人工物のオントロジー構築と学術論文からの因果関係抽出への 応用

Causal Knowledge Extraction from Scientific Papers by Engineering Ontology

梶川裕矢

Yuya KAJIKAWA

東京大学大学院工学系研究科

Graduate School of Engineering, The University of Tokyo

〒113-8656 東京都文京区弥生2-11-16工学部9号館

E-mail: kaji@biz-model.t.u-tokyo.ac.jp

本研究では、学術論文からの因果関係抽出におけるオントロジー利用の効果を検討する。まず、既存の概念モデルを統合した人工物オントロジーを提案し、妥当性の検証を行なった。結果、技術文書中の技術用語の90%以上が記述可能であった。次に、得られたオントロジーを用いて、学術論文からの因果関係抽出実験を行なった。抽出に用語間の共起関係のみを用いた場合に比べ、オントロジーを用いた場合、精度は有意に改善した。

The effect of the utilization of ontology is examined in order to extract causal relationships from scientific papers. At first, engineering ontology was proposed by integrating existing conceptual models about artifacts, and its validity was tested. Then, an experiment to extract causal relationships was performed by using the obtained ontology. The precision in the extraction is significantly improved compared to the case using co-occurrence of terms from 28.6% to 63.3%.

キーワード: 情報抽出, オントロジー, 因果関係
information extraction, ontology, causal relationship

1 緒言

学問の進歩は専門化・細分化を伴う、専門化・細分化することで、一人の専門家が特定の課題を深く探求し、そこから有益な知見を引き出すことが可能となる。しかし、そのような学問の特性と、それ

に加え、現在の学術論文の出版数の爆発的な増加は、一人の専門家が学術の全体像を俯瞰することを困難としている。結果、本来活用されるべき知識がうまく活用できていない^[1]。

そのような状況下で、専門家の研究活動を支援する情報システムの構築が強く

望まれている。その中の一つに学術情報からの因果関係抽出システムがある。特に工学分野においては、因果関係は探求すべき知識の中で本質的なものであり、大量の文献群から因果関係を抽出するシステムを開発することでその学問分野の進歩に寄与し得ると考えられる。また専門家が気づいていない因果関係をうまく抽出し、組み合わせることで新たな因果関係を発見できる可能性がある^[2]。

因果関係の抽出法としては、専門用語の文章中での共起に基づく手法が一般的に用いられている。しかし、単に共起を用いた場合、組み合わせの数が非常に多くなりノイズを含んでしまうという問題点がある。従って、何らかの手法で共起情報に限定を加える必要がある。そのような手法としては当該文章中での手がかり句を用いる手法と、当該文章によらないオントロジーを用いる手法がある。

前者の手法として、動詞に着目する手法がある。例えば、Onoらは、interact, associate, bind, complexという4種の動詞を手がかりとしてタンパク質相互作用を抽出している^[3]。また石川らは「ことにより」という表現に着目することで、特許から発明の手段と効果の関係を抽出している^[4]。後者の手法としては、因果関係を持ち得る概念の組み合わせを予め規定しておくことが考えられる。例えば、タンパク質と病名の組み合わせのみを因果関係として抽出するといったことである。Weeberら^[5]はこれをsemantic filteringと呼んでいる。

しかし、両手法とも、適用されるルールがドメインに固有のものとなる可能性がある。従って、様々な領域で機能する

因果関係抽出システムを構築するためには領域によらない汎用的な手法の開発、もしくは、各領域に固有の知識構造であるドメインオントロジーの獲得が必要である。

我々は後者の立場から研究を行っている。対象は人工物工学であり、人工物工学に固有の知識の構造を抽出し、それを因果関係抽出システムへと実装することを目的としている。

2 人工物工学における知識構造

既往の研究において、我々は人工物に関する知識を記述するためのフレームワークを提案している^[6]。これを模式的に表したものが図1である。

図1 人工物のオントロジー^[6]。

人工物工学における最終的な目標は、我々の目的に合った機能を発現する人工物を実現することである。機能は材料(Material)の物性(Property)の組み合わせとして表現できる。材料の物性は材料の構造(Structure)によって決まるが、それはプロセス(Processing)とその条件(Process Condition)によって定まる。またその人工物を産業化する際にはプロセスのパフォーマンス(Performance)が重要である。構造が物性を決めるメカニズムを

物性モデル(Property Model), プロセスのパフォーマンスを決めるメカニズムをプロセスモデル(Process Model)とここでは呼ぶ。また出来た製品(Product)を分析する工程(Analysis)もまた重要である。図1には記載していないが、分析もまた固有の条件(Analysis Condition)と分析精度等のパフォーマンス(Analysis Performance), その間の理論(Analysis Model)がある。

以上のフレームワークでどの程度人工物工学で用いられる専門用語を記述可能であるか確認するために, NEDOが発表している技術戦略マップ中に評価パラメーターとして記載されている専門用語を対象に, それらを提案オントロジー中の各概念に格納可能かどうかの検討を行った。技術戦略マップ中に現われる技術用語の90%以上を図2のオントロジーで記述可能であった^[6]ことから, これをオントロジーとして採用する。

3 因果関係抽出実験

前節で提案したオントロジーを用いて, 文献からの因果関係抽出実験を行った^[7]。対象としたのは, 材料科学の一分野である化学気相成長法 (Chemical Vapor Deposition, CVD) である。この領域設定は筆者が専門領域としており, 結果の妥当性を判断できるという理由から選定した。

コーパスとしては, Wileyが出版しているCVD分野の専門誌であるChemical Vapor Depositionの1998年から2003年までの5年間に出版された257論文のフルテキストを用いた。総語数は約75万語であ

る。

上記コーパスに対し, C-value^[8]を用いて語抽出を行ったのち, 類義語を手動でシソーラスとして同一の概念に格納した。次に, 概念間の一文中での共起情報を取得した。

超伝導転移温度 (superconducting transition temperature, T_c) は, 98文において他の概念との共起が認められたが, 因果関係の妥当性を分析したところ28文が妥当であった (精度28. 6%)。例えば, "YBCO film deposited at 700°C has T_c of 40 K." という文におけるYBCO filmと T_c の関係はHas-a関係であり因果関係ではない。

これに対し, 図1で規定したオントロジーを用いた場合, 30文が因果関係を述べた文の候補として抽出され, そのうち, 17文が妥当であった (精度63. 3%)。抽出された概念間の因果関係を模式的に表したものが図2である。

図2 分析結果の例^[7]

本論文では, 領域固有の知識構造を抽出し構築したオントロジーを用いることで, 因果関係抽出の際の精度が大きく改善することを示した。しかし, これを大規模なシステムに拡充する際には, オントロジーの自動生成が重要な鍵となり, 今後の研究課題である。

4 結言

本研究では、学術論文からの因果関係抽出におけるオントロジー利用の効果を検討した。まず、既存の概念モデルを統合した人工物オントロジーを提案し、妥当性の検証を行なった。結果、技術文書中の技術用語の90%以上が記述可能であった。次に、得られたオントロジーを用いて、学術論文からの因果関係抽出実験を行なった。抽出に用語間の共起関係のみを用いた場合に比べ、オントロジーを用いた場合、その精度は有意に改善した。

参考文献

[1] Kajikawa, Y.; Abe, K.; and S. Noda: "Filling the gap between researchers studying different materials, different methods: A proposal of structured keywords", *Journal of Information Science* Vol.32, No.6, pp.511-524, 2006.

[2] Swanson, D.R.: "Two medical literatures that are logically but not bibliographically connected", *Journal of the American Society for Information Science*, Vol.38, No.4, pp.228-33, 1987.

[3] Ono, T.; Hishigaki, H.; Tanigami, A.; Takagi, T.; "Automated extraction of information on protein-protein interactions from biological literature", *Bioinformatics*, Vol.17, No.2, pp.155-161, 2001.

[4] 石川大介, 石塚英弘, 宇陀則彦, 藤原 譲: 特許文献における因果関係の抽出と統合, *情報知識学会誌*, vol.14, No.4, pp.105-118, 2004.

[5] Weeber, M.; Klein, H.; de Jong-van den Berg L.T.W.; Vos, R.: Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. *J Journal of the American Society for Information Science and Technology*, Vol.52, No.7, pp.548-557, 2001.

[6] Kajikawa, Y.; Usui, O.; Hakata, K.; Yasunaga, Y.; Matsushima K.: "Structure of knowledge in science and technology roadmaps", *Technological Forecasting and Social Change*, Vol.75, No.1, pp.1-11, 2008.

[7] Kajikawa, Y.; Sugiyama, Y.; Mima, H.; Matsushima K.: "Causal knowledge extraction by natural language processing in material science: a case study in chemical vapor deposition" *Data Science Journal*, Vol.5, 108-118, 2006.

[8] Mima, H.; Ananiadou, S: "An application and evaluation of the C/NC-value approach for the automatic term recognition of multi-ward units in Japanese", *Terminology*, Vol.6, No.2, pp.175-194, 2000.