

# 知識グラフ作成のための統合知識基盤の構築に向けて

Towards Implementation of Integrated Knowledge Infrastructure to Create Knowledge Graphs

市瀬 龍太郎 \*1\*2

Ryutaro Ichise

Natthawut Kertkeidkachorn\*2

Natthawut Kertkeidkachorn

趙 麗花 \*2

Lihua Zhao

\*1国立情報学研究所

National Institute of Informatics

\*2産業技術総合研究所

National Institute of Advanced Industrial Science and Technology

Knowledge graphs are recently used for many AI tasks such as question and answering. However, there are still problems to create precise knowledge graphs. In this paper, we discuss three problems and propose a solution called integrated knowledge infrastructure consisting of four modules.

## 1. はじめに

機械に知的な作業を行わせるためには、知識が大きな役割を果たす。たとえば、質疑応答システムでは、質問に対する回答の作成に、知識を利用するし、Web 検索システムでは、ユーザに必要な情報を提供する際に、知識を利用している。このような知識を構築する際には、知識グラフ (Knowledge Graph) を用いられることが多い。知識グラフとは、〈主語、述語、目的語〉の3つ組 (トリプル) を用いて、エンティティとその関係を表現したものである。

知識グラフを作成する際には、通常、Linked Data から必要な情報を取得し、それらを整理の上で、別のデータベースなどの情報を統合することで作成される。しかし、そのようなプロセスを経て作成しても、精密な知識グラフを作成することは困難であることが多い。作成が困難な点にはいくつかの理由がある。まず、第1に、利用した Linked Data やデータベースなどの資源自身が不完全である点である。利用するアプリケーションにも依存するが、知識グラフを作成する際には、様々な場合に対応できるようにするために、大量のエンティティに関する情報が集められる。そのため、全てのエンティティの全ての関係について、予め完全な情報を準備することは難しい。第2に、知識グラフは時間的に変化し、新しい情報を常に入れる必要がある点である。知識グラフは静的に存在するものではなく、常に情報が追加、更新される必要がある。そのために、最初の Linked Data やデータベースに記述されていなかった知識でも、新たに知識が生ずれば、その知識の記述方法を必要に応じて新たに定義し、知識を迅速に追加することが必要となる。第3に、知識グラフとして用いられた知識には、誤った知識が混入している場合がある点である。知識を作成する際には、人手による作業が行われるため、誤った情報が混入してしまうことが避けられない。本稿では、精密な知識グラフを作成するために、知識グラフの作成方法に関して、上記の問題点の考察を行い、必要な技術要素の同定、および、それを統合した精密な知識グラフを作成するためのフレームワークを提案する。

## 2. 知識グラフ作成のための諸問題

前章において、知識グラフを作成する際の問題点として、知識の資源自身が不完全である問題、新しい情報を追加する問題、誤った知識の混入問題の3つの問題を提起した。本章では、それらの問題について考察を行う。

### 2.1 知識資源の不完全性

この問題は、知識グラフを作成する際に、用いる資源が不完全なことによって発生する問題である。例えば、既存の Linked Data を用いると、多数のエンティティに対する情報が入手できる反面、全ての情報が完全に網羅されているわけではない。例えば、多数の人物に関する情報が書かれていても、全ての人物に対して、国籍や生まれた場所などが記述されているわけではない。そのような欠けた情報は補完する方法が別途必要となる。

補完するためのアプローチとして、知識グラフ補完 (Knowledge Graph Completion) という方法が知られている。知識グラフ補完では、記述されている既知の情報を学習データとして用いることで、未知の関係の予測 (Link Prediction) を行う。例えば、ある人物に対して、国籍の記述がなかったとしても、生まれた場所の記述があれば、その情報を元にして、精度高く国籍の予測ができる。そのために、知識埋め込み手法 (Knowledge Embedding) などの技術が必要となる。

### 2.2 新しい情報の追加

この問題は、新しい知識を知識グラフにどのように追加するかに関する問題である。第2.1章で述べた手法により、新たな情報を追加することも可能であるが、予測に十分な知識が知識グラフ上に存在しない場合もある。そのため、新たな資源から知識を抽出することが必要となるが、データベースのように知識が整理された資源は多くはない。そこで、大量に存在し、新たな情報が随時入手可能なテキスト情報から知識を作ることが望ましいであろう。テキスト情報からトリプルを作成するには、領域に依存せずにテキストから関係を抽出できる技術 (Open IE: Open Information Extraction) や照応解析 (Coreference Resolution) の技術が必要となる。その後、作成されたトリプルを知識グラフに統合する必要がある。統合の際の問題は、2つに分けることができる。一つは、既存のオントロジーで表現できる知識を対象とし、単に欠けている知識を既存のオントロジーで追加する場合である。もう一つは、既存のオントロジーでは、表現できない知識を対象とし、オントロジー自体を変更

連絡先: 市瀬 龍太郎, 国立情報学研究所情報学プリンシプル研究系, 〒101-8430 東京都千代田区一ツ橋 2-1-2, Tel:03-4212-2000, E-mail:ichise@nii.ac.jp

する必要のある場合である。例えば、人物を表現する場合に、出身校を表現する述語がオントロジーとして定義されていない場合は、出身校を定義するようなオントロジー記述を新たに追加しなければ、十分な知識を表現できない。前者の場合には、オントロジーマッチング技術 (Ontology Matching) などにより解決することができるが、後者の場合には、オントロジー学習技術 (Ontology Learning) [市瀬 09] などが必要となる。

### 2.3 誤った知識の混入

この問題は、知識グラフに誤った知識が混入する問題である。実際の Linked Data データでは、このような問題はしばしば見られる。例えば、DBpedia では、ある会社の立地する国名がケルンであると記述されている。しかし、ケルンは国名ではなく地名であり、正しくは、ケルンが存在するドイツとするべきである。このような誤った知識に対応するためには、知識グラフにおける誤りを検知する技術 (Error Detection) や、誤った知識に対して修正する技術 (Error Correction) が必要となる。

## 3. 統合知識基盤

前章で考察したように、精密な知識グラフを作成するプラットフォームは、3つの問題に対処しなければならない。そのため、既存の知識グラフから足りない知識を補完する手法、テキストから知識グラフに欠けている知識を補うための関係抽出手法、得られた知識を統合するためのオントロジー統合手法、そして、正しい知識グラフになるように、誤りを検知、修正する手法を統合する必要がある。

本稿では、これらを統合するための統合知識基盤を提案する。そのアーキテクチャを図 1 に示す。本アーキテクチャは、主に4つのモジュールより構築されている。1つ目は、図 1 の右上にある知識補完 (Knowledge Completion) モジュールで、既存の知識グラフから足りない部分を補完した知識グラフを生成する。この部分は TorusE [Ebisu 18] の技術を用いることができる。2つ目は、図 1 の左上にある関係抽出 (Relation Extraction) モジュールで、テキスト情報からトリプルを作成する。その際に、専門家が作ったブートストラップデータを用いることで、よりの確なトリプルの作成が行える。この部分は、T2KG [Kertkeidkachorn 18] の技術を用いることができる。3つ目は、図 1 の左下にあるオントロジー統合 (Ontology Integration) モジュールで、新たなトリプルを知識グラフに統合する。その際に、既存のオントロジーにない概念が含まれる場合には、オントロジーの学習を行う。この部分は、T2KG [Kertkeidkachorn 18] や FITON [Zhao 14] の技術などを用いることができる。4つ目は、図 1 の右下にある知識検証 (Knowledge Verification) モジュールで、知識グラフに誤りがないかを調べ、状況に応じて修正を行う。この部分は、ALDErrD [Rahoman 16] や誤り修正 [Lertvittayakumjorn 17] の技術などを用いることができる。以上のような統合を行うことで、精密な知識グラフを作成することが可能となる。

## 4. まとめ

本稿では、精密な知識グラフを作成するために必要な課題を整理し、そのための技術について考察を行った。その考察に基づき、知識グラフを生成するためのプラットフォームとして、統合知識基盤のアーキテクチャを提案した。統合知識基盤は、既存の知識グラフに対して、自動的に誤りの検知、修正を行い

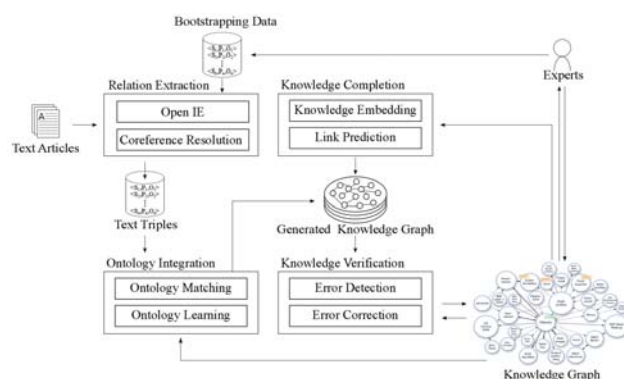


図 1: 統合知識基盤のアーキテクチャ

知識グラフの改善をする機能、知識グラフ上で欠けている知識を補完する機能、テキストから知識を抽出し新たに追加する機能などがあり、精密な知識グラフを作成するための基盤技術として、大きな役割を果たすことが期待される。今後は、第 3 章で述べた技術をモジュール化し、本稿で提案したアーキテクチャで統合するとともに、知識グラフが必要となる特定ドメインを例として、知識グラフを実際に整備することを試みる予定である。

## 謝辞

本研究の一部は、新エネルギー・産業技術総合開発機構 (NEDO) の支援により実施されたものである。

## 参考文献

- [Ebisu 18] Ebisu, T. and Ichise, R.: TorusE: Knowledge Graph Embedding on a Lie Group, in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, AAAI (2018)
- [Kertkeidkachorn 18] Kertkeidkachorn, N. and Ichise, R.: An Automatic Knowledge Graph Creation Framework from Natural Language Text, *IEICE Transactions on Information and Systems*, Vol. 101-D, No. 1, pp. 90–98 (2018)
- [Lertvittayakumjorn 17] Lertvittayakumjorn, P., Kertkeidkachorn, N., and Ichise, R.: Resolving Range Violations in DBpedia, in *Proceedings of the 7th Joint International Semantic Technology Conference*, Vol. 10675 of *Lecture Notes in Computer Science*, pp. 121–137, Springer (2017)
- [Rahoman 16] Rahoman, M.-M. and Ichise, R.: Automatic Erroneous Data Detection over Type-Annotated Linked Data, *IEICE Transactions on Information and Systems*, Vol. E99-D, No. 4, pp. 969–978 (2016)
- [Zhao 14] Zhao, L. and Ichise, R.: Ontology Integration for Linked Data, *Journal on Data Semantics*, Vol. 3, No. 4, pp. 237–254 (2014)
- [市瀬 09] 市瀬 龍太郎: オントロジー学習 - 計算機によるオントロジー構築 -, 電子情報通信学会誌, Vol. 92, No. 9, pp. 791–795 (2009)