デジタル人文学とメタデータ
Digital Humanities and Metadata

大野邦夫†
Kunio OHNO†

†株式会社モナビITコンサルティング　†Monavis IT Consulting Co. LTD.

Email: k−ohno@star.ocn.ne.jp

1. はじめに

今回の年次大会のテーマを「デジタル人文学とメタデータ」としたが、仮想美術館をテーマにした3月11日の第一回DMH研究会の際に、私の発表[1]を含め、メタデータ・オントロジに関する議論が活発になされたことに起因する。さらにメタデータに関しては、データ科学による自然言語処理の高度なツールの出現によりその概念が変貌を迫られているのではないかという議論も存在する。要するにword2vecやBERTと言ったツールにより、メタデータ付与が自動化される可能性が云々されているのである。本報告ではその技術状況を紹介すると共に、それらのツールが人文学領域に及ぼす可能性についても考察したいと考える。

従来のメタデータは、データに関するデータという概念で、図書館の図書カードが代表的なモデルである。分類番号、著者、タイトル、出版社、発行年月日、配置書架などが記されているが、要するに書籍に関する情報が属性と値のペアで記述される。この方式の標準的なメタデータは、ダブリンコアである。DCMI (Dublin Core Metadata Initiative)は、ウェブ上のリソースを記述する共通のメタデータ標準として、1995年3月に米国オハイオ州のダブリン(Dublin)で開催されたOCLC/NCSA Metadata Workshopでの討議結果を”Dublin Core metadata”と呼んだところに由来する。Dublin Coreの中核となるのは、基本となる15の属性（プロパティ）を定義したDublin Core Metadata Element Set[DCMES]で2003年2月には、ISO15836として国際標準化されている。この属性は、title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, right であり、これらの属性に対する値の形式で記述するデータ形式である。

2. メタデータの検討経緯

2.1 属性と値によるメタデータ記述

属性と値で項目を管理する手法は、ミンスキーのフレーム理論的な知識構造に対応するが、データモデル的には抽象データ型としての構造体である。オブジェクト指向プログラミングのクラス定義は、クラス変数やインスタンス変数を構造体として扱い、クラス継承に伴う意味継承を活用する系統的な管理を可能とする手法であった。

マークアップ言語のSGML、XMLは、抽象データ型を階層的に扱うデータモデルである。オブジェクト指向のクラスが、集合論理に基づく概念の階層を形成するのに対して、SGML、XMLは実装データ（インスタンス）としての階層を記述する枠組みである。XML表現であるRDFやOWLは、データ形式としてのXML表現に論理的な方向付けや、クラスとしての意味的な階層を導入した概念記述の語彙であり、メタデータ記述に対するひとつの方向付けを与えた。現在具体的に用いられているメタデータは、概略この枠組みと言えるであろう。

2.2 セマンティックWebとメタデータ

西暦2000年前後に、メタデータが話題になったことがあった。そのトリガーは、W3Cが提言したセマンティックWebの階層（図1）に、RDFとRDF Schemaの層があり、メタデータ層として扱われたことに起因すると思われる。

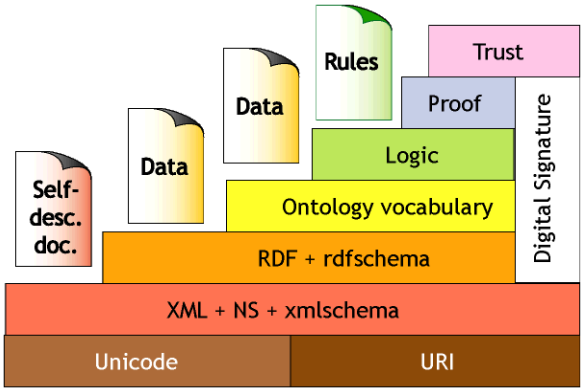

図1　セマンティックWebの階層モデル

そのような経緯から、上位のオントロジ層とセットになって議論され検討した経緯がある。画像電子学会でもメタデータSG（Study Group）が設けられ、ネットワーク管理、デバイス管理、コンテンツ管理の枠組みの下で、メタデータ・オントロジ分野の標準化検討が試みられた[2]。

2.3 OMGによるメタデータの標準化

セマンティックWebで話題になる以前に、オブジェクト指向技術の標準化団体であるOMG（Object Management Group）でメタデータの標準化に関する検討が行われた[3]。OMGは、基本的通信の枠組みとしてのCORBA（Common Object Request Broker）に基づく標準化分野を、OS（オブジェクトサービス）、CF（共通ファシリティ）、AO（アプリケーションオブジェクト）に区分したが、CFの中にMOF（メタオブジェクトファシリティ）という領域を設定し、その枠組みとして表1に示すM0〜M3のレベルを設定した。この枠組みは、オブジェクト分析設計における抽象レベルの標準化を意図しており、UML（Unified Modeling Language）の実装に寄与することを狙ったものであった。しかしながらOMGの組織改変により具体的な標準化は行われなかった[4]。

なおMOFの経緯に基づくオブジェクト分析設計における具体的な実装が検討され、XMI（XMLMetadataInterchange）という名称で標準化が検討された。従来XMLに対して批判的

表1　OMGのMOFにおけるメタデータ階層

メタレベル	MOF用語	事例
M3	メタ・メタモデル	MOFモデル
M2	メタ・メタデータ メタモデル	UMLメタモデル CWMメタモデル
M1	メタデータ モデル	UMLモデル、ウェアハウスメタモデル
M0	データ	モデル化されたシスム、ウェアハウスDB

だったOMGも方針を変えてセマンティックWebに準拠した枠組みの標準化を目指したものであった。XMIの具体的な概念として、図2に自動車の色とドア形式に関する事例を示す。

データレベルのモデル（MOFのM0）はUMLのクラス図で提示される。インスタンスとしてのXMI Documentでは、AutoのタグのColorとDoorが要素名となり、その値が与えられる。そのメタレベルのモデル（MOFのM1）としてXMIDTDが定義されているが、XMIが検討された当時、XML SchemaがW3Cで正式に標準化されていなかった経緯がある。XMLによるDTDとインスタンスの表現をUMLのクラス図を経由して、CORBAのIDL（Interface Definition Language）やC、Javaのクラス実装に結び付けることが意図されていた。

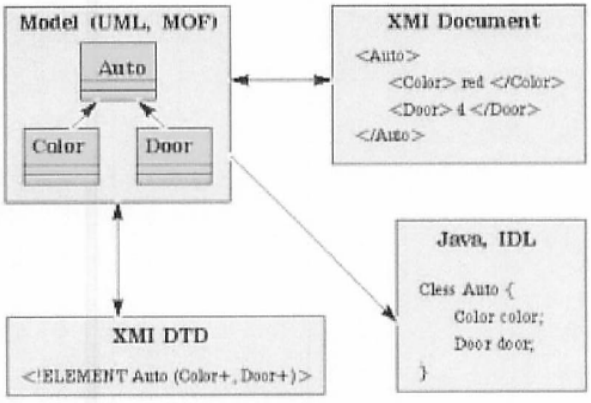

図2　XMIによるモデル変換の概念

2.4 MDA

なおこの枠組みは、MDA（ModelDrivenArchitecture）として発展し、UMLによる仕様と具体的なプログラム実装を双方向的に一元管理可能なアーキテクチャを可能とすることを試みたもので、セマンティックWebのオントロジ層にも対応付けることを指向した野心的な試みであったが、プロトタイプ止まりで成功したとは言えなかった。

だが、MDAの発想は、その後ProtegeなどのGUIツールに基づくOWL（Ontology Web Language）によるオントロジ実装[5]の端緒であった。Protegeは、UMLのクラス図の代わりにGUIを通じてオントロジのクラス階層を構築するツールであるが、アクティビティ図やシーケンス図が不在なのでプログラム実装のためにはUMLのような系統性に欠けていた。

個人的には、UMLのクラス図をCommon Lispのクラス定義を記述するCLOSに実装して階層的・系統的なクラスとして構築するオントロジの方が実用的な印象を持った。種々の処理がCLOSの総称関数としてクラスの枠組み内で定義でき、かつ一般の関数同様に扱えるからである。OWLによる実装は、ApacheのJenaを用いるのであるが、非習熟のJavaプログラマには極めて煩雑になる。OWLの特徴とされる論理記述は、実用的には殆ど無意味な印象であった。CLOSを用いたオントロジ実装は、履歴書オントロジ[6][7]、PIMオントロジ[8]や造り酒屋オントロジ[9]として検討した経緯がある。UMLからCLOSとXMLを含む実装コードを系統的に構築し、XMLデータをXSLTでHTMLに変換し、Webに表示することを実現したので、XMIとMDAの狙いを部分的に実装したと考えている。

3. データ科学時代のメタデータへの展望

3.1 深層学習の自然言語分野へのインパクト

他方、2010年代に入って、深層学習による判別技術が急速に進展した。既に5年前の過去の話であるが、深層学習が人材育成と職業変化に及ぼす可能性[10]と、深層学習と以前のAIとの対比[11]とについて考察したことがあるが、その時点では自然言語に対する深層学習の導入が、以前のAIの役割を超えるであろうと考えた。深層学習の基本的なアルゴリズムとしては、CNN、Auto−Encoderが挙げられるが、深層学習を自然言語処理に適用するためには、中間ノード層をバッファ的にデータを蓄積し、それを再帰的にフィードバックして逐次処理するRecurrent Neural Network手法、略称RNNが開発され、回帰結合ニューラルネットワークという日本語名称で検討されつつあることを把握したがそれ以上は不明であった。

だが自然言語への適用はその背後で着実に進展し、ビッグデータとしての大量の文書を背景に、グーグルの研究所がその実用化に成功してグーグル翻訳で効果的に使用される段階を迎えている。その基本的な手法として、word2vecとBERTが知られており、そのアウトラインを最近素人なりに把握した。以下はその紹介である。

3.2 単語の分散表現とword2vec

word2vecは、文字通りニューラルネットワークを通じて、特定分野の単語群をベクトル空間に構築する手法である。大規模な文書群（コーパス）を入力してその単語群のベクトル空間を生成するが、このベクトル空間は典型的には数百次元からなり、コーパスの個々の単語はベクトル空間内の個々のベクトルに割り当てられる。コーパス内で同じコンテキストを共有する単語ベクトルは、ベクトル空間内の近くに配置される[12]。このような単語の扱いは分散表現と呼ばれている[13]。

Wikipediaによると、word2vecでは、CBoW（Continuous Bag−of−Words）モデルおよびskip gramモデルという二つのモデル構造のいずれかを使用し、単語の分散表現を生成する。CBoWモデルでは、周囲のコンテキスト単語から現在の単語を予測するが、コンテキスト単語の順序は問わない。skip−gramモデルでは、現在の単語を使用して、周囲のコンテキスト単語を予測する。現在の単語に近ければ近いほど、コンテキスト単語の重みを大きくする。skip−gramモデルはCBoWモデルと比較すると処理量が膨大になるが、出現頻度が低い単語に対しても適用範囲が拡大する。個別の単語ではなくドキュメント全体からの単語の埋め込みを構築するためのword2vecの拡張が検討され、この拡張は、paragraph2vecまたはdoc2vecと呼ばれ実装されている[14]。

3.3 生化学分野への適用

この手法を使用すると、分野ごとに単語の類似性がベクトル空間の距離として定量化される。この原理を用いて、バイオインフォマティクスのための単語ベクトル、バイオベクターが提案され、タンパク質（アミノ酸配列）に対するタンパク質ベクター（ProtVec）、遺伝子配列に対する遺伝子ベクター（GeneVec）といった生物学的配列の総称として確立されている。プロテオミクスおよびゲノミクスにおける機械学習の実装において幅広く用いられている。BioVectorsが生化学的および生物物理学的解釈に基づいて生物学的配列を分類できることが示唆されている[15]。

さらに、バイオベクターを用いる医療分野における放射線医学分野では、単語ベクトルに対してインテリジェントな単語埋め込み（IWE）を実践している。Intelligent Word Embedding (IWE) は、word2vecに類義語辞書マッピング手法を組み合わせて、口語表現や語彙の曖昧さに関する整合性を試みている[16]。この分野の日本の研究は目立たないが、分散表現に関しての報告が行われている[17]。

4. BERT

4.1 双方向の符号化表現

word2vecのような手法を用いて、単語の分散表現をベースとする自然言語処理が行われたが、従来は文章内の単語を逐次前後方向に分析する手法であった。それを双方向に行うアイデアで飛躍的な進展をもたらした技術がBERTである[18]。

BERTは、Bidirectional Encoder Representations from Transformersの略で、Transformerによる双方向の符号化表現で、2018年10月にGoogleのJacob Devlinらの論文で発表された自然言語処理モデルである。翻訳、文書分類、質問応答など自然言語処理分野の処理のことを「タスク」と呼ぶが、BERTは、多様なタスクのベンチマークにおいて優れた性能を発揮した。

一般に自然言語処理は、文を語彙を要素とする一次元のベクトルと見なし、先に述べた通り単語を高次元のベクトルに置き換える分散表現技術を用いて入力される。単語データの並びを「シーケンス」と呼び、これは文章に相当するが、BERTは入力されたシーケンスから別のシーケンスを予測する技法である。

4.2 MLMとNSP

BERTは事前学習モデルであり、入力されたラベル無し、すなわち分類名称がついていないシーケンスをTransformerを通じて処理することによって学習する。実際には、TransformerがMLM（Masked Language Model）とNSP（Next Sentence Prediction）という2つの手法を並列に処理することにより学習する。図3に示す通り、事前学習の後にファインチューニングを行う。事前学習を初期値として分類名称が付与されたラベルありデータでファインチューニングを行ない、詳細な処理を実現する。

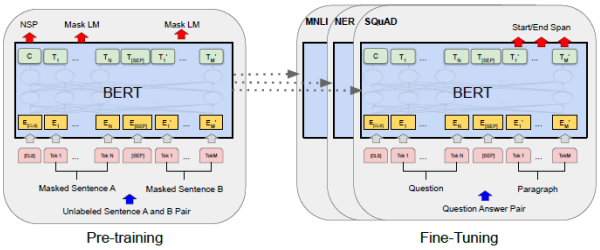

図3　BERTにおける事前学習とファインチューニング

従来の自然言語処理モデルでは、文章を進行方向でしか処理しなかったため、目的の単語の前の文章データから予測する必要があった。しかし、BERTは双方向のTransformerによって学習するため、従来の手法に比べ精度が向上した。それを実現しているのがMasked Language Modelである。

入力文の15%の単語を別の単語で置き換え、文脈から置き換える前の単語を予測する。具体的には、選択された15%のうち、80%は[MASK]に置き換えるマスク変換、10%をランダムな別の単語への変換、残りの10%はそのままの単語で残す。このようにして置換された単語を前後の文脈から推測するタスクで解くことにより、単語に対応する文脈情報を学習する。

4.3 文の隣接可能性判別

Masked Language Modelにおいて単語に関する文脈情報の学習は実現できるが、文単位の学習はできない。そこで、2つの入力文に対して、その2文が隣接する可能性をNext Sentence Predictionによって学習する。これにより、2つの文の関係性を学習することが可能になる。

文の片方を50%の確率で他の文に置き換え、それらが隣接するか（isNext）隣接しない（notNext）か判別することによって学習する。そのために2文を[SEP]というトークンで分け、isNextかnotNextかを分類するために[CLS]というトークンが用意される。

BERT以前にはELMo、OpenAI GPTといった言語処理モデルが存在した。図4に示すようにELMoは浅い双方向モデルであり、OpenAI GPTは未来の単語しか予測することができない単一方向モデルであった。そのために文脈を処理することができなかった。

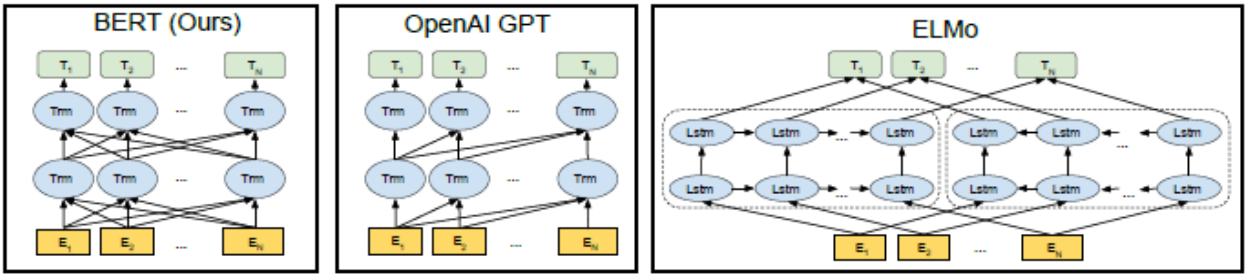

図4　BERTとOpenAI－GPT、ELMoの処理比較

4.4 文脈における意味把握

例えば、接続詞における隣接文相互の意味的関係は、BERT導入以前の処理モデルでは踏み込んだ分析は不可能であいまいにならざるを得なかった。さらに文中の否定語の存在による意味の逆転においてもBERTは有効性を発揮した。

BERT導入以前のGoogle検索では、否定語が存在しても否定する対象を具体的に把握するのは困難であった。

このようにして、自然言語分野における語彙や文章の意味的な分析技術は着実な進展を遂げている。この技術が世界中の記録文書に適用され得ることを考えると、巨大なインパクトを感じざるを得ないがかなりの時間が必要であろう。

5. 考察

人文学は英語のHumanitiesの日本語であり、人間に関する学問であるが、そのカテゴリは、Wikipediaによると、哲学、論理学、倫理学、美学、宗教学、歴史学、考古学、人文地理学、文化人類学、民俗学、言語学、文学、芸術学、教育学、心理学、人間科学が列挙されている。

人文学における従来の研究方法の基本は文献学的方法であり、解釈の論理的整合性だけが研究者の主張に妥当性をあたえるとされてきた。だが、自然科学的な実験や観察、統計もまた人文科学の方法として使用されるようになり、さらに従来の文献学がコンピュータ科学やデータ科学に置き換えられ情報管理の方法論が急速に変化しつつあるのが現状である。word2vecによる単語や語彙の分散表現、BERTによる文脈処理の適用と言った技術は、今後人文学における古典書籍などに導入されて、種々のデータが獲得されると思われる。博物館や美術館の領域においても、その電子コンテンツ化された情報に関しては、分散表現やBERTが系統的に適用され、新たな事実が判明することが期待される。

分散表現による語彙の分布とBERTによる意味的分析を通じて、冒頭で述べた通り、メタデータ付与が自動化される実現性は十分に存在する。現に組織名、人名、地名、日付表現などに関しては、固有表現抽出（named entity recognition）という手法により、文中の単語の属性を割り当てる手法が一般化しており、そのための辞書も開発されている[19]。従ってメタデータの自動付与は時間の問題であり、むしろ従来のメタデータ領域を包含する新たな情報管理の体系が構築されると思われる。これは学問・研究分野のみならず、最近話題になっているDXなどのビジネス分野などにも展開すると思われる。

とは言え従来のメタデータが完全に廃れることは無いであろう。従来のメタデータの象徴とも言えた図書カードは、コンピュータに置き換えられたが、図書館や書店における図書の書架の検索には依然として、著者、タイトル、出版社、発行年月日など、かつての図書カードの属性が用いられている。このように一般人における書籍や資料の検索は、文化であり従来の習慣が急に変わるものではない。従って、分散表現やBERTによる処理などは、従来の分析手法を補完する形式で徐々に導入されると思われる。

6. おわりに

以上、デジタル人文学とメタデータについて、発展しつつあるデータ科学領域の観点から考察した。画像とは対極と思われる分散表現やBERTのような自然言語分野を、画像電子学会が取り組むのは異色かもしれないが、文章と図形・画像を連携させた複合文書[20]は、文字だけのドキュメントよりも認知的に分かりやすいので、却って着手しやすいのではないかと思われる。その観点を包含したデジタル人文学と画像情報の関係については、一昨年のDSGワークショップで考察している[21]。この分野への関心が高まることを期待したい。なお本検討を行うにあたり、歴史的経緯は主に画像電子学会メタデータSGでの検討に伴い把握した内容である。その設立・運営に貢献頂いた前国士舘大学教授の小町祐史様、およびIIJ Innovation Instituteの新麗様に感謝します。さらに生化学分野におけるデータ科学情報の提供を頂いた、前国立遺伝学研究所長の桂勲様に御礼申しあげます。

文献

[1] 大野邦夫；”Webサービスにおける仮想ミュージアムへの考察～物語性とキュレーションの観点から～ ”, 画像電子学会第1回デジタルミュージアム・人文学研究会資料(2021.3)

[2] 大野邦夫, 小町祐史；”オントロジ応用技術国際標準化の事前検討”, 画像電子学会第13回VMA研究会資料(2004.7.9)

[3] 大野邦夫；”メタデータ概念と記述”, 精密工学会研究会（2001.11）

[4] 河込和宏, 中村秀男, 大野邦夫, 飯島正; “分散オブジェクトコンピューティング”, 共立出版, pp.287−323（1999）

[5] AIDOS; “オントロジ技術入門”, 東京電機大学出版局（2005）

[6] 大野邦夫, 須藤僚；”拡張可能な履歴書管理システムの情報環境に関する研究～ジョブカード様式を事例とするXMLとLispの比較”, 平成21年度職業能力開発総合大学校紀要（2010.3）

[7] 大野邦夫, 角山正樹；”拡張可能な履歴書管理システムの 実装に関する検討”, 平成22年度職業能力開発総合大学校紀要（2011.3）

[8] 大野邦夫, 王研；”Common Lispによるパーソナル情報の管理とWeb表示に関する研究”, 画像電子学会研究会in鹿児島(2011.3)

[9] 大野邦夫;”オブジェクト指向プログラミングによる意味的クラス継承に関する考察 ～造り酒屋オントロジモデルの検討から得られた可能性と限界～”,情報処理学会研究報告, DC116−5（2020.3）

[10] 大野邦夫；”深層学習技術の適用に関する一考察 ～ 第2世代AI事業関係者の個人的な見解”, 第3回画像関連学会連合会大会講演論文（2016.11）

[11] 大野邦夫；”深層学習時代のIoTサービスと人材育成への展望と課題”, 画像電子学会第7回DSGワークショップ(2016.11)

[12] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J.: Distributed representations of words and phrases and their compositionality, Proc. of NIPS, pp. 3111−3119（2013）

[13] 岡崎 直観; “言語処理における分散表現学習のフロンティア ”, 人工知能 31巻2号（2016.3）

[14] Le, Quoc. ”Distributed Representations of Sentences and Documents”. arXiv:1405.4053 [cs.CL]。

[15] Asgari, Ehsaneddin; Mofrad, Mohammad R.K. (2015). “ Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics ”. PLOS One 10 (11): e0141287. arXiv:1503.05140. Bibcode: 2015PLoSO..1041287A. doi:10.1371/journal.pone.0141287. PMC: 4640716. PMID 26555596.

[16] Banerjee, Imon; Chen, Matthew C.; Lungren, Matthew P.; Rubin, Daniel L. (2018). “ Radiology report annotation using intelligent word embeddings: Applied to multi−institutional chest CT cohort ”. Journal of Biomedical Informatics 77: 11?20. doi:10.1016/j.jbi.2017.11.012. PMC: 5771955. PMID 29175548.

[17] Hitoshi Iuchia, Taro Matsutani, Keisuke Yamada, Natsuki Iwanob, Shunsuke Sumi, Shion Hosoda, Shitao Zhao, Tsukasa Fukunaga and Michiaki Hamada; “Representation learning applications in biological sequence analysis”, bioRxiv preprint doi: https://doi.org/10.1101/2021.02.26.433129

[18] Jacob Devlin, Ming−Wei Chang, Kenton Lee, Kristina Toutanova; “BERT: Pre−training of Deep Bidirectional Transformers for Language Understanding”, arXiv.org > cs > arXiv:1810.04805, Cornell University, (2018)

[19] 春遍雀來；”深層学習用辞書データベース～DeepLEX”, 日中韓辞典研究所（2020.12）

[20] 大野邦夫；”複合文書の標準化経緯 − その登場からHTML5に至るまで − ”, 画像電子学会誌,Vol.47, No.4, pp.488−491 (2018)

[21] 大野邦夫；”デジタル人文学と画像情報”, 画像電子学会第10回DSGワークショップ(2019.12)