建築オントロジーの構築に向けた基礎的検討
埋め込み表現を用いた階層関係の抽出
A Fundamental Study on Construction of Architectural Ontology
Extraction of concept hierarchies using embeded representation

○恒川　裕史[*1]
Hiroshi Tsunekawa[*1]

*1 株式会社竹中工務店技術研究所　主任研究員
Chief Researcher, Research and Development Institute, Takenaka Corporation

キーワード：オントロジー; 自然言語処理; ディープラーニング; 機械学習
Keywords: Ontology; natural language processing; deep learning; machine learning.

1. はじめに

第 3 次人工知能ブームの中、建築分野でも人工知能の活用が進んでいる。その中心は機械学習、特にディープラーニング（DL）と呼ばれるニューラルネットの技術である。機械学習では、いわゆるビッグデータを利用・分析することで判断や最適化を行う。ところが、実際にこうした手法を建築に応用しようとすると、データが思うように集まらないといった障害が起こる。DL で想定するデータ数の規模が数万件から数十万件なのに対して、実務で集められるデータの数は圧倒的に足りない場合が多い。一方、建築の分野では、これまでに積み重ねられてきた経験や研究の成果を技術資料や指針の形でまとめていることが多く、これを人工知能が理解してくれれば、わざわざ機械学習でデータから学習するまでもない。

こうした技術資料は人間が読んで理解できるテキストとして作成されており、このようなテキストを対象にした人工知能技術を自然言語処理と呼んでいる。DL の登場に伴い、近年、この分野の進歩は著しく、それまで使い物にならなかった自動翻訳が実用レベルになるなど、注目を集めている。人工知能が技術資料を理解した場合に期待されることの一つに、ユーザの質問に答えてくれることがあり、これは質問応答と呼ばれている。Devlin らが提案した BERT[1]は事前の教師なし学習で獲得したネットワークを用いた転移学習で高い性能を発揮し、SQuAD[2]というデータセットの中の限定されたタスクでは人間の平均スコアをも超えたと話題になり、その後も XLNet[3]など新しい手法が提案されている。

このように技術進歩の著しいニューラルネットによる自然言語処理であるが、事前の学習のおかげで学習データが少なくて済むとは言うものの、依然として大量の学習データが必要である。しかも現在の技術は決してテキストの内容を理解しているわけではなく、間違いが事故や災害に直結する建築設計生産に使うには不安が大きい。

人間の知識を人工知能技術の中で記述するという試みは第2次人工知能ブームの中でエキスパートシステムの開発と言う形で試みられた。これは一定の成果を挙げたものの、知識のメンテナンス・再利用が難しいなどの原因で下火になっていった。この時使われたフレームやルールといった知識表現からの反省を経て知識の共有と再利用を実現するための枠組みとして提案されたのがオントロジーである。オントロジーは、「人工システムを構築する際のビルディングブロックとして用いられる基本概念／語彙の体系」と定義されており[4]、全体性、網羅性、体系性を備えた対象表現である[5]。

筆者らは法令の網羅的探索のためにオントロジーを適用しており[6]、今回も法令のテキストを題材にオントロジー構築のための検討を行ったので報告する。

2. 検討の方針

オントロジーの半自動的な構築をオントロジー学習と呼ぶ。Figure 1に、オントロジー学習のレイヤー[7]を示す。用語の抽出、同義語の同定、同義語をまとめた概念の形成、その概念の階層の設定、関係の構築とルールの抽出である。ここでは、オントロジー学習の概要をまとめ、本研究での方針を説明する。

Asim ら[8]は、オントロジー学習で使われる技術を、言語学的技術、統計的技術、演繹論理プログラミングに分けて

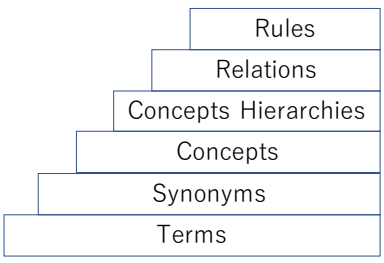


Figure 1. Ontology learning layer cake[7]

整理している。まず非構造テキストに対して前処理を行うが、ここでは言語学的技術が活躍する。構文解析を行い品詞などの情報を付加する。この際、日本語の場合は形態素に区切るという作業も必要になる。抽出された用語から stop word と呼ばれる不要な単語を除外し、活用形を整理した上で見出し語を抽出する。

前処理が終わると次は用語および概念の抽出である。言語学的技術では、構文解析の結果を用いてこれを実現する。用語候補から用語を選ぶには、用語らしさを測るスコア付けが必要であり、ここで統計的技術が使われる。概念は意味の似通った用語、関連のある用語の集合であり、関連する用語は文書の中で共通して現れるということを前提として共起解析が行われ、例えば相互情報量などの測度を用いて用語間の関連性を推定する。概念の形成につながる類似語を探す方法としては、用語文書行列の特異値分解を行い、主要な関連構造を抽出する LSA[9]や、ベイズ推定を用いる LDA(Latent Dirichlet Allocation)、コサイン類似度などを用いた階層型クラスタリングや、k-means 法などの非階層型クラスタリング手法も用いられる。階層型クラスタリングは、用語・概念抽出の後に行われる関係抽出でも使われる。関係抽出では、マーケットバスケット分析で知られる ARM(Association Rule Mining)と言う方法も使われる。

本研究では、ニューラルネットによる自然言語処理技術の進展を踏まえて、オントロジーの最大の特徴である用語・概念の階層関係抽出に焦点を当てる。具体的には、建築関連の法令文から用語・概念の抽出を行った後、前述した BERT による階層関係の抽出を試みる。BERT では事前の教師なし学習と後半の教師あり学習でともにニューラルネットを使うが、本研究では事前学習で得られた単語の埋め込み表現を使い、教師あり学習には別の統計的手法を用いた。なお、教師あり学習の教師データには、建設情報標準分類体系（JCCS）[10]を用いた。

3. 用語の抽出

法令データは、電子政府の e-Gov 法令検索から XML 形式でダウンロードし、タグや目次を削除したものを用いた。e-Gov 法令検索では事項別分類索引から「建築・住宅」を選択し、検索された 128 件の法令を使用した。検索された法令には、建築基準法、建築士法、建設業法などの基本的な法令から住生活基本法、公営住宅法などが含まれており、全体で 8 万行、8MB 弱のテキストが得られた。作成したテキストファイルを形態素解析器である Chasen[11]にかけ、形態素への分割と品詞の付与を行った。辞書は ipadic[12]を使用した。この Chasen の出力を専門用語自動抽出システム termex[13]にかけて用語を抽出した。termex では、単名詞と複合名詞を対象として、重要度を計算して用語を抽出する。以上の作業により、13,340 語の用語が抽出された。Table1 に、抽出された用語の例を重要度順に示す。

Table 1. Example of extracted terms

term	importance	frequency
法	55100485440	6371
建築物	24095884051	3576
施行	17057241600	3200
住宅	13167554400	1120
事業	6031067856	779
国土交通大臣	5510884915	3088
建築	4148535240	190
建築基準法	2821634440	386
施行者	2796961827	356

この結果に JCCS の用語を加えたものを用語とした。JCCS の用語は 10,297 語、両者共通の用語は 1,307 語であった。なお、JCCS と抽出した用語とで ipadic との共通の用語はそれぞれ 2,705 語、3,027 語だった。

4. 階層関係の抽出

4.1. 埋め込み表現の作成

埋め込み表現の作成には、柴田ら[14]が公開している日本語 Pretrained モデルを用いた。これは、日本語 Wikipedia 約 1,800 万文を使い、Juman++で形態素解析を行った後で subword に分割し、事前学習したものである。事前学習は、ランダムに置換した単語を推定する Masked LM と、2つのセンテンスが連続したものであるか否かを推定する Next Sentence Prediction により行われる。なお、公開モデルの中で fine tuning タスクでの推定精度が最も高い LARGE WWM 版（1024次元）の24層、16ヘッドのモデルを用いた。

前章で作成した法令文データを Juman++で形態素解析し、Pretrained モデルを用いて名詞のみを埋め込み表現に変換した。Alammar[15]によれば、文脈を考慮した埋め込み表現としては、出力層に近い 4 層を連結して用いた場合に最も性能が良いとのことだが、連結するとモデルが大きく複雑になるため、4 層の平均を用いた。前章で抽出した用語のうち Pretrained モデルに含まれない複合名詞を法令文データから検索し、用語を構成する個々の単語の埋め込み表現を平均したものを当該用語の埋め込み表現とした。word2vec などと異なり、BERT の埋め込み表現は文脈を考慮したものであるため、同じ単語でもすべて表現が異なる。Figure2 は、作成した埋め込み表現のうち JCCS に含まれる

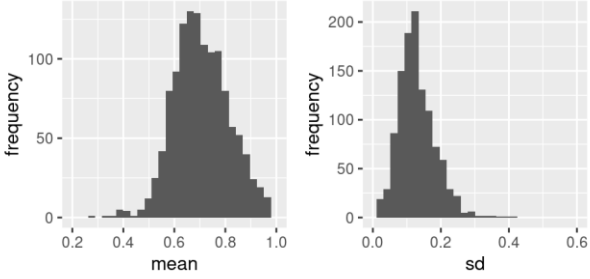


Figure 2. Distribution of similarity of terms(original)

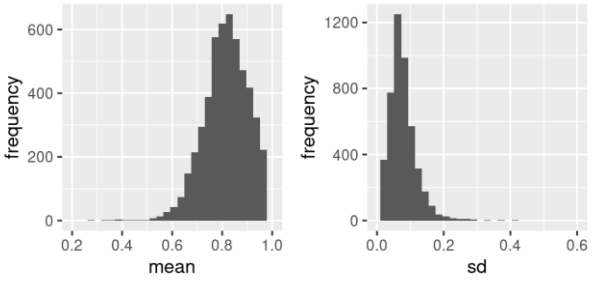


Figure 3. Distribution of similarity of terms(clustered)

用語ごとのコサイン類似度の平均と標準偏差の分布である。図から類似度はある程度高いものの、ばらつきがあり、平均値を取ることが必ずしも良くない可能性が推察される。そこで、文脈による違いを考慮するため、クラスターの数が過大にならない範囲で、できるだけコサイン類似度の最低値が 0.6 以上になるように k-means 法でクラスタリングを行った。類似度の分布を Figure3 に示す。類似度の平均が上がり、標準偏差が小さくなっていることが分かる。

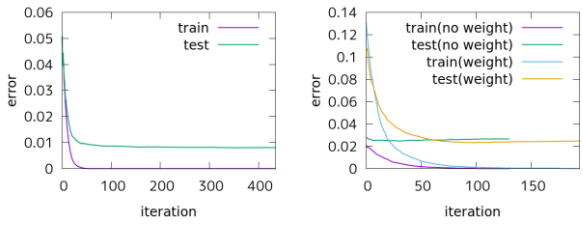
4.2. 階層関係の学習

階層関係の教師データの正例として、前述の JCCS を用いた。JCCS で上下関係にあり、埋め込み表現が得られている用語ペアは 668 語からなる 577 組であった。

正例は 2 種類の方法で作成した。一つは、計算された素の埋め込み表現を使う方法(original)、もう一つはクラスタリングされた単語の平均値を使う方法(clustered)である。ただし、前者は総組合せ数が 43,346,094 組と膨大になるため、位置的に近い単語の組合せを優先させ、同じ単語の組合せを 100 以下に限定することで 53,032 組の正例を作成した。後者は総数で 12,896 組となったため、そのまま採用した。

負例は、埋め込み表現が得られている JCCS の単語 1,703 語の正例以外の組合せ 2,472,189 組の中から、JCCS での上位概念と、それに対応する下位概念以外との組合せ 210,081 組をすべて、それ以外の組合せからは 1 割をランダムにサンプリングした。なお、クラスタリングしたケースでは、選ばれた単語の組合せに対して当該単語クラスターの複数の平均値の中からランダムに選び、クラスタリングしていないケースでは単語全体の平均値を使用した。

BERT などでは後半の教師あり学習もニューラルネットを使っているが、本研究では勾配ブースティングに基づく xgboost[16]を使用した。ブースターは決定木、入力は用語ペアの埋め込み表現、目的変数は階層関係か否かの 2 クラス分類で確率を出力するものとした。各学習データとテストデータを 8:2 の比率で分け、テストデータでのエラーが最も小さくなる繰り返し回数で全データを用いた学習を行い、階層関係予測モデルを作成した。なお、学習/テストデータを分ける際には、同じ用語は同じ側に分けるようにした。また、正例と負例とのデータ数がアンバランスであるので、学習の際重みをつけたもの(weight)とつけないもの(no weight)を試した。Figure4 に学習時の誤差の収束状況を示す。どちらも学習データの誤差はほぼ 0 に近づいているが、テストデータの誤差はクラスタリングしたものの方が大きくなっている。Table2 にテストデータの再現率と精度を示す。original の方がともに良くなっており、クラスタリングしたものでは重みを付けることで正例の再現率が上がっていることが分かる。



(a) Original (b) Clustered

Figure 4. Learning curve

Table 2. Learning results

	TRUE		FALSE	
	recall	precision	recall	precision
original	0.91	1.00	1.00	0.99
clustered	0.25	0.85	1.00	0.98
clustered(weight)	0.40	0.75	1.00	0.98

4.2. 階層関係の抽出

JCCS に定義された概念の更に下位の概念の抽出を試みた。ただし、JCCS では上位下位関係と全体部分関係が区別されておらず、本研究でも両者を含めて階層の中で下に位置する概念を探索することを目的とした。具体的には、前節で作成した階層関係予測モデルに、JCCS での下位概念 566 語を上位、4.1 節で埋め込み表現を作成した用語 10,185 語を下位として階層関係を探索した。クラスタリングモデルでは、クラスタリングした埋め込み表現も候補に加えた。確信度が 0.5 を超えたものを抽出結果とした。

original モデルでは、確信度が 0.5 を超える組合せは抽出されなかった。念のため確信度を下げて確認したところ、0.1 以上で 2 件、上位はともに「生産計画」下位が「消費量」「中小」であり、学習の成績は良かったが、抽出での正解はなかった。

clusterd(no weight)モデルでは、5,100組が抽出された。建設工事を上位として4,561語が抽出され、その中には正解と思えるものもあったが、全候補の半数が抽出されており、自動抽出の狙いにはほど遠いため、建設工事を上位とする組は除外し残りの539組を対象とした。正解は Table3に示す6組のみであった。PRED がモデルの出力である確信度である。平均精度[17]は0.015であった。232の鳥取県知事は、JCCS での都道府県―鳥取県という関係との類似から複合

名詞の鳥取県知事が抽出されたものと思われる。重みを付けたモデルでは、10,566組が抽出され、建設工事を除くと5,241組で、正解は187件であった。Table4に例を示す。JCCSには記録の下位としてカルテ、ビデオなどが挙げられているが、写真と図面が抽出されており、興味深い。また、複数の政令がリストアップされており、成績を押し上げた。平均精度は全体で0.29で、100位までででは0.73となった。

Table 3. Hierarchie extraction results: clustered(no weight)

#	UPPER	LOWER	PRED	PRECISION
88	経理	発注	0.754	0.0114
91	維持	修繕	0.752	0.0220
232	都道府県	鳥取県知事	0.627	0.0129
285	独立行政法人	産業技術総合開発機構	0.591	0.0140
384	区分	分野	0.551	0.0130
423	独立行政法人	独立行政法人国立文化財機構	0.532	0.0142

Table 4. Hierarchie extraction examples: clustered(weight)

#	UPPER	LOWER	PRED	PRECISION
34	記録	写真	0.984	0.706
45	政令	都市計画法施行令	0.980	0.667
50	記録	図面	0.979	0.640
56	政令	地方自治法施行令	0.977	0.643
66	政令	景観法施行令	0.972	0.606
87	政令	建設業法施行令	0.967	0.506
178	施工管理	建築施工管理	0.945	0.331
455	修繕	小修理	0.884	0.193
501	図面	断面図	0.876	0.182
675	設置	接着	0.846	0.145

5. おわりに

近年進歩の著しいニューラルネットによる単語の埋め込み表現を用いた建築用語の階層関係抽出を試みた。文脈による埋め込み表現の違いを考慮し埋め込み表現をクラスタリングし、更に教師付き学習の正例と負例のアンバランスを重みで調整することで、ある程度の精度での階層関係抽出を行うことができた。しかしながら、その精度は十分なものではなく、多くの手作業が必要である。その原因として、学習データの質と量の課題が挙げられる。質の面では、今回教師データとした JCCS は階層関係と言えど様々な関係が混在しており、そうした異質の関係が影響した可能性はある。また、元々オントロジーを意識した階層ではないため、概念の定義に違和感があるか所もある。また、埋め込み表現は文脈を考慮してクラスタリングしたが、同じように JCCS も同じ用語が異なる文脈で使われているか所があり、本来であれば複数のクラスターの中からそれぞれ文脈的に該当する JCCS の用語を選ぶべきではあるものの、今回は考慮していない。量的には、文脈により埋め込み表現が異なることで単純な数的には十分な数の教師データが得られているように見えるが、所詮は同じ用語のバリエーションに過ぎず、探索空間全体をカバーするには足りなかった。階層関係の教師データとしては、例えば WordNet などの言語資源が挙げられる。WordNet は上位下位関係やその他の関係を記述した言語資源で教師データとしての期待があるが、一方で同じ類義語が上位語や下位語にも表れるなど使いにくさもあり、その利用には工夫が必要かと思われる。

東京大学中川研究室・横浜国立大学森研究室で開発された用語抽出システムを使用しました。

[参考文献]

1) Devlin, D. et al.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proc. of the 2019 Conf. of the North American Chap. of the ACL: Human Language Tech., Vol.1, pp. 4171-4186, 2019
2) Rajpurkar, P. et al.: SQuAD: 100,000+ Questions for Machine Comprehension of Text, Proc. of the 2016 Conf on Empirical Methods in Natural Language Processing, 2016
3) Yang, Z. et al.: XLNet: Generalized Autoregressive Pretraining for Language Understanding, 33rd Conf. on Neural Information Processing Systems, 2019
4) Riichiro Mizoguchi: Knowledge acquisition and ontology, Proc. of KB&KS'93, Tokyo, pp.121-128, 1993
5) 武田英明：人工知能におけるオントロジーとその応用,情報知識学会 第 9 回(2001 年度)研究報告会講演論文集,情報知識学会,2001
6) 恒川 裕史,美馬 秀樹:オントロジーを用いた建築関連法令の検索と俯瞰に関する研究, 第 31 回情報・システム・利用・技術シンポジウム,日本建築学会,pp.79-84,2008.
7) Buitelaar, P. et al.: Ontology Learning from Text: An Overview, Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, 2005
8) Asim,M.N. et al.: A survey of ontology learning techniques and applications, Database (2018), Vol. 2018
9) Landauer, T.K. et al.: An introduction to latent semantic analysis, Discourse Processes, Routledge, vol.25, no.2-3, pp.259-284, 1998
10) 社会基盤情報標準化委員会: 建設情報標準分類体系 (JCCS) Ver.2.0 の公開について, JACIC, https://www.jacic.or.jp/hyojun/jccs-ver2r.html, (accessed 2020-9-18)
11) Chasen -- 形態素解析器, https://chasen-legacy.osdn. jp/, 奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座(松本研究室), (accessed 2020-9-18)
12) IPAdic legacy, https://ja.osdn.net/projects/ipadic/, (accessed 2020-9-18)
13) 湯本紘彰ほか:出現頻度と連接頻度に基づく専門用語抽出, 第 145 回自然言語処理研究会, 情報処理学会, 2015
14) 柴田知秀ほか: BERT による日本語構文解析の精度向上, 言語処理学会 第 25 回年次大会, pp.205-208, 名古屋, (2019.3)
15) Jay Alammar: The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning), http://jalammar.github.io/illustrated-bert/, (accessed 2020-9-18)
16) Chen, T. et al.:XGBoost: A Scalable Tree Boosting System, Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp.785-794,2016
17) 岸田 和明: 情報検索における評価方法の変遷とその課題, 情報管理, Vol.54, No.8, pp.439-448