The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

未知の属性の問合せに応答可能な対話システムを目指した知識グラフの拡充

Knowledge Graph Expansion for Constructing Dialogue Systems Responding to Queries about Unknown Attributes

藤岡 勇真　　駒谷 和範
Yuma Fujioka　　Kazunori Komatani

大阪大学産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University

Dialogue systems cannot respond to queries including unknown words. Constructing a perfect knowledge base is practically impossible; that is, filling all the attributes in relational databases is quite labor-intensive. We are trying to construct a system that responds reasonably by exploiting existing system knowledge even when a user's query contains unknown attributes. More specifically, we use knowledge graphs as the system's backend and infer unknown attributes of entities by generalizing them with collaborative filtering and tensor factorization. As a first step for it, we extracted a knowledge graph from Wikidata, a commonly-used LOD (linked open data), in the food and dish domain, but it was sparse for generalization. We expand the knowledge graph with external datasets, specifically recipe data contributed to a Web site, and evaluate it by a task to find similar words.

1. はじめに

対話システムは，未知の情報に関する問合せに上手く応答することが出来ない．対話システムに対するよくある入力例として，ある語に関する情報の問合せを考える．このような問合せが入力された場合，システムはその知識ベースから必要な情報を検索し，適切な応答文を生成する．しかし現状の対話システムは，要求された情報が知識ベースに存在しなかった場合，話題を変えたり「わかりません」といった，文脈上不自然な応答をしてしまう．また，必要な情報が全て網羅されている知識ベースを構築するのは非常に困難である．ゆえに，より良い応答の出来る対話システムを目指すうえで，知識ベースに無い情報の問合せへの対応は避けられない課題だと言える．

本研究では，質問された語の情報がシステムにとって未知だった場合でも合理的に応答出来るシステムの構築を目指している．図 1 に目標とするシステムの枠組みを示す.「ミネストローネってイタリアの料理なんですか？」と，ユーザからある語に関する問合せがシステムに入力される．この問合せには，入力語である「ミネストローネ」の「国」という属性に関する情報が必要になる．知識ベースにその情報がないことが判明すれば，システムは既知の情報を用いて，欠損した語の属性や類似した語の推測を試みる．図 1 では,「ミネストローネ」の「国」が「イタリア」であるかどうかを既知の情報から推測し，その結果から「ミネストローネはたぶんイタリアの料理だと思いますよ」という，断定はしないが可能性をほのめかせる応答を行っている．

対話システムにおける入力文中の未知語の処理に取り組んだ研究はこれまでにも行われている [大塚 13][大野 18]．これらの研究では，ユーザへの質問や発話内容の解析によって未知語の情報の獲得を試みているが，本研究では既に知識ベース上にある情報を利用して必要な情報の推測を目指す．

図 1 のような応答生成の実現に向けての第一歩として，システムがバックエンドで持つ知識ベースとして知識グラフを使用する．知識グラフとは，グラフ構造を持ったデータベースである．知識グラフはデータ間の様々な関連性を表現するのに長けた知識モデルであり [Angles 08]，近年こういったグラフ構造をもつデータベースが徐々に増えている．

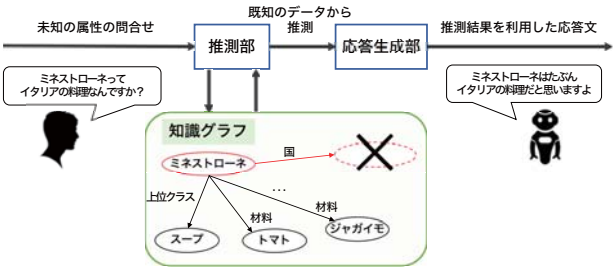

図 1: 目標システムの全体像

本稿では，システムが持つ知識グラフが料理の一般的な情報を持つとして，その構築を行う．まず大規模知識ベースである Wikidata[*1] からのデータ抽出によって知識グラフを構築した．しかし，構築した知識グラフ上には料理の一般的な材料情報が十分でなく，知識グラフがあまりにスパースだと推測や解析が困難になると考えられる．そこで，楽天公開データ [*2] 中の楽天レシピに関するデータを使って，料理の一般的な材料情報を知識グラフの拡充を行う．知識ベースの拡充手法には Web 上のテキストデータからの抽出などがあるが [高橋 07]，本研究では構造化されたレシピデータを用いて信頼性の高いデータの抽出を試みる．拡充の有効性は，入力語に類似した語を知識グラフ上から得るというタスクにより検証した．

2. システムが持つ知識グラフ

2.1 知識グラフ

目標システムは，料理に関する一般的な情報が格納された知識グラフを持つとしてその構築を行う．ここで知識グラフとは，データとデータの間にある関係がグラフ構造によって表現されたデータベースを指す．一般に，グラフ中のそれぞれのノードが 1 つのデータを表現し，ノード間に張られるエッジは

連絡先: 藤岡勇真，大阪大学産業科学研究所，大阪府茨木市美穂ヶ丘 8-1，06-6879-8416，fujioka@ei.sanken.osaka-u.ac.jp

∗1 https://www.wikidata.org
∗2 https://rit.rakuten.co.jp/data_release_ja/

その 2 つのデータ間に何らかの関係があることを示している．また，データ間の構造的な関係を表現するには向きを持つ有向エッジが用いられ，エッジそのものにラベルを付与し関係の種類，すなわちどんな属性であるかを表すこともある．

2.2 Wikidata を用いた知識グラフの構築

Wikidata は，Web の仕組みに基づきリンクされたオープンなデータである Linked Open Data として公開されている．Wikidata は大規模知識ベースの 1 つとして知られ，多言語対応，人手による管理体制などが主な特徴として挙げられる．Wikidata は，Resource Description Framework(RDF) 用のクエリ言語である SPARQL を用いたデータ検索が可能である．

RDF とは，概念や実体を抽象化したリソースの関係を，主語・述語・目的語という三つ組によって記述するデータモデルであり，この三つ組は一般的にトリプルと呼ばれている．RDF は，トリプルを有向エッジ，リソースをエンティティ，述語に記述されたプロパティを有向エッジのラベルとする，ラベル付き有向グラフとして表現することが出来る．SPARQL ではそのグラフパターンを指定することでデータ検索を実現しており，トリプルとしてデータを抽出することが出来る．

Wikidata からの料理に関する情報抽出の手順を示す．はじめに，目標システムが料理としてみなすエンティティの集合である料理群を設定した．具体的には，

1. 『料理，料理のサブクラス，type of food or dish』のインスタンス
2. 『料理，type of food or dish』のサブクラス

という 2 つの条件を 1 つでも満たすエンティティの集合である．

この条件を満たすエンティティを Wikidata から収集すると，その数は 3278 個であった (重複を除く)．

そして，取得した料理群の要素となるエンティティ(以下，料理エンティティと記述する) が主語もしくは目的語となっているトリプルを全て取得した．ここで取得した 110255 個のトリプルには，他のデータベース上における固有 ID や画像・発音音声へのリンクなどといった，エンティティ間の関連性を表さないプロパティを持つものが多く存在していた．そのため，前述のようなプロパティを除く 10 種のプロパティを持つ 11495 個のトリプルによって知識グラフを構築した．

3. 楽天公開データを用いた知識グラフの拡充

前節で構築した知識グラフは，一般的な料理の材料情報が不足していた．そこで，楽天公開データ中の料理レシピ検索サイト「楽天レシピ *3」に投稿された約 80 万件分のレシピデータから，料理の一般的な材料の情報を抽出して知識グラフに拡充する．

このレシピデータはレシピ名・料理名・カテゴリ・材料などの情報が各レシピごとにデータ化されたものである．同じ料理名が記述されたレシピでも，使っている材料はレシピ毎に異なる．またレシピに記載されている材料の中には，地域性や投稿者のアレンジなどに由来する，その料理の材料としては一般的でない材料が含まれている場合もある．ゆえに，料理名に結びつけられた使用材料をそのまま全て知識グラフに拡充することは，一般的な料理の材料情報の拡充という目的にそぐわない．そこで，ある料理で使われている材料 m の一般性が高いほど，その料理のレシピにおける材料 m の使用率が高いと仮定し，使用率に基づく拡充対象の選別を行う．

*3 https://recipe.rakuten.co.jp/

表 1: 楽天レシピデータにおける「オムライス」の材料使用率

使用された材料名	使用率 U_{rate} (%)
「卵」	79.9
「ケチャップ」	69.9
「ご飯」	44.7
「玉ねぎ」	38.4
……	……
「しいたけ」,「ほうれん草」	1.6
「豆乳」,「みりん」,「ソース」	1.5
「塩麹」,「とろけるチーズ」,「冷凍コーン」	1.3
……	……

「料理名が c でその材料が m である」という情報を知識グラフに拡充する手順は，大きく分けて次の 3 つのステップからなる．

1. 全てのレシピデータから，料理名と使用材料の組を抽出する．
2. 料理名 c のレシピにおける材料 m の使用率 $U_{rate}(c,m)$ を算出する．
3. 一定の条件を満たす c,m の組を知識グラフへの拡充候補とし，そうでない場合は拡充候補から棄却する．

まず，第 1 ステップである料理名と使用材料の組の抽出について述べる．料理名 c における材料 m の使用率を求めるために，各レシピデータから料理名と材料の組を抽出する．i 番目のレシピデータに記述されている料理名 $c^{(i)}$ と使われている材料からなる集合 $M^{(i)}$ を抜き出し，i 番目のレシピデータの料理名・材料の組を表す集合 $S^{(i)} = \{c^{(i)}, M^{(i)}\}$ を抽出する．この $S^{(i)}$ の作成を，$i = 1, 2, ..., 796067$ について行う (796067 はレシピ総数)．

第 2 ステップである，料理名 c のレシピにおける材料 m の使用率 $U_{rate}(c,m)$ の算出について述べる．料理名 c における材料 m の使用率 $U_{rate}(c,m)$ を式 (1) のように定義する．

$$U_{rate}(c,m) = \frac{|\{S^{(j)}|c^{(j)} = c, M^{(j)} \ni m\}|}{|\{S^{(j)}|c^{(j)} = c\}|} \qquad (1 < j < n) \tag{1}$$

式 (1) の分母は料理名が c であるレシピの総数を表している．そして分子は，料理名が c であるレシピの中で，材料 m が使用されているレシピの総数を表す．

使用率 $U_{rate}(c,m)$ の実例として表 1 に，c を「オムライス」とした時の，材料ごとの使用率の一部を示す．表 1 は「オムライス」の食材を使用率の昇順に並べたもので，使用率が高い食材と低い食材の一部を抜粋している．使用率が高い材料には「卵」,「ケチャップ」,「ご飯」といった,「オムライス」の一般的な材料が並んでいる．一方で，使用率が低い材料には,「しいたけ」,「豆乳」,「塩麹」等といった「オムライス」の材料としては一般的ではない材料が多く含まれている.「オムライス」の例は，使用率が高いほどその料理に一般に使われる材料であるという仮説を支持する結果になっている．

次に第 3 ステップである，材料情報の拡充の判定について述べる．使用率 $U_{rate}(c,m)$ に対するしきい値を θ とし，$U_{rate}(c,m) \geq \theta$ が成り立つ c,m の組を知識グラフへ拡充する．この判定で，使用率の低い材料が知識グラフに拡充されることを防ぐ．しきい値 θ の値は 4.2 節の調査結果より決定する．

例外として，多くの料理に材料として頻出した「水」,「油」,「砂糖」,「塩」の 4 つの材料は，料理の特徴的な材料としては不適当であるとし，これらの材料を含む c, m の組は拡充対象から除外している．また，Wikidata を用いて構築した知識グラフ上のエンティティの中に，c, m というラベルが付与されたエンティティが片一方でも存在していない場合でも，拡充対象から除外している．これは実装上設けた制限であり，構築した知識グラフ上に新たなエンティティを作成することは行っていない．あくまで Wikidata を用いて構築した知識グラフ上のエンティティ間に，材料の関係を結びつける形になる．

第 3 ステップで述べた条件を満たす全ての c, m の組について,「c の材料が m である」という事実を表すトリプルを作成し，知識グラフにトリプルを拡充する．以上の処理によって，楽天公開データから料理の材料情報をトリプル形式で抽出し，Wikidata と楽天公開データの料理に関する情報を統合する．

4. 知識グラフ上での類似語推測性能の調査

4.1 調査方法

知識グラフの拡充の効果を，入力語に類似した語を知識グラフから得るというタスクにより検証する．一般的な料理の材料情報が楽天公開データから拡充されていれば，その情報によって知識グラフ上の料理エンティティは正しく特徴付けられる．よって拡充を行わない場合に比べて，容易に知識グラフ上から類似語を得ることが出来ると考えている．反対に一般的でない料理の材料情報が拡充されていれば，料理エンティティが一般的でない情報で特徴付けられ，類似した語を得るのが困難になると考えている．

入力語に類似した語をシステムが見つけられているかどうかの指標として，類似度に基づくランキングにおける順位を用いる．この調査では，テストセット中の 30 種の料理を入力語と仮定している．ここでテストセットとは，30 種の料理と各料理に類似した料理を 1 から 3 種人手で選定してペアとして結びつけた 30 個のペアを指す．類似した料理の合計は 57 個となった．類似した料理を入力語に対する類似語とし，入力語と紐づけた類似語を類似度ランキング上で高く順位付け出来るかで構築した知識グラフの評価を行う．

類似度ランキングの作成に際して，知識グラフ上の各エンティティを表現するベクトルを定めた．知識グラフ上のエンティティを表現するベクトルは，知識グラフにおけるエンティティ間の接続関係を 1 と 0 で表した隣接行列を用いて表現した．

入力語を表現するベクトルと料理エンティティを表現するベクトルに関するコサイン類似度に従ってランキングを作成する．ランキング対象とする料理エンティティは，エンティティを表現するベクトルの要素の総和が 4 以上かつ Wikidata から取得したエンティティのラベル情報に，日本語のラベルが含まれている 513 個の料理エンティティである．

4.2 類似度ランキングにおける順位による評価としきい値の分析

テストセット中の入力語とする 30 種の料理全てに対して類似度ランキングを生成した．類似度ランキングの生成は，拡充していない知識グラフと，$U_{rate}(c, m)$ に関するしきい値 θ に基づいて拡充した知識グラフに対して行った．θ は 0 から 0.28 まで 0.01 刻みに変化させ，調査結果からしきい値を決定した．θ の変化に対する，作成した類似度ランキングにおける似た料理全 57 種の平均順位，および拡充されたトリプルの数の変化を図 2 に示す．

θ が 0 から 0.03 に増加するにつれて平均順位は高くなっていき，$\theta = 0.03$ の時に平均順位が最高になった．$\theta > 0.03$ の区間では，θ が増加するにつれて緩やかに平均順位が低下する傾向が見られた．$\theta = 0$ として 7685 個全てのトリプルを知識グラフに拡充するよりも，$\theta = 0.03$ として使用率の低い材料を除外した 4035 個のトリプルを知識グラフに拡充した方が平均順位が高くなるということがわかる．拡充するトリプルを約半数削っているにもかかわらず平均順位は高くなっており，使用率の低い材料の情報を知識グラフに加えることが，類似語の推測性能に悪影響を与えることがわかる．料理の一般的でない材料情報によって，料理エンティティが特徴付けられたためだと考えられる．

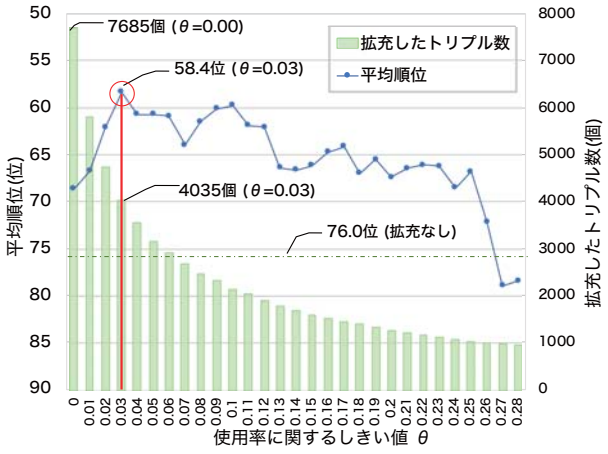

図 2: 類似度ランキングにおける平均順位と拡充トリプル数の θ に対する変化

そして，しきい値 θ が高くなるにつれて，拡充トリプル数の減少量が少なくなっている．このことから，使用率が高い材料の種類数に比べて，使用率が低い材料の種類数の方が多いことがわかる．

また，楽天公開データからのデータ拡充をしていない知識グラフでの結果は 76.0 位であったが，$\theta < 0.27$ においてはその値を上回っており，拡充の有無による差が見られる．

表 2: 類似度ランキングにおける類似語の順位の平均・標準偏差

	平均順位 (位)	標準偏差 (位)
拡充あり ($\theta = 0.03$)	58.4	76.7
拡充なし	76.0	68.1

表 3: 拡充結果の詳細 ($\theta = 0.03$)

構築された知識グラフの総トリプル数	15530 (個)
楽天レシピを用いて拡充したトリプル数	4035 (個)
材料情報を拡充した料理の種類数	573 (種)
拡充した料理の平均追加トリプル数	7.0 (個)

以上の調査から θ を，平均順位が最も高くなる 0.03 に決定して拡充を行った．拡充前後の平均順位，標準偏差を表 2 に示す．拡充後の平均順位の値が拡充前に比べ 17.6(位) 上昇しており，拡充の効果が見られる．一方，ランキング対象とした料理エンティティが 513 個であることを考慮すると，58.4 位と

いう平均順位はさほど高精度な結果ではないと考えられる．拡充結果の詳細を表 3 に示す．結果として楽天公開データから，4035 個のトリプルを知識グラフに拡充した．

4.3 類似度ランキングの実例

類似度ランキングの実例を示し，類似した料理が想定通りに推測出来た例とそうでない例での傾向の差異について述べる．

表 4: 「ボルシチ」に関する類似度ランキング

順位	料理名
1	◆ミネストローネ
2	ヴィシソワーズ
3	ハヤシライス
4	シチー
5	◆ビーフストロガノフ
6	カレーライス
7	ポトフ
8	クリームシチュー
.....	.....

まず，類似した料理が想定通りに推測出来た例として,「ボルシチ」に関する類似度ランキングから，上位 10 位までを抜粋したものを表 4 に示す．表 4 内で太字で示されている「ミネストローネ」と「ビーフストロガノフ」は，テストセット上で「ボルシチ」に類似した料理として結びつけられた料理である.「ミネストローネ」は 1 位,「ビーフストロガノフ」は 5 位に順位付けられており，想定通りの推測結果が出ている．上位 10 位以内に順位付けされている他の料理についても，煮込み料理が主に並んでおり，概ね似ていると思われる料理が多い．

表 5: 「トルティージャ」に関する類似度ランキング

順位	料理名
1	バカリャウ・ア・ブラース
2	オリヴィエ・サラダ
3	フラットブレッド
4	フォカッチャ
5	ビリヤニ，ザルツブルガーノッケルン
7	マギリッツァ，けんちん汁
.....	.....
47	◆オムレツ，他 22 種
.....	.....
133	◆卵焼き，他 7 種
.....	.....

反対に，類似した料理が上手く推測出来なかった例として,「トルティージャ」に関する類似度ランキングから，一部を抜粋したものを表 5 に示す．表 4 と同様に,「オムレツ」と「卵焼き」はテストセット上で「トルティージャ」に類似した料理として結びつけた料理である．ランキングにおいて「オムレツ」は 47 位,「卵焼き」は 133 位に順位付けられており，類似した料理を想定通りに推測出来ていない．上位に順位付けられている料理も，サラダや汁物，パンなどが見られ，似ていると思われる料理が少ない．

「ボルシチ」と「トルティージャ」の類似度ランキングを見比べてみると,「トルティージャ」の類似度ランキングには同率に順位付けされている料理が多い傾向にあるということがわかる.「トルティージャ」の類似度ランキングを見ると，5 位,7 位,47 位,133 位には同率に順位付けされている料理エンティティが存在する．この原因として考えられるのは，定めたベクトルが持つ情報量の少なさである．表 6 に,「ボルシチ」と「トルティージャ」，そしてテストセット上でこの二つの料理と結びつけた料理エンティティを表現するベクトルの要素の和を示す．料理エンティティを表現するベクトルの要素の和は，結びつく他エンティティの数を表す.「ボルシチ」とそれに紐づけた料理エンティティのベクトルの要素の和に比べ,「トルティージャ」とそれに紐づけた料理エンティティのベクトルの要素の和は小さい傾向にあることがわかる.「トルティージャ」やそれに紐づけた料理エンティティを表すベクトルに，他エンティティとの差を表す情報を十分に組み込めておらず，類似性という点でも，単純なコサイン類似度では差がつかない料理が同率順位に並んでいるのだと考えられる．

表 6: 料理エンティティを表すベクトルの要素の和

ボルシチ	27	トルティージャ	7
ミネストローネ	24	オムレツ	11
ビーフストロガノフ	27	卵焼き	20

5. おわりに

未知の属性の推測にむけて，今後は構築した知識グラフ上における関係を推測する手法の確立を目指す．またエンティティ間の接続関係だけでなく，プロパティなど他の情報を考慮した推測が効果的かどうか検証する．得られた推測結果を応答生成に反映させる枠組みについても検討する必要がある．

謝辞

本研究では，楽天株式会社が国立情報学研究所の協力により研究目的で提供している「楽天公開データ」を利用した．

参考文献

[Angles 08] Angles, R. and Gutierrez, C.: Survey of graph database models, *ACM Computing Surveys*, Vol. 40, No. 1, pp. 1–39 (2008)

[高橋 07] 高橋 潔考, 鈴木 遼, 下村 芳樹, 舘山 武史, 吉岡 真治, 武田 英明 : Web 情報を用いた設計知識情報データベースの拡充手法, 横幹連合コンファレンス予稿集, Vol. 2007, pp. 45–45 (2007)

[大塚 13] 大塚 嗣巳, 駒谷 和範, 佐藤 理史, 中野 幹生 : データベース検索音声対話システムにおける店舗属性値取得のための質問生成, 人工知能学会第 27 回全国大会 (2013)

[大野 18] 大野 航平, 武田 龍, ニコルズ エリック, 中野 幹生, 駒谷 和範 : 対話を通じた未知語のクラス獲得に向けた暗黙的確認の提案, 人工知能学会論文誌, Vol. 33, No. 1, pp. 1–10 (2018)