジーを参考にした分類を適用した。その結果、ある程度高精度なアノテーションを行うことはできたが、用語の分類と関係アノテーションの間に予期しない不整合が発生することが観察された。これは、主に言語仕様が厳密でないことが原因であるが、どのように分析すべきか、今後検討する必要がある。また、このデータを使った自動認識実験を行った。その結果、日本語、英語ともだいたい同程度の精度で、検索等の応用にはある程度利用可能なものの、高度な推論を行うには不十分なレベルである。今後は、自動認識精度の向上と、ここで対象としている関係について推移律等の推論規則の定義について研究を進める予定である。

また、各研究の間で綿密な情報交換を行うため、以下のように定期的に研究会合を開催した。

日　時：2014 年 4 月 4 日
テーマ：学術論文アノテーション、連体修飾節のアノテーションの進捗報告

日　時：2014 年 5 月 9 日
テーマ：連体修飾節のアノテーションの進捗報告、学術論文アノテーションのオントロジー化について

日　時：2014 年 7 月 4 日
テーマ：ACL、LREC 参加報告、SDRT に基づく談話関係の定義について

日　時：2014 年 9 月 8 日
テーマ：連体修飾節による文間関係の自動認識について、SDRT に基づく談話関係のデータ分析

日　時：2014 年 10 月 9 日
テーマ：SDRT に基づく談話関係アノテーションについて

日　時：2014 年 11 月 7 日
テーマ：学術論文アノテーションの自動認識について、SDRT に基づく談話関係アノテーション進捗報告

日　時：2014 年 12 月 8 日
テーマ：SDRT に基づく談話関係アノテーション進捗報告

日　時：2014 年 1 月 15 日
テーマ：論述問題解答プロセスの分析について、SDRT に基づく談話関係アノテーション進捗報告

〔5〕 研究成果物

① 知見・成果物・知的財産権等

なし

② 成果発表等

＜論文発表＞

〔学術論文〕

1. Toward a Discourse Theory for Annotating Causal Relations in Japanese. Kaneko, Kimi. Bekki, Daisuke. (2014). In Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing (PACLIC28), pp.460-469, Phuket, Thailand.
2. Building a Japanese Corpus of Temporal-Causal-Discourse Structures Based on SDRT for Extracting Causal Relations. Kaneko, Kimi. Bekki, Daisuke. (2014). In Proceedings of the EACL2014 Workshop on Computational Approaches to Causality in Language (CAtoCL), pp.33-39, 26th April, Gothenburg, Sweden.
3. Annotation of Computer Science Papers for Semantic Relation Extraction. Tateisi, Yuka, Shidahara, Yo, Miyao, Yusuke, Aizawa, Akiko. (2014). In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014). European Language Resources Association (ELRA). Reykjavik, Iceland.

〔一般講演〕

1. 形式意味論に基づく出来事間関係認識に向けて－リソース構築の展望とテンス「タ」のアノテーション. 宇津木舞香. 稲田和明. 金子貴美. 戸次大介. 乾健太郎. (2015). 言語処理学会第 21 回年次大会発表論文集(CD-ROM), B7-3, 京都大学, 2015/3/16-21.