科学知識発見を目的とした特許のアノテーション

日浦 隆博 [1,2], 吉田 奈央 [3], 松井 陽子 [2], 河野 誠也 [2,1], 野中 尋史 [4], 吉野 幸一郎 [2,1]
[1] 奈良先端科学技術大学院大学, [2] 理化学研究所 GRP, [3] ANNOTAN, [4] 愛知工業大学
hiura.takahiro.hu6@is.naist.jp
{yoko.matsui, seiya.kawano, koichiro.yoshino}@riken.jp
annotan.tsukuba@gmail.com, hnonaka@aitech.ac.jp

概要

科学知識発見や仮説生成を行おうとする場合、その用途に特化した知識ベースと知識推論モデルが必要となる。本研究ではその第一歩として、特許文書へ科学知識発見を目的としたアノテーションを行うためのアノテーションスキーマを構築した。アノテーションスキーマの構築にあたっては、特許の非専門家でもアノテーションが可能なように文法や手がかり語に着目した定義を行った。

1 はじめに

深層学習や言語モデルなどの研究の発展により、言語を用いた知識推論の可能性が取り沙汰されるようになった。とりわけ大規模言語モデル（LLM）が持つ推論能力には大きな期待が寄せられており、新しい仮説の生成や知識発見といった応用に目が向けられている。しかし、LLM を用いて仮説生成や知識発見を実現しようとする場合、いくつか問題が指摘されている。

まず、LLM は一般文書で学習されているため、特定のドメインにおける仮説生成や知識発見においては一般的な内容の出力が大半となる。仮説生成や知識発見を目的とする場合、論文や特許など知識を記述したドメインテキストをかなりの量（最低でも数 B トークン）で準備することが必要である。

また、LLM は知識推論を目的関数として学習されているわけではないので、しばしば誤った推論が出力される。一般的な常識推論を対象とした場合、LLM に与えるプロンプトを工夫したとしても、その出力は 3 割程度誤りが含まれる [1]。このため、知識推論を目的関数とした新たなモデルの構築 [2]、あるいは LLM のチューニングが不可欠である [3]。

専用の知識推論モデルを構築することを想定する場合、多くの知識推論モデルでは、ConceptNet [4] や ATOMIC [5] に代表されるように、事態（ノード）とその関係（エッジ）を活用した 3 つ組で知識を表現する。この知識表現は常識推論において必要な関係が定義されていることが多いが、仮説生成や知識発見には必ずしも特化していない。科学的知識発見を目的とする場合、これに利用される事態と関係を再定義する必要がある。

本研究ではこれらの問題を踏まえつつ、科学知識発見を目的として特許文書を対象としたアノテーションスキーマを構築する。特許文書は技術文書であり、また特許の請求範囲を明確化するために可能な限り一意に解釈できるよう知識が記述される。また、日本における特許は特許庁が一括管理して公開しており、その大半が電子化・公開されていることから、論文よりも統計的手法を適用する対象としやすい。一方で法的文書としての性質も持ち、その記述は一般的な文の記述様式とやや異なる。これらの性質を踏まえ、まず本研究ではアノテーション対象の知識について定義し、さらにそれらを、特許に対するドメイン知識が持たないアノテータでも付与可能なスキーマとして構築することを目指す。

2 知識とその関係

特許や論文に記述される知識は様々であり、その目的に応じて認定基準もまた多様である。科学知識発見を目的とする場合、科学知識を構成する重要な要素として何らかの操作、あるいは操作による様態の変化が考えられる。例えば、ある手段によって何らかの効果を得る場合、この「手段」によって「効果」が「引き起こされる」関係にあることが科学知識の一つの単位となる（図 1）。

特許で一般的な装置の説明においては、装置にはなんらかの「目的」が存在し、その目的を実現する上での「問題」、その「問題」に対する「解法」、その「解法」を与えた場合の「効果」が重要な知識と

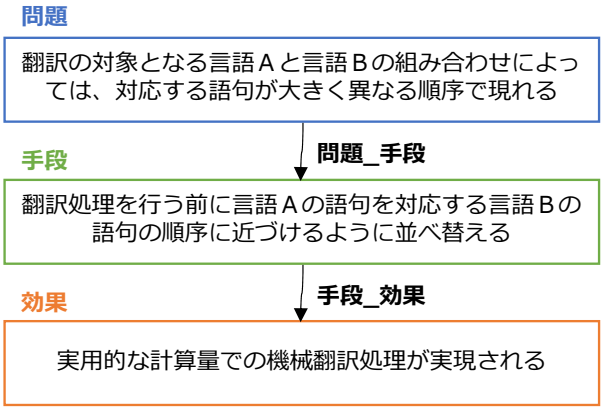

図 1　特許に記載された問題、手段、効果の例 [6]

なる。これら 4 種類を今回考慮する事態（ノード）の基本とする。操作や様態の変化を表す事態は、動詞、形容詞、あるいは事態性名詞（末尾に「する」を補うことで事態化する名詞）を主辞に持つ。この主辞は述語項構造における述語とほぼ同一視でき、主辞に従属する必要最低限の格要素と合わせて節を構成する。この主節を最小限の単位とする。

この主節に対して対象を限定する従属節が存在する場合、その範囲限定を行う「条件」は重要な情報を含むため、知識の一単位として認定する。例えば図 1 の「問題」における「翻訳の対象となる言語 A と言語 B の組み合わせによっては、」は、その後の「対応する語句が大きく異なる順序で現れる」という問題が生じる際の重要な「条件」である。

また、特許特有の記述方式として、連体修飾によって用言を繰り返し記述するパターンが頻出する。例えば、「コーパスから学習した分類器を利用する」という表現では、「コーパスから学習する」「分類器を利用する」という 2 つの事態が連接している。こうした連接は知識の範囲限定において重要であるものの、特許文書においては複数回に渡り連続して用いられる場合も多い。こうした連接を無分別に連結することはかえって知識発見という目的の妨げになるため、連体修飾による連接は必要最小限の範囲で事態の 1 単位として扱うことが望ましい。

「条件」と同様に、特許ではその内容に関わる分野での常識や共通認識が記述され、そうした記述が重要な知識となる場合も多い。こうした内容は「常識」として認定することが必要である。

科学的知識発見や仮説生成を行おうとする場合、知識の認定だけでなく、ある特定の知識と特定の関係を持つ別の知識への関係を知りたい。例えばある「手法」に対する「効果」を予測したい場合、先行する知識である「手法」に対して「因果」の関係にある知識を生成したい。特定の「問題」に対して解法となりうる「手法」についても同様である。こうした三つ組を用いる知識推論エンジンは近年一般的に利用されるようになっており [2]、こうした関係を大量に用意することで科学知識発見に特化した知識推論エンジンを構築することができる。「因果」は科学知識発見におけるもっとも重要な関係であり、主として科学知識における「目的」「問題」「手段」「効果」の関係として付与できる。

これらの知識に対して「条件」が知識が指す範囲の限定を行う場合、どの「条件」がどの知識を限定しているか知る必要がある。この条件部分は対象の知識から離れたスパンで記述される場合があるため、その関係を明示する手段が必要である。また、構文構造における係り受け関係や、連体修飾が存在する場合、知識を分割して認定せざるを得ないケースが存在するため、これらの関係を明示する手段も必要となる。

3　アノテーションスキーマ

2 節で示した内容を特許文書の非専門家でもアノテーション可能なよう、アノテーションの手順を知識の認定と関係の認定の 2 段階に分離する。

3.1　知識の認定

アノテーションの第一段階では、知識の候補となりうるテキスト範囲を示すスパンと、そのスパンの役割であるタグを付与する。スパンおよびタグの認定においては、対象の事態が「専門知識」として見なされるかに留意しつつ、一般的な知識については認定から除く。例えば、「コップを落とす」と「コップが割れる」のようなテキストの関係は「手段」と「効果」として捉えることもできるが、一般的な常識的知識の範疇と見なしてアノテーション対象としない。スパンに対して与えるタグとして、以下の 7 種類（小分類含め 8 種類）を定義する。

1. 手段: method
2. 目的: purpose
3. 問題: problem
4. 効果: effect
   (a) 技術効果: tech-effect
   (b) ユーザ効果: user-effect
5. 条件: condition
6. 常識: common

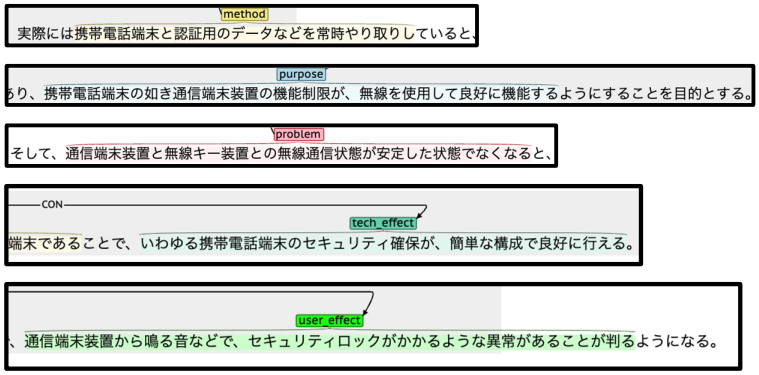

図 2　科学知識のスパンとタグの例

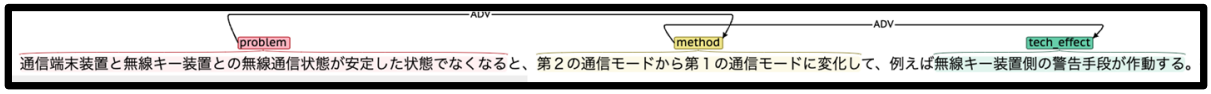

図 3　知識同士の因果関係（ADV）の認定例

7. その他: others

1-4 は科学知識の中核であり、5-7 はそれらの補足的内容である。1-4 の例を図 2 に示す。これらの知識は複数の異なる知識と別の関係をもつ場合があり、そうしたケースでは二重にスパンとタグを付与する。ここで「技術効果」と「ユーザ効果」は非専門家にとって区別が付きにくい場合があり、そうしたケースはスーパークラスである「効果」を付与する。

3.2　関係の認定

アノテーションの第二段階で、知識の候補同士の関係を認定する。この際、知識のスパンに付与されているタグは関係と重要な関わりを持つため、必要に応じてタグを変更したり、スパンを増やしたりしても構わない。関係に与えるタグとして以下の 4 種類を定義する。

1. 因果: ADV
2. 条件: CON
3. 依存: DEP
4. 連接: CHAIN

ADV は主として科学知識同士の関係に対して与えられるもので、本アノテーション結果を用いる際の主たるアプリケーションとして想定している。この際、head となる知識が原因となり tail となる知識が結果として導かれるよう関係の方向を認定する（目的→手段、問題→手段、手段→結果など）。ADV 認定の例を図 3 に示す。CON は知識における「条件」と主たる関係があり、head が tail の指示範囲を限定する。ただし、CON は「条件」以外の知識同士の関係として付与される場合もある。この判定には「なら」「場合にのみ」「であるから」などの手がかり語を主として用いる。CON は ADV の部分集合となるため、「条件」以外の知識同士の関係として付与する場合は CON の認定を優先する。DEP は係り受け関係を指し、本来同一のスパンとして認定されるべき知識同士が文中の離れた箇所に存在する場合に、head の知識には「その他」を付与し、本来連続するべき tail の知識と DEP で繋ぐ。CON と DEP の認定例を図 4 に示す。CHAIN は特許特有の関係で、特に連体修飾する知識同士の関係を示す。特許において連体修飾は頻出するため、連体修飾全てを 1 単位として認定すると認定スパンが非常に長大になり、アノテーションが困難になる。そこでスパンのレベルではこれらをいったん分割し、必要に応じて CHAIN で繋ぐ。CHAIN の認定例を図 5 に示す。

4　特許に記述される知識

特許は発明に関する技術的詳細が記述されており、また公開情報であることから、古くから自然言語処理、情報抽出の研究対象として注目されてきた。本節ではこれまでの特許を対象とした情報処理、知識抽出について説明し、それらに対する本研究の位置づけを述べる。

4.1　事態・事物についての知識

最も一般的な知識抽出、あるいは知識抽出のためのアノテーションは、特許中のモノやコトに対するラベルの付与である。特許はその性質として特許の権利範囲を限定するための発明品、機能、目的など

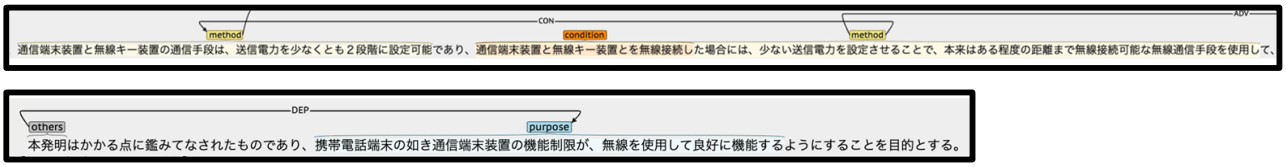

図 4　条件（CON）・依存（DEP）関係の認定例

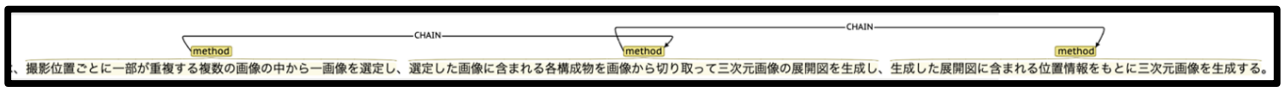

図 5　連体修飾に対する連接（CHAIN）の認定例

が列挙されているため、それらを対象とした情報抽出が行われることが多い。特許に関する属性定義の先行研究・取り組みとしては化学分野の化学物質の構造やプロセスに関するものが主流となっている。公開コンペティションにおけるデータセットとしては化学式・構造と化学物質に関する各種属性に関する CLEF-IP[7] や TREC-CHEM[8] が挙げられる．この他、合成プロセスに関する属性定義と抽出に関する研究としては [9] や [10] がある。しかしながら、知識発見のためには化学分野に限らず技術分野を横断して属性を定義したうえでデータセットを整備する必要がある。技術分野を問わず属性を定義した公開データセットを用いたコンペティションとしては NTCIR-8 PATMN タスクがある [11]。このタスクでは特許中の知識として重要な技術と効果を Technology と Effect としてアノテーションしている。さらに NTCIR-8 PATMN の Technology や Effect の定義を細分化し自動抽出する手法を提案したものとしては Nonaka らの研究 [12] や坂地 [13] らの研究がある。一方でこれらの研究は知識発見の観点からは課題を抱えていた。具体的には、「Technology が特許全体の技術内容に網羅的にアノテーションするものではない」、「従来技術の内容や各属性の関係などの属性も定義されていない」など知識発見のための情報源としては不足がある問題があった。そこで、本研究では従来のアノテーション基準を踏襲しつつ、特に特許に記述された科学技術の知識を抽出するためのラベルセットを定義した。

4.2　関係知識

特許は複数の構成要素が複雑に絡み合って構成されるため、事態・事物に加えてそれらの関係を理解することが重要である。しかしながら、特許は一種の法的文書であり、独特の記述形式を持っている。そのため、専門家以外の人にとっては極めて読みにくいものになっている。そこで、特許文書の可読性を向上させるための研究がこれまで多く行われてきた。Shinmori ら [14] や Bouayad-Agha ら [15] は、特許文書における談話関係を、談話標識と統語構造解析の結果に基づいたルールベース手法により抽出した。しかしながら、明示的な談話標識の存在を仮定するこれらの研究では、抽出できる知識の種類や、その抽出性能には限界があった。また、こうした特許ドメインにおける談話関係の解析タスクは、基本的には特許請求の範囲（請求項）の解析を目的として設計されており、発明の詳細内容から科学技術に関する関係知識を発見することを目的としたものではなかった。そこで、本研究では、アノテーションする関係の範囲を科学技術の発見に関する関係知識を得るために特化したアノテーションスキーマを構築した。

5　まとめ

本研究では、科学知識の発見を目的として、特に特許文書を対象とした知識および関係のアノテーションスキーマ構築を行った。特許文書は一般文書とは記述様式が異なり、読解に専門知識を要する。そこで非専門家でもアノテーションが可能なよう、アノテーションスキーマ構築においては文法や手がかり語に注目した定義を行った。今後は本定義に従ってアノテーションを拡充し、実際の事例に応じてアノテーションマニュアルを更新・公開する[1)]。また、構築されたデータを元に科学知識発見のための知識推論モデル構築を行う [16]。

1）　https://github.com/caesarwanya/TAURO-hypothesis

謝辞

本研究は、JST ムーンショット型研究開発事業 JPMJMS2236 および JST 戦略的創造研究推進事業（ACT-X）JPMJAX22A4 の支援を受けたものです。

参考文献

[1] Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. Symbolic knowledge distillation: from general language models to commonsense models. In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 4602–4625, Seattle, United States, July 2022. Association for Computational Linguistics.

[2] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. Comet: Commonsense transformers for automatic knowledge graph construction. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 4762–4779, 2019.

[3] Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. Large language models still can't plan (a benchmark for llms on planning and reasoning about change). In **NeurIPS 2022 Foundation Models for Decision Making Workshop**, 2022.

[4] Hugo Liu and Push Singh. Conceptnet—a practical commonsense reasoning tool-kit. **BT technology journal**, Vol. 22, No. 4, pp. 211–226, 2004.

[5] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning. In **Proceedings of the AAAI conference on artificial intelligence**, Vol. 33, pp. 3027–3035, 2019.

[6] 須藤克仁, 永田昌明, 星野翔, 宮尾祐介. 語順並べ替え装置、翻訳装置、方法、及びプログラム. 特開 2015-153182(P2015-153182A), JP, 2015.

[7] Florina Piroi, Mihai Lupu, Allan Hanbury, and Veronika Zenz. Clef-ip 2011: Retrieval in the intellectual property domain. In **CLEF (notebook papers/labs/workshop)**, 2011.

[8] Mihai Lupu, Jimmy Huang, Jianhan Zhu, and John Tait. Trec-chem: large scale chemical information retrieval evaluation at trec. In **Acm Sigir Forum**, Vol. 43, pp. 63–70. ACM New York, NY, USA, 2009.

[9] Kohei Makino, Fusataka Kuniyoshi, Jun Ozawa, and Makoto Miwa. Extracting and analyzing inorganic material synthesis procedures in the literature. **IEEE Access**, Vol. 10, pp. 31524–31537, 2022.

[10] Fusataka Kuniyoshi, Kohei Makino, Jun Ozawa, and Makoto Miwa. Annotating and extracting synthesis process of all-solid-state batteries from scientific literature. **arXiv preprint arXiv:2002.07339**, 2020.

[11] Hidetsugu Nanba, Atsushi Fujii, Makoto Iwayama, and Taiichi Hashimoto. Overview of the patent mining task at the ntcir-8 workshop. In **NTCIR**, pp. 293–302, 2010.

[12] Hirofumi Nonaka, Akio Kobayashi, Hiroki Sakaji, Yusuke Suzuki, Hiroyuki Sakao, and Shigeru Masuyama. Extraction of effect and technology terms from a patent document (theory and methodology). **Journal of Japan Industrial Management Association**, Vol. 63, No. 2, pp. 105–111, 2012.

[13] 坂地泰紀, 野中尋史, 酒井浩之, 増山繁. Cross-bootstrapping: 特許文書からの課題・効果表現対の自動抽出手法. 電子情報通信学会論文誌 D, Vol. 93, No. 6, pp. 742–755, 2010.

[14] Akihiro Shinmori, Manabu Okumura, Yuzo Marukawa, and Makoto Iwayama. Patent claim processing for readability-structure analysis and term explanation. In **Proceedings of the ACL-2003 workshop on Patent corpus processing**, pp. 56–65, 2003.

[15] Nadjet Bouayad-Agha, Gerard Casamayor, Gabriela Ferraro, Simon Mille, Vanesa Vidal, and Leo Wanner. Improving the comprehension of legal documentation: the case of patent claims. In **Proceedings of the 12th International Conference on Artificial Intelligence and Law**, pp. 78–87, 2009.

[16] 日浦隆博, 河野誠也, Angel Ferdinando Garcia Contreras, 吉野幸一郎. 敵対的生成ネットワークを用いた記号的知識蒸留. 言語処理学会第 30 回年次大会論文集, 2024.