高分子材料オントロジーの構築に向けた Wikidata からのドメイン概念関係抽出方法の検討

Consideration on Extraction Method of Relationships among Domain Concepts from Wikidata for Construction of Polymer Material Ontology

佐藤壮[1]　久米慧嗣[1]　古崎晃司[1]

Takeshi Sato[1], Satoshi Kume[2], Kouji Kozaki[3]

[1]大阪電気通信大学
[1]Osaka Electro-Communication University

Abstract: Abstract: Ontology development for a specific domain is a task to take a large cost. To reduce the costs, ontology development methods are needed. In this study, we investigate a semi-automatically method to extract concepts for domain ontologies from Wikidata, a large-scale public Linked Open Data (LOD) to develop a preliminary ontology. Our previous work extracted domain concepts with their conceptual hierarchies from Wikidata. In this study, we use these extracted concepts as input to extract conceptual relationships among domain concepts from Wikidata This paper discusses a method to extract relationships as triples related to domain concepts, and examines how extracted relationships could be used as conceptual relationships for the domain ontology.

1. はじめに

オントロジーとは概念とその概念関係を記述したものであり、その根幹となす箇所の作成手順は「用語の列挙」、「クラスの定義」、「プロパティ（概念関係）の定義」となっており、このプロセスを往復しながら行われる[1]。しかし、このような作成手順は現在においても手動で行われることが多く、オントロジーの自動構築手法は確立されていない。また、本研究で扱うような広範囲のドメインオントロジーを手動で作成するとなると膨大な時間を要することは容易に想像できる。そこで本研究プロジェクトでは外部知識から対象ドメインの概念、概念関係を機械的に抽出し、対象ドメインオントロジーの初期版を半自動構築することを目標としている。

現行研究プロジェクトの先行研究として RDF（Resource Description Framework）形式で公開された対象ドメインを限定しない大規模な LOD（Linked Open Data）の一つである Wikidata から対象ドメインの概念抽出と構造を表す概念関係を抽出し、初期オントロジーの半自動構築を行う手法の開発を進めている。先行研究では、高分子材料分野を対象とし、オントロジーを構成する概念階層を Wikidata から週出する手法の開発を行った[2]。これはオントロジー作成の「用語の列挙」および、「クラスの定義」におけるクラス階層の定義の部分に相当する。そこで本研究では、現在の初期オントロジーに無いセマンティックな概念関係を Wikidata から機械的に抽出し、「プロパティの定義」に相当するプロセスを実現する手法の開発を目標とする。

2. 研究アプローチ

2.1 対象ドメインの概念抽出(先行研究)

現行研究プロジェクトの先行研究として、Wikidata からの高分子材料分野を対象としたドメイン概念の抽出と、それらの概念階層を表す概念構造関係の抽出を行った[2]。対象研究ではまず、対象ドメインの語彙をドメイン文書（高分子データベース PolyInfo の説明文書）から収集し、LOD（Wikidata）に存在するエンティティとのリンキングを行い、188 のエンティティを検索エンティティとして得た。続いて Wikidata から SPARQL クエリにより検索エンティティの上位概念関係（instanceOf / subClassOf）の検索を行い、検索エンティティが下位に複数存在するエンティティを共通上位エンティティとして定義した。この共通上位エンティティ同士のグラフ関係を共通パスとして定義し、共通パスを除くグラフを再構成した。この結果、下位概念に検索エンティティをもつ共通上位エンティティを 144 の展開共通上

位エンティティとして決定し、その下位概念に有する検索エンティティからの最短段数（距離）の最高値を展開段数として決定した。最後に展開共通上位エンティティごとに決定した展開段数の下位概念を検索し、対象ドメインのオントロジー概念候補として抽出した。

この結果、6,306,735 のエンティティと 6,404,672 の上位概念関係を持つ高分子オントロジーの初期版を構築することができた。

2.2 本研究のアプローチ

本研究では上記の先行研究で得られた高分子オントロジーの初期版に含まれる概念をもとに LOD（Wikidata）への検索を行い、対象とする高分子ドメインに関わる概念関係の抽出を行う。このときこの高分子オントロジーに含まれる概念を対象ドメイン性をもつ概念であると仮定し、それらの概念（Wikidata 上のエンティティ）がもつトリプルに含まれる概念関係は対象ドメイン性を有する関係であつと想定して抽出する。

3. ドメイン概念関係の抽出手法

本項ではオントロジーにおける対象ドメインの概念関係（プロパティ）の抽出手法について述べる.

3.1 対象ドメインのトリプル取得

本研究では先行研究（2.1）で得られたオントロジーの初期版に含まれる概念を対象ドメインの専門語彙を表すもの（以下対象ドメインエンティティと呼ぶ）とし、これらのエンティティをもとに、対象ドメインに関わる概念関係（オントロジー候補概念関係と呼ぶ）を Wikidata から抽出する。なお、RDF グラフ上ではこれらの対象ドメインエンティティがラベルとして持つ文字列が対象ドメインに関する語彙に相当する。

以下、オントロジー候補概念関係を抽出する手順について述べる。まず、Wikidata から主語または目的語に対象ドメインエンティティを含むトリプルの一覧を取得する。次に取得したプロパティを次の条件ごとに場合分けを行う。場合分けのイメージを図 1.に示す。

① 主語のみに対象ドメインエンティティをもつトリプル
② 目的語にのみ対象ドメインエンティティをもつトリプル
③ 主語、目的語両方に対象ドメインエンティティをもつトリプル

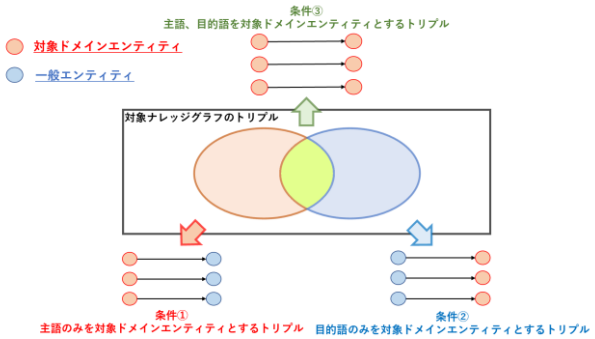

図 1. 条件ごとの場合分けイメージ図

なお、各条件で取得したトリプルは原理的に重複しない。

3.2 各条件で取得されたトリプル検討

本項では 3.1 の条件ごとに場合分けされた取得トリプル群を比較し、それにより得られたプロパティについて評価を行う。そして、どのような条件下において対象のドメイン性を持つプロパティが取得できるかを検討する。

本研究では次の項目についての検討を行う。

(1) 全体的なトリプル取得状況の比較
(2) オントロジー候補概念関係が含まれる割合の取得条件ごとの比較
(3) 共通取得プロパティの検討
(4) 非共通取得プロパティの検討
(5) 主語・目的語のクラス階層を用いた比較

まず、全体的なトリプルの取得状況の比較を行う。具体的には、⑴では各条件での取得されたトリプル(プロパティ)の総数の比較検討を行い、⑵では各条件で取得されたプロパティの種類数の比較検討を行う。

続いて、取得プロパティに対して対象ドメインのオントロジー候補概念関係であるかどうかについて、3 段階の主観的な評価を目視により行い、トリプルの取得条件ごとに比較する。⑶、⑷では各条件により取得されたプロパティの種類を比較し、⑶では条件①と条件③で取得したプロパティの種類の一覧において、共に出現したプロパティ（以下、共通プロパティと呼ぶ）に各評価のプロパティがどれだけ含まれているかを確認し、検討を行う。同様に条件②と条件③、条件①と条件②の共通プロパティについても検討を行う。⑷では条件①、条件②で取得したプロパティの種類の一覧において先程の共通プロパティを除くプロパティの種類一覧（以下、非共通プロパティと呼ぶ）に対して各評価のプロパティがどれだけ含まれているかを確認し、検討を行う。同様に条件②と条件③、条件①と条件②の共通プロパテ

ィについても検討を行う。

検討項目⑸では条件③により取得されたトリプルの主語、目的語エンティティがオントロジーの初期版における概念（クラス）階層において、どのクラスに含まれているかを確認し、その分布について検討、評価を行う。具体的には、今回の先行研究にて抽出した高分子オントロジーの初期版の概念階層の分布から、特に対象ドメイン性を有する概念として同定された material、physical_characteristics、process の 3 つのクラスの下位概念について検討する。

4. 結果と考察

本項では 3.1 で示した各条件での取得トリプルに対して 3,2 で示した項目について検討を行う。

4.1 検討項目の結果と検討

4.1.1 検討項目⑴:取得トリプル(プロパティ)数の比較

各条件における取得トリプル数を表 1.に示す。対象ドメインオントロジーに存在する概念構造関係（wdt:P31 instance of,wdt:P279 subclass of）と比較しても、多くのプロパティを取得することが出来た。表の括弧内の数値は目的語にリテラルデータを含まないトリプルの総数を示す。リテラルデータは常に目的語として現れる為、目的語を対象ドメインエンティティとして取得した条件②、条件③ではこのようなトリプルは含まれない。条件①において約 8 割（78.1%）近くがこのリテラルデータを含むトリプルとなった。これは全取得トリプル数の約 3 割(31.5%)に値する。また、条件③の取得トリプル数は全取得トリプル数と比較して、約 1 割(12.6%)程度まで減少してしまう。リテラルデータを含まない全取得トリプルと比較すると 18.4%となる。

表 1. 取得トリプル(プロパティ)数

対象データ	取得トリプル数
対象ドメインオントロジー(概念構造関係のみ)	6,404,672
取得トリプル(全取得トリプル)	94,314,858(64,596,910)
取得トリプル(条件①：主語のみ)	38,044,058(8,326,110)
取得トリプル(条件②：目的語のみ)	44,375,762
取得トリプル(条件③：主語、目的語)	11,895,038

※ () 内の数値は目的語にリテラルを含まないトリプル総数

4.1.2 検討項目⑵:取得プロパティの種類数の比較

各条件における取得プロパティの種類数を表 2.に示す。表の括弧内の数値は 4.1.1 と同様に取得トリプルにリテラルデータを含まない場合のプロパティ種類数を表している。まず、全プロパティ種類数の内 72.3%がリテラルデータを含むトリプルによるものであり、取得プロパティの大半の種類がこれに属するといえる。これは当トリプルに識別子（Identifier）に属するプロパティが多いことが理由に上げられる。

リテラルデータを含まないデータとして比較した場合、条件①、条件②で取得したプロパティの種類数が酷似していることが分かる。しかし、２つの取得プロパティ群からこのようになる原因を見つけるに至らなかった。また、4.1.1 において全体取得トリプル数と比較して少ない割合となった条件③であったが、取得プロパティ種類数として比較した場合全体の約 6 割(60.8%)と比較的大きな割合となっている。

表 2. 取得プロパティ種類数

対象データ	取得プロパティ種類数
取得トリプル(全取得トリプル)	4220(1168)
取得トリプル(条件①：主語のみ)	3982(930)
取得トリプル(条件②：目的語のみ)	929
取得トリプル(条件③：主語、目的語)	710

※ () 内の数値は目的語にリテラルを含まないトリプル総数

4.1.3 検討項目⑶:共通取得プロパティ

本項目 4.1.3 と次の項目 4.1.4 では、各条件における取得プロパティの種類についての比較を行う。取得プロパティを目視により次の 3 つの段階で評価し、検討を行う。また、条件①のプロパティ評価に関しては目的語にリテラルデータを含まない取得トリプルデータ(930 種類のもの)を使用している。これは目的語がリテラルデータとなるトリプルと目的語がエンティティとなるトリプルでは共通するプロパティが表れない為である。

評価A) 対象ドメインに必要なプロパティ
評価B) 対象ドメインに必要になる可能性があるプロパティ
評価C) 対象ドメインに必要のないプロパティ

評価を行ったプロパティの例を図 2.に示す。

評価A	評価B	評価C
分子機能	役割	ジャンル
材料	引用文献	国
特性	専門分野	スポーツ
用途	崩壊性物質	曜日
物理的相互作用	原因	脚本
共役酸	研究する分野	遺産保護指定
共役塩基	発見時期	重要な出来事
以下のポリマー	表記法	参加イベント
以下のモノマー	使用するもの	作曲
副産物	以下を含める	場所

図 2.プロパティの評価一例

共通取得プロパティについて評価を行った結果を表 3.に示す。すべての条件においてプロパティの種類はおおよそ均等に存在していることが分かる。

表 3. 共通取得プロパティの評価

比較条件	共通プロパティ数	評価Aプロパティ数	評価Bプロパティ数	評価Cプロパティ数
条件① vs 条件③（主語のみ）（両方）	574	22(3.9%)	23(4%)	587(91.6%)
条件② vs 条件③（目的語のみ）（両方）	641	25(3.9%)	25(3.9%)	591(92.2%)
条件① vs 条件②（主語のみ）（目的語のみ）	723	23(3.2%)	24(3.3%)	676(93.5%)

4.1.4 検討項目⑷:非共通取得プロパティ

共通取得プロパティについて評価を行った結果を表 4.に示す。どの条件下においても評価 A、評価 B のプロパティが殆ど出現していないことが見て取れる。

表 4. 非共通取得プロパティの評価

比較条件	非共通プロパティ数	評価Aプロパティ数	評価Bプロパティ数	評価Cプロパティ数
条件① vs 条件③（条件①のみで出現）	356	1(0.3%)	2(0.6%)	353(99.2%)
条件① vs 条件③（条件③のみで出現）	136	3(2.2%)	3(2.2%)	130(95.6%)
条件② vs 条件③（条件②のみで出現）	288	1(0.3%)	2(0.7%)	285(99%)
条件② vs 条件③（条件③のみで出現）	69	0(0%)	1(1.4%)	68(98.6%)
条件① vs 条件②（条件①のみで出現）	207	0(0%)	1(0.5%)	206(99.5%)
条件① vs 条件②（条件②のみで出現）	206	3(1.5%)	3(1.5%)	200(97.1%)

4.1.5 検討項目⑸:主語と目的語のクラス間関係

条件③により取得されたトリプルのうち、主語、目的語のオントロジー概念候補エンティティが material、physical_characteristics、process の下位概念であるものに限定したときの主語、目的語のクラス関係ごとに総トリプル数を集計した。対象クラス間関係の総取得トリプル数を表 5.に示す.主語、目的語共に対象クラスエンティティであるトリプルは全体の 0.8%と非常に少ないことが分かった。(99.2%は主語、目的語のどちらか片一方、もしくは両方に対象クラスエンティティを含まないトリプル)

表 5. 対象クラス間関係数

主語と目的語の対象クラス関係	対象トリプル数
“material”-”physical_characteristics”	34964
“material”-”process”	53334
“physical_characteristics”-”process”	11756

条件③により取得された全トリプルに含まれる対象クラスのエンティティ数を表 6.に示す。主語に含まれる material クラスのエンティティ割合は全体の 90.1%と全体のほとんどを占める結果となった。また、目的語内では process クラスのエンティティが 41.2%と最も多く現れた。

表 6. トリプル内の対象クラスのエンティティ数

対象クラス	対象クラスエンティティ数（主語）	対象クラスエンティティ数（目的語）
material	5066546(90.1%)	90237(33.5%)
physical_characteristics	10099(0.2%)	68118(25.3%)
process	547823(9.7%)	111090(41.2%)

※ () 内の数値は対象クラスエンティティの割合

5. おわりに

本研究では先行研究により抽出されたオントロジー概念候補を入力として使用し、Wikidata から対象ドメインに関係性のあるトリプルを抽出し、結果から対象ドメイン概念関係抽出方法の検討を行った。全体取得トリプルの総数は既存のクラス概念関係と比較して多い結果となったが、全取得トリプル数の内、条件③による取得トリプルが少ない結果となった。取得プロパティの種類数に関しては全体の約 7 割がリテラルデータを目的語に含むトリプルのプロパティであり、このプロパティ群に対して評価を別途行う必要があると考えられる。各条件による取得データの組み合わせによる共通、非共通プロパティの評価では非共通と比べ、共通して出現するプロパティ群で対象ドメイン性を有するプロパティが多く出現することが確認された。また、条件③により取得されたプロパティだけでなく、条件①と条件②の共通プロパティにおいても同様に評価の高いプロパティ数が出現していることから、これら条件の共通プロパティは概念関係抽出において重要になると考えている。条件③における取得トリプルの主語-目的語間クラス関係の評価では、双方に対象クラスエンティティが含まれるトリプルが非常に少ないことが分かった。しかし、material クラスのエンティティは主語に偏る傾向があることが確認された。

本研究では対象ドメインの概念関係抽出方法の検討にとどまるため、今後の研究では今回の検討結果から適切な概念関係抽出方法を検討し、それにより得られたプロパティの評価を行う必要があると考える。

謝辞

本研究の一部は，国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の助成による.

参考文献

[1] 來村徳信:オントロジーの普及と応用 pp. 39-44, (2012)

[2] 久米慧嗣, 古崎晃司: 高分子材料オントロジーの構築に向けた Wikidata からのドメイン概念, 人工知能学会第二種研究会資料, SIG-SWO-052-03 (2020)