Wikidata からの遠距離教師あり学習に基づく大規模関係知識獲得

松田耕史 [1]　鈴木正敏 [1]　乾 健太郎 [1,2]
[1] 東北大学　[2] 理化学研究所 AIP センター
{matsuda,m.suzuki,inui}@ecei.tohoku.ac.jp

1 はじめに

大規模なテキストデータから，構造化された知識を発見するタスクを，知識獲得タスクという．知識獲得は言語処理における中心的なタスクであるとともに，言語処理における基礎解析にとって逆側の車輪とも呼べる．つまり，知識獲得によって得られた知識を用いて，各種の言語処理タスクの性能向上を目指すことが可能である．同様に，精緻な言語処理によって，知識獲得のカバレッジを向上させることが可能になるだろう．構造化された関係知識ベースの活用は，質問応答 [5]，チャットボット [11]，含意関係認識 [2] などの応用タスクにおいて性能の向上に寄与することが報告されているが，関係知識ベースの自動的な構築や拡張は依然として重要な課題である．

本稿においては，Wikidata [10] と呼ばれる知識三つ組で構成された多言語知識ベースからの遠距離教師あり学習によって，Wikipedia の本文から知識三つ組を大規模に獲得する，言語非依存な方法を提案する．具体的には，Wikipedia アブストラクト内のアンカーリンクのうち，Wikidata 内の知識三つ組とマッチするものに対して，関係を表す意味ラベルを自動付与することで大規模な学習データを生成する．このデータから学習したモデルを関係ラベルが未知のアンカーに対して適用することで知識獲得を行い，その質を評価した．提案する手法は特定の言語に依存しないので，どのような言語でも適用できる可能性があるが，本稿においては，本手法を日本語版の Wikipedia アブストラクトに適用した結果，約 180 万タプルに及ぶ新たな知識を 74% の精度で抽出することができたので，その詳細を報告する[*1]．

[*1]知識ベースのビューワーは http://www.cl.ecei.tohoku.ac.jp/kbsearch/ で公開している．

2 関連研究

テキストからの知識獲得問題は古くから取り組まれている問題である．そのためのアプローチは，主に，教師ありの手法，教師なしの手法，弱教師ありの手法に大別される [7]．教師ありの手法は一般に高い性能を達成可能である反面，教師データの作成に非常に大きなコストがかかる．近年では，クラウドソーシングに基づいた教師データの作成も模索されている [4] が，大規模な教師データの作成は依然として現実的とはいえない．教師なしの手法として，パターンベースの方法や，語の頻度分布にもとづく方法が知られている [8] が，is-a, part-of といった個別の関係ごとにパターンを開発しなければならず，多様な関係を網羅することは現実的ではない．そこで我々は，弱教師ありの手法に着目する．

具体的には，Wikidata と呼ばれる, 大規模なデータベースから遠距離教師づけ (Distant Supervision[6]) を行うことで，大規模な教師データの生成を実現する．Wikipedia の本文と Wikidata のアブストラクトのアラインメントについては，最近 Elsahar らの研究がある [3]．彼らの提案はデータセットの構築方法に関するものであり，今回は同様の発想で生成したデータからの関係抽出器の実装と評価に焦点をあてている点に差異がある．

3 Wikidata

Wikidata[*2] は，自由・共同作業・多言語・二次情報を特徴とする、構造化データのデータベース（知識ベース）である [10]．すべてのエンティティは Q から始まるエンティティ ID で識別され，エンティティ間の関係は P から始まるプロパティによって記述されている．

すべての知識は，(s(*subject*), p(*property*), o(*object*)) からなる ID の三つ組で構成されている．たとえば,「地球は太陽系の一部である」という事実（fact）を Wikidata

[*2]https://www.wikidata.org/wiki/Wikidata:Main_Page

上の表現で表すと，(*Earth* (Q2), part of (P361), *Solar System* (Q544)) となる．ここでは理解を容易にするために英語のラベルを付与したが，Wikidata 上の知識は，日本語や英語といった特定の自然言語に依存していないという特色を持っている．エンティティの名前も関係の名前も，単に「ある言語におけるラベル」であり，エンティティには様々な言語で表現された名前が付与されている．この多言語対応機能は Wikipedia の言語間リンク機能をベースにしており，Wikipedia 上で言語間リンクが存在する記事のクラスタが一つのエンティティに対応している．現時点において，16 億（大半はエンティティの言語ラベルであり，関係知識と呼べるものは 10% 程度である）におよぶ事実が記述されており，現在も盛んにアップデートが続いている．

4 提案手法

関係分類器の学習と適用は，Wikipedia のアブストラクトに含まれるアンカーリンクの単位で行う．知識獲得タスクにおいては，対象とする知識の粒度が問題となるが，アンカーリンクという単位は明示的にマークアップされており，かつ曖昧性の解消が不要であるという利点がある．

4.1 緩和された Distant Supervision による教師データの生成

典型的な Wikipedia の文においては，省略された主語は記事のトピックエンティティを指しているとみなせることが多い．通常の Distant Supervision においては，ある文内に共起しているエンティティペアを用いて関係を検索するが，Wikipedia の文の性質を用いると，1 文中に共起しているという Distant Supervision の制約を緩和することが可能である．緩和された設定においては，以下のような流れでラベル付きデータの生成を行う．

アブストラクトに含まれる文 S, アンカーテキスト a に対して，以下の処理を行う．記事のタイトルが表すエンティティ s とアンカーリンク先が指すエンティティ o で，$(s, *, o)$ となるようなプロパティ p を Wikidata から検索する．プロパティが存在した場合，(s,p,o,S) からなるタプルをアンカー分類器の学習データに追加する．存在しなかった場合は，モデルの適用を行うテストデータに $(s,\mathtt{NIL},o,S)$ というタプルを追加する．

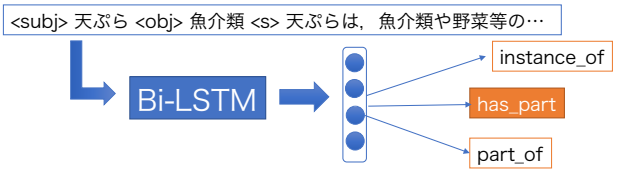

図1: 事例のエンコードと分類器の構成

4.2 アンカー分類器による未知のアンカーリンクテキストからの知識抽出

今回の問題を，文中のアンカーリンクで示されるスパンを関係ラベルに対して分類するマルチクラス分類問題であるとみなした．具体的には，図1のように主語エンティティと目的語エンティティ，文をセパレーターで連結した系列を作成し，この系列に対する分類問題として解く．

上記の方法で抽出された大規模な訓練データを用いてアンカーテキストの分類器を訓練し，分類器を，4.1の手続きによってプロパティラベルが付与されなかったアンカーに対して適用する．結果として，プロパティラベル上の確率分布が得られるが，特定のプロパティラベルを持つ確率がしきい値以上であるアンカーを，あらたに知識ベース上に追加する．

5 実験設定

5.1 データ

日本語版の Wikipedia は 2018/09/03 版の Circusserach ダンプに対してアンカーを付与したものを用いた[*3]．また，Wikidata は，2018/10/06 版の RDF ダンプを用いた．前処理として句点を基準に文の分割を行った．この際，文長があまりに長い場合は経験的に箇条書きなどのパーズエラーが含まれている場合があるので，500 文字を超える長さの文はラベル付けの対象から除外した．

5.2 アンカー分類器

任意の系列分類器が利用可能であるが，今回は双方向 LSTM モデルを利用した．前方から後ろ向きに LSTM でエンコードした際に得られた最終状態の隠れ変数ベクトル v と，後ろから前向きにエンコードした際に得られた最初の隠れ状態のベクトル v' を連結し，事例のベクトル表現を得た．このベクトル表現を線形層と softmax 層に通すことによって，プロパティラベルの確率分布を得る．

[*3] Wikipedia 本文の前処理に用いたスクリプトは https://github.com/singletongue/WikiCleaner で公開している

学習されたモデルを新たな事例に適用して知識獲得を行う際には，特定のラベルの出力確率が閾値 (今回は 0.7 を用いた）を超える場合に，主語，目的語ペアがそのラベルを持つとみなすことにした．どのラベルも確率値が閾値を超えなかった場合は，用いた関係語彙の中には適切な関係が存在しなかったという意味を表す，NIL という特別なラベルを付与することにした．

LSTM のユニット数は 300，バッチサイズは 64，最適化には ADAM を用いた．

6 実験結果

前節で述べたラベル付きデータの生成法を日本語版 Wikipedia のアブストラクト全体に対して適用した結果，82 万事例のラベル付きデータが得られた．ラベルの種類数は 1169 種類であった．この事例集合からアンカー分類モデルを学習し，680 万事例の（Distant Supervision によって NIL ではないプロパティラベルの付与が行われなかった）アンカーリンクに対して適用した．

6.1 自動ラベルが付与された事例のノイズ

Distant Supervision に基づいてラベル付与された事例にどれほどのノイズが含まれるかを調査するために，NIL ではないプロパティラベルが付与された事例を 200 事例目視で分析して，タプルと文のアラインメントが妥当に行われているかを確認した．その結果，明らかに妥当ではないプロパティラベルの付与が行われているペアの数は 200 例中 5 例だった．具体例を表1に示す．

ほとんどの場合，我々が提案する緩和された Distant Supervision は妥当なプロパティラベルを付与できているが，1 番目の例のように，文中の主語がタプルの主語エンティティと同一であるという仮定が強すぎる例が散見される．また，2, 3 例目のように，文から含意されない誤った関係を出力する例もいくつかあったが，全体として数は少なかった．

6.2 完成した知識ベースの概要と精度評価

680 万事例の未知のアンカーテキストに対してモデルを適用し，特定のラベルの出力確率が 0.7 以上の事例を取り出すと，約 180 万事例の新たなタプルが得られた．モデルの訓練がどの程度成功しているかを見積もるため，一部のデータを開発データとしてヘルドアウトしておき，開発データにおける正解率を測定したところ，エンコーダーに CNN を用いた場合は 0.79 ，エンコーダーに双方向 LSTM を用いた場合は 0.82 であった．この

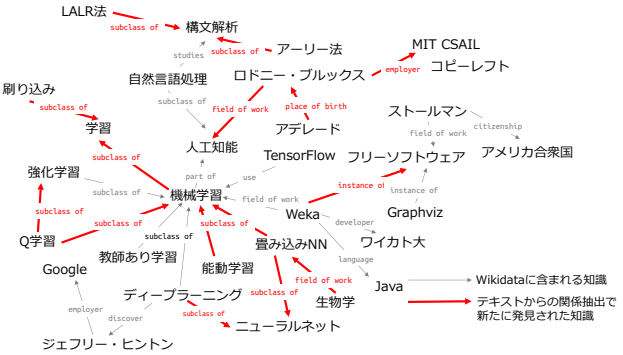

図2: Wikipedia のアブストラクトから自動獲得された知識の例: 赤色で表示されたエッジが元の Wikidata には含まれない知識である．

値を絶対的に評価することは難しいが，ラベルの種類数の多い問題であるものの，ある程度の性能で学習できていることがわかる．

実際の Wikipedia のアブストラクトから抽出された知識の例を表2に示す．ドラゴンドラというエンティティに関する知識は元の Wikidata には 1 件も含まれていないが，テキスト中に表現されている知識がアンカーテキストという粒度で獲得できていることが見てとれる．

実際にどのようなタプルが獲得されているかを，「機械学習」というエンティティを中心に取り出した例が図2である．元の Wikidata にはスパースで記述されていなかった知識が，テキストから得られる知識に基づいて補完されている様子がわかる．

獲得された知識の正確性を評価するために，新たに抽出されたタプルから 200 件をランダムに取り出し，以下のような条件を満たすかどうか，目視で評価を行った．

現実世界における知識の真偽は問わずに，文のみからタプルで表される知識が含意されるか．という基準で知識ベースを評価した場合，73.5% の事例が原文に記述されている関係を正しく含意していると評価された．誤りのパターンとしては，(i) Wikidata には，適切な関係が存在しないため，NIL を出力することが正しい，(ii) 生誕日と死没日など，似てはいるが異なる関係を出力してしまう，といった類型がみられた．

6.3 応用例: 雑談対話ボットの知識源として

本論文の手法で構成された知識ベースの一部分を，雑談対話チャットボットシステムの知識源として利用した例が Zunkobot [11] である．具体的には，自動構築された知識ベースから，ヒューリスティクスに基づいて料理関係の知識が記載された部分を抽出し，ユーザーの発話に含まれる食材名や料理名をトリガーとして知識ベ

表1: Distant Supervision に基づくアンカー付与が誤っていた事例

主語エンティティ (記事タイトル)	Wikidata から得られた関係	本文 (下線部が目的語エンティティ)
梅村清弘	educated at	父は中京大学 学長・梅村清明.
田島泰彦	educated at	上智大学 文学部新聞学科教授.
熊本県道 194 号和仁菊水線	owned by	熊本県道 194 号和仁菊水線は、熊本県 から...(中略)... に至る一般県道である.

表2: 学習したモデルによるテキストからの知識獲得の例

主語エンティティ (記事タイトル)	本文 (下線部が目的語エンティティ)
ドラゴンドラ	ドラゴンドラは、新潟県 (headquarters location) 南魚沼郡 (located in) 湯沢町 (located in) 苗場スキー場 (NIL) と かぐらスキー場 (NIL) の田代エリアを結ぶ、プリンスホテル (operator) が運営する索道 (instance of) である.

ースを検索し，「**ペスカトーレ** といえば **魚介類** ですよね」といった，会話の継続と話題の転移を狙った発話を行う．知識ベースの有無を切り替えた実験を行っていないため，純粋に知識ベースの有効性に関する評価は行っていないが，当該ボットはリアルタイムコンペにおいて高い評価を得た[*4].

7 まとめと今後の課題

Wikidata からの遠距離教師あり学習を行うことにより，Wikipedia アブストラクトからの関係知識獲得を大規模に行うことができた．

今回は学習データの生成にも，実際の知識抽出にも Wikipedia のアブストラクトを用いたが，本文からの知識抽出も興味深い課題である．今回は単一ラベルの問題として定式化したが，エンティティペアは複数の関係を持つ可能性があるため，これは近似であり，より現実に即したモデリングがありえる．また，1節で述べたように，Wikidata は特定の言語に依存しないデータベースであるので，本手法も言語に依存せず有効なことが期待される．今後は多言語を同時に解析することでより信頼性の高い知識を獲得することが可能なモデルも検討したい．また，知識ベースを連続空間に埋め込むモデル（[1, 9] など）との統合も興味深い方向性である．

謝辞

本研究は JST CREST(課題番号: JPMJCR1301) の支援を受けて行った．

[*4] https://dialog-system-live-competition.github.io/dslc1/result.html

参考文献

[1] Antoine Bordes et al. “Translating Embeddings for Modeling Multi-relational Data”. In: *NIPS 26*. 2013, pp. 2787–2795.

[2] Qian Chen et al. “Neural Natural Language Inference Models Enhanced with External Knowledge”. In: *Proceedings of the 56th Annual Meeting of the ACL*. 2018, pp. 2406–2417.

[3] Hady Elsahar et al. “T-REx: A Large Scale Alignment of Natural Language with Knowledge Base Triples”. In: *Proceedings of the LREC 2018*. Ed. by Nicoletta Calzolari (Conference chair) et al. May 2018.

[4] Angli Liu et al. “Effective Crowd Annotation for Relation Extraction”. In: *Proceedings of the NAACL-2016*. Association for Computational Linguistics, 2016, pp. 897–906.

[5] Todor Mihaylov and Anette Frank. “Knowledgeable Reader: Enhancing Cloze-Style Reading Comprehension with External Commonsense Knowledge”. In: *Proceedings of the 56th Annual Meeting of the ACL*. 2018, pp. 821–832.

[6] Mike Mintz et al. “Distant supervision for relation extraction without labeled data”. In: *Proceedings of the Joint Conference of the ACL-IJCNLP*. 2009, pp. 1003–1011.

[7] Sachin Pawar, Girish K. Palshikar, and Pushpak Bhattacharyya. “Relation Extraction : A Survey”. In: *CoRR* abs/1712.05191 (2017). arXiv: 1712.05191.

[8] Stephen Roller, Douwe Kiela, and Maximilian Nickel. “Hearst Patterns Revisited: Automatic Hypernym Detection from Large Text Corpora”. In: *Proceedings of the ACL*. Association for Computational Linguistics, 2018, pp. 358–363.

[9] Ryo Takahashi, Ran Tian, and Kentaro Inui. “Interpretable and Compositional Relation Learning by Joint Training with an Autoencoder”. In: *Proceedings of the ACL*. 2018, pp. 2148–2159.

[10] Denny Vrandečić and Markus Krötzsch. “Wikidata: A Free Collaborative Knowledge Base”. In: *Communications of the ACM* 57 (2014), pp. 78–85.

[11] 阿部香央莉 et al. “Zunkobot：複数の知識モジュールを統合した雑談対話システム”. In: *SIG-SLUD* B5.02 (Oct. 2018), pp. 112–117.