

# 肝細胞がん診療ガイドラインに関する知識グラフと質疑応答システムの構築

(Construction of a Knowledge Graph and  
Question-Answering System for Clinical Practice  
Guidelines for Hepatocellular Carcinoma)

Author: Runpeng XIA

## 概要

近年、深層学習技術は著しい発展を遂げ、様々な科学研究分野で利用されるようになっている。医療分野に存在する諸課題の解決に対しても深層学習技術が応用されつつある。日本の医療機関における肝細胞がんの診療指針として重要な役割を担う肝細胞がん診療ガイドラインは、臨床研究者にとって不可欠な文書である。このガイドラインには役立つ知識が豊富に含まれているが、そのボリュームや構造化されていない情報のため、肝細胞がんの診療に関する情報を検索するのが不便であり効率性も良くない。この課題を解決するために本研究では、深層学習技術を活用して肝細胞がん診療ガイドラインに関する知識グラフおよび質疑応答システムを構築した。構築した知識グラフと質疑応答システムに対して評価を行なった結果、両者の有効性が証明された。これにより肝細胞がん診療ガイドラインに含まる知識の構造化と可視化が実現され、情報検索の効率性と利便性の向上という目的も達成された。

**キーワード：**肝細胞がん診療ガイドライン、知識グラフ、質疑応答システム、深層学習技術、自然言語処理

# Abstract

In recent years, deep learning technologies have undergone remarkable development and have been employed across various scientific research fields. These technologies are increasingly being applied to address the challenges present in the medical domain. The Clinical Practice Guidelines for Hepatocellular Carcinoma play a critical role as a directive for the treatment of hepatocellular carcinoma in Japan, serving as an indispensable resource for clinical researchers. Although the guidelines are replete with valuable knowledge, the volume of content and lack of structured information render the retrieval of hepatocellular carcinoma treatment information inconvenient and inefficient. To address this challenge, we have constructed a knowledge graph and a question-answering system for Clinical Practice Guidelines for Hepatocellular Carcinoma by leveraging deep learning technologies. Upon evaluating the constructed knowledge graph and question-answering system, their effectiveness was demonstrated. This has enabled the structuring and visualization of knowledge contained in the Hepatocellular Carcinoma Treatment Guidelines, achieving our objective of enhancing the efficiency and convenience of information retrieval.

**Keywords:** Clinical Practice Guidelines for Hepatocellular Carcinoma, Knowledge Graph, Question-Answering System, Deep Learning, Natural Language Processing

# 目次

概要.....	2
ABSTRACT.....	3
目次.....	4
1 序論.....	6
1.1 背景.....	6
1.2 研究目的.....	8
1.3 本論文の構成.....	8
2 関連先行研究.....	9
2.1 知識グラフ.....	9
2.2 質疑応答システム.....	12
2.3 基礎的な深層学習技術.....	13
2.3.1 単純パーセプトロン .....	13
2.3.2 多層パーセプトロン .....	14
2.3.3 誤差逆伝播法 .....	16
2.3.4 回帰型ニューラルネットワーク .....	18
2.3.5 長・短期記憶 .....	21
2.4 自然言語処理における深層学習技術.....	24
2.4.1 Word2Vec .....	24
2.4.2 トランスフォーマー .....	27
2.4.3 GPT .....	31
2.4.4 BERT .....	34
2.4.5 BART .....	36
2.4.6 大規模言語モデル .....	39
2.5 知識グラフ構築への深層学習技術の応用と課題 .....	41
2.5.1 知識グラフ構築への深層学習技術の応用 .....	41
2.5.2 知識グラフ構築に存在する課題 .....	43
2.6 質疑応答システム構築に関する先行研究と課題 .....	44
2.6.1 質疑応答システム構築への深層学習技術の応用 .....	44
2.6.2 質疑応答システム構築における課題 .....	45
3 肝細胞がん診療ガイドラインに関する知識グラフの構築 .....	46
3.1 REBEL による関係トリプル抽出の試み .....	46
3.1.1 REBEL の概要.....	46
3.1.2 REBEL の構築に用いたベースモデルとデータセット.....	47
3.1.3 REBEL による抽出性能の検証.....	48

3.2 多段階関係トリプル抽出アルゴリズムによる知識グラフ構築 .....	50
3.2.1 提案手法 .....	50
3.2.2 固有表現認識モデルの構築 .....	51
3.2.3 固有表現認識によるワードクラウドの作成およびキーコンセプトの抽出 .....	55
3.2.4 構文解析による対象文の選別 .....	60
3.2.5 大規模言語モデルによる属性と属性値の抽出 .....	61
3.3 構築した知識グラフの評価 .....	65
<b>4 肝細胞がん診療ガイドラインを対象とした質疑応答システムの構築 .....</b>	<b>66</b>
4.1 構築手法 .....	66
4.2 質疑応答システムの性能評価 .....	67
<b>5 考察 .....</b>	<b>70</b>
5.1 構築した知識グラフに関する考察 .....	70
5.2 構築した質疑応答システムに関する考察 .....	71
<b>6 結論 .....</b>	<b>73</b>
6.1 研究内容のまとめ .....	73
6.2 結論と今後の展望 .....	74
<b>謝辞 .....</b>	ERROR! BOOKMARK NOT DEFINED.
<b>参考文献 .....</b>	<b>75</b>
<b>APPENDIX .....</b>	ERROR! BOOKMARK NOT DEFINED.

# 1 序論

## 1.1 背景

肝細胞がん (Hepatocellular Carcinoma、HCC) は、肝臓の主要な細胞である肝細胞ががん化して肝臓にできる腫瘍性疾患である[1]。日本における原発性肝がんの年間新規罹患数は約4万人であり、その中の約95%を肝細胞がんが占めており、部位別のがん罹患数でみると男性では第5位、女性では第10位となっている[2]。肝細胞がんは日本においても死亡数が高いがんの一つとして数えられ、2020年度のがん死亡数統計によれば年間約2万5,000人が肝細胞がんで死亡しており、部位別のがんによる死者数では5番目に多いがんとなっている[3]。肝細胞がんの生存率改善に早期発見と適切な治療が重要な要因であり、肝細胞がんの診断・治療の適切な進め方に関する有用な情報を記載するのは肝細胞がん診療ガイドラインである。

肝細胞がん診療ガイドライン (Clinical Practice Guidelines for Hepatocellular Carcinoma) [4]は、日本肝臓学会 (The Japan Society of Hepatology) によって作成・出版され、肝細胞がんの診療に携わる肝臓専門医・一般医師を含む全ての医師を対象としてエビデンスとコンセンサスに基づいて日本における肝細胞がんの標準的なサーベイランス法、診断法と治療法を提示するものである。日本における肝細胞がん診療レベルを全体として向上させることを目的とし、肝細胞がん診療ガイドラインは地域や施設による格差の解消、肝細胞がん患者の生存期間の延長と生活の質の改善を目指している。肝細胞がん診療ガイドラインは、最新の知見の提供や診断と治療の標準化、医療リスクの管理など多くの医療現場のニーズを満たしているため、日本全国の様々な医療施設で肝細胞がんの診療に携わる医師にとって必携の一冊となっている。

肝細胞がん診療ガイドラインの本文は、「診断およびサーベイランス」、「治療アルゴリズム」、「予防」、「手術」、「穿刺局所療法」、「肝動脈塞栓療法」、「薬物療法」、「放射線療法」、「治療後のサーベイランス・再発予防・再発治療」の9章に分かれている。各章はその章のタイトルに示した分野について書かれ、図1に示すように「はじめに」と、複数個のクリニカルクエスチョン (Clinical Question) およびその解説から構成されている。「はじめに」は、その章の前書きである。クリニカルクエスチョンは、専門家によって選定され、肝細胞がん診療において臨床的な意義が高い質問である。肝細胞がん診療ガイドラインに合計56個のクリニカルクエスチョンが記載されているので、ガイドライン自体もかなり分厚い文書となっている。

## ●はじめに

肝切除外は腫瘍に対して最も根治的な治療であり、近年、その安全性は飛躍的に向上している。しかし、National Clinical Database (NCD) の解析による、肝切除外全体の手術閑死率は2.4%と報告されており、さすがに安全性を向上させる必要がある。そのためには、肝切除外の適応となる肝細胞癌手術、肝機能や肝子機能の評価などさらに選ばれた安全で合意的な手術形式の選択が必要となる。肝切除外に関する検討の大きな変化は種々の手術機器と手術手技の改良と進歩による腹腔鏡下肝切除外の普及である。本邦においては、2010年に高齢先進医療として認可され、2010年に肝部分切除術と肝外側区域切除術が、さらに2010年には腹腔鏡下肝腫瘍剥離を行わなければ肝切除外が認められ、その手術者数も急速に増加している。しかし、腹腔鏡下肝切除外は特に広範な切除外において完全に確立された手術手技ではなく、そのリスクも否定できない。ここ数年の安全性に対する懸念等により、本ガイドラインへの掲載に躊躇する様な議論があつたが、腹腔鏡下肝切除外の初期および後成績などのエビデンスが報告され、国際コンセンサス会議においても腹腔鏡下肝切除外が費示されてきている経緯を踏まえ、今回の改訂では腹腔鏡下肝切除外の適応についてのCQを加えることとなった。ただし、これは腹腔鏡下肝切除外をやみくもに推奨するものではなく、むしろ慎重な導入、確実な実験とその結果に対する対

## CQ81

穿刺局所療法はどのような患者に行うのが適切か？

## 推奨

穿刺局所療法の適応はChild-Pugh分類AあるいはBの症例で、腫瘍径3cm以下、腫瘍数3個以下である。(強い推奨)

## ■背景

肝切除外、穿刺局所療法、TACE のいずれもが施行可能な患者が存在した場合、治療アルゴリズムでは原則として、肝切除外、穿刺局所療法、TACE の順に推奨されている。セカンドラインとしての穿刺局所療法の位置づけは、ファーストラインである肝切除外、セカンドラインであるTACEとの推定される手後の差によって決定されるが、その後の差は肝機能と腫瘍条件によって変化することが推定される。穿刺局所療法の適応を他治療との比較から設定可能か検討した。

図1 肝細胞がん診療ガイドラインの例[4]

しかし、図1からわかるように、肝細胞がん診療ガイドラインは文書で非構造化データであり、確かにその中に医学分野の知識が大量に含まれているがそれらの知識は構造化されておらず知識間の関連性が分かりにくく、一つの専門用語が複数の章に現れることもよくあり同じトピックに関連する知識の検索は不便である。ガイドラインの中から知識を獲得するために、分厚いガイドライン本文を細かく読まなければならず、情報検索の効率性が低いという現状を認めざるを得ない。また、肝細胞がん診療ガイドラインに医学分野の外来語や専門用語が頻出し長文も多用されていて構文が複雑であるという特徴があり、日本語ネイティブではない外国籍の研究者にとって読みにくく、そのような人たちの情報検索の効率性をさらに下げてしまう恐れがある。こういった問題を解決するために、肝細胞がん診療ガイドラインに含まれる知識を構造化・可視化して肝細胞がん診療に関する情報検索の効率化を向上させる必要がある。

情報工学分野において知識の構造化や可視化を行うための手法として、混ざり合っている集団から似ているデータ同士をグルーピングすることでデータの中に潜む構造を表出させるクラスター分析や、複数の概念やプロセスを順序立てて視覚化するフローチャートなどの手法がある。これらの手法の中、知識グラフ (Knowledge Graph) はテキストの中の知識を構造化すると同時に複数の知識間の複雑な関係性を明瞭化した上で直観的に理解しやすい形で知識を可視化するという長所があるため、本研究では知識の構造化・可視化のための手法として知識グラフの構築を提案する。また、情報検索の効率性を向上させるために、特定の質問に迅速かつ正確に答える能力を持つ質疑応答システム (Question Answering System) が一般的に用いられており、本研究でも肝細胞がん診療に関する情報検索の効率化を実現するための手法として適切であると考え採用することにする。さらに、質疑応答システムの構築における専門性の欠如という課題を解決するために知識グラフの質疑応答システムへの組み込みも試みる。

以上のことから、本研究では、知識の構造化と可視化を実現し、情報獲得の効率性を向上させるために、肝細胞がん診療ガイドラインを対象とした知識グラフと質疑応答システムの構築を提案する。

## 1.2 研究目的

本研究の目的は、肝細胞がん診療ガイドラインに関する知識グラフおよび質疑応答システムを構築することにより、ガイドラインに含まれる知識の構造化と可視化を図り、専門的正確性を保ちつつ肝細胞がん診療に関する情報検索の効率性と利便性を向上させることである。この目的を達成するために、本研究ではまず、知識グラフおよび質疑応答システムの構築にあたって重要となる深層学習技術を調査し、深層学習技術を活用した知識グラフ・質疑応答システム構築に関する先行研究を整理し、存在する課題を明らかにする。次に、深層学習技術を用いて肝細胞がん診療ガイドラインに関する知識グラフおよび質疑応答システムを構築し、知識グラフに対して正確性と完全性の二つの観点から評価し、質疑応答システムに対して正解率を評価基準として評価を実施する。最後に、構築した知識グラフと質疑応答システムの有効性について考察する。

## 1.3 本論文の構成

本論文の全体的な構成は以下のようになる。

第1章では、本研究の背景である肝細胞がん診療ガイドラインの概要について説明し、存在する課題を論じた上で本研究の目的および論文の構成を説明した。

第2章では、知識グラフと質疑応答システムを紹介し、知識グラフ・質疑応答システム構築にあたって基礎となる深層学習技術を概観する。また、深層学習技術を用いて知識グラフと質疑応答システムを構築する際の課題について、関連する先行研究をまとめながら論じる。

第3章では、知識グラフの構築方法を提案し、肝細胞がん診療ガイドラインに関する知識グラフを構築する。構築した知識グラフを正確性と完全性という二つの観点から評価する。

第4章では、質疑応答システムを構築する方法を説明し、肝細胞がん診療ガイドラインを対象とした質疑応答システムを構築する。また、肝細胞がん診療ガイドラインの中のクリニカルクエスチョンに対する正解率を評価基準として構築した質疑応答システムの評価を行う。

第5章では、第3章と第4章で得られた評価結果を総合し、構築した知識グラフと質疑応答システムの有効性について考察を行う。

最後に、第6章で本研究の研究内容をまとめた上で結論を述べる。

## 2 関連先行研究

本章では、知識グラフおよび質疑応答システムの構築に関する先行研究および存在する課題について記述する。2.1節と2.2節では、知識グラフと質疑応答システムの概要について説明する。2.3節と2.4節では、知識グラフ・質疑応答システムの構築にあたって重要な深層学習技術を、基礎的な深層学習技術と自然言語処理における深層学習技術の二つの側面から説明する。2.5節と2.6節では、深層学習技術を知識グラフと質疑応答システムの構築に応用した先行研究を紹介し、深層学習技術を用いて肝細胞がん診療ガイドラインに関する知識グラフ・質疑応答システムを構築する際の課題についても明らかにする。

### 2.1 知識グラフ

知識グラフは、肝細胞がん診療に関する複数の概念間の複雑な関係性を明瞭化し、これらの概念およびその相互関係を直観的に理解しやすい形で可視化することを可能にするため、本研究では肝細胞がん診療ガイドラインを体系的に構造化し視覚的に表現する方法として、知識グラフの構築を提案する。本節では、知識グラフの概要について説明する。

知識グラフには一意的で厳格な定義が存在しないが、グラフに基づく知識表現手法の一種であると広く認知されている[5]。本研究では、知識グラフ (Knowledge Graph) を、グラフ構造を用いて知識と知識間の関係を表現することによって、知識をモデリングするデータモデルであると定義する。一般的に、知識グラフは複数個のノードと任意の二つのノードを結ぶエッジによって構成される有向グラフとして表現される。知識グラフにおけるノードは記述対象を表すものであり、人物や書籍のような具体的な概念から、「知識グラフ」や「資本主義」のような抽象的な概念を表す実体に至るまで、多岐にわたる対象がこのノードによって表現され得る。一方で、知識グラフにおけるエッジは、二つのノード間の関連性や相互作用を示すものであり、友人関係や配偶者といった人間関係の表現から、「所有する」や「診断される」といったより複雑な相互作用の描写に至るまで、幅広い関係性を表現することが可能である。知識グラフを構成する基本単位は、二つのノードとそれらを結ぶエッジから成る三つ組（トリプル）である。このトリプルは、「実体—属性—属性値」や「実体—関係—実体」のような形で知識を格納することができ、複数のトリプルが相互につながることによって、知識の集合体である知識グラフが形成される。

簡単な知識グラフの例として、図2は選挙の仕組みに関する知識グラフを示している。この知識グラフは、選挙システムを論じる上で欠かせない「有権者」、「候補者」、「政党」といった中核的な概念をノードとして含んでおり、それぞれのノード間は「持つ (has)」、「～である (is\_a)」、「できる (can)」といった関係性を示すエッジで結びつけられている。知識グラフによって、選挙における多様な概念とそれらの間の関係性が視覚化され、選挙のプロセスがいかにして多層的な主体間の相互作用によって成り立つ

ているのかが明確に理解できる。例えば、候補者が人間であること、有権者が選挙権を有していること、選挙管理委員会が選挙活動を監督していることといった情報が、この知識グラフから容易に把握できる。

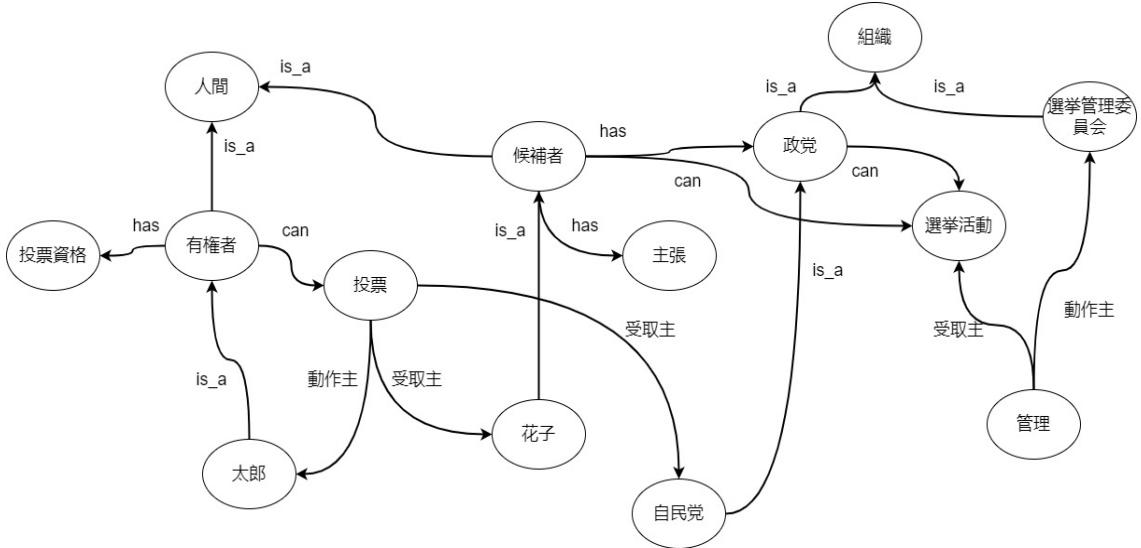


図 2 知識グラフの例

知識グラフという用語が最初に登場したのは、1980 年代後半、オランダのフローニングン大学とトゥウェンテ大学が共同で行った「Knowledge Graph」というプロジェクトである。このプロジェクトでは、異なる情報源からの知識を統合し、自然言語を表現するための知識ベースシステムを形式的に記述するための用語として知識グラフが導入された [6]。当時の研究者たちは、人間の相互作用を含む限定された関係性に焦点を当てた定性的モデリングとしての知識グラフを提案しており、そのアイディアは 1956 年に発明されたセマンティックネットワークに起源を持つとされている。

2012 年、Google は文字列のマッチングではなく、意味的な関連性を持った検索結果を提供するために、「Google Knowledge Graph」と呼ばれる知識グラフを組み込んだ検索機能を発表した[7]。図 3 の示す通り、ウェブ上の情報を基に構築された知識グラフは、検索時に表示されるナレッジパネルに反映され、これによって「Google Knowledge Graph」はより関連性の高い検索結果を提供することができるようになった。その結果、「Google Knowledge Graph」は世界的に最も知られた知識グラフとなったと同時に、知識グラフの一般的な認知度を大きく高めた。それ以降、DBpedia や Wikidata など、多くの知識グラフや知識ベースが公開されている。近年、知識グラフは、セマンティック検索、質疑応答システム、推薦システム、言語理解の補助、ビッグデータ分析の補助、機械学習モデルの解釈性向上など、多様な分野でますます重要な役割を果たしている。

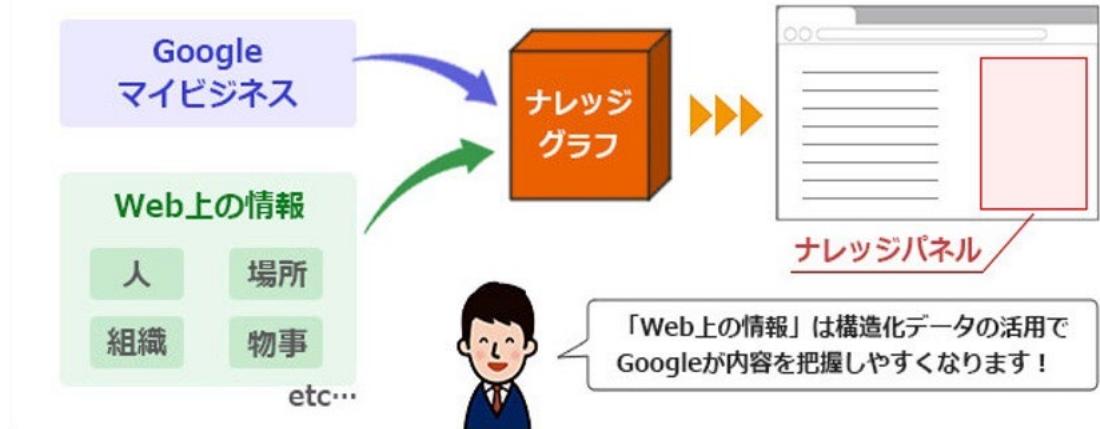


図3 Google Knowledge Graph の仕組み[8]

知識グラフの構築方法は、構築元の情報源の性質により、大別して構造化データからの構築と非構造化データからの構築の2種に分けられる。構造化データとは、特定の構造を持ったデータのことである。例えば、リレーションナルデータベースに管理されるデータは、一般的にカラムによって格納するデータの構造を決定するので構造化データである。ウィキペディアのインフォボックスやテーブルも構造化データの一種として数えられる。非構造化データとは特定の構造を持たないデータのことであり、テキストや画像、動画が該当するが、知識グラフは主にテキストデータを構築の対象とする。構造化データからの構築は、キーと値のペア（Key-value pair）を用いて比較的簡単に実行できるのに対して、非構造化データからの構築は情報抽出（Information Extraction）という分野に該当し、自然言語の多様性や必要情報の取捨選択といった難点が存在するため現在多くの研究が進められている[9]。本研究の研究対象も、非構造化データの肝細胞がん診療ガイドラインである。非構造化データから知識グラフを構築するには、前述した知識グラフの基本的な構成単位であるトリプルの抽出が鍵となり、非構造化データから抽出したトリプルに基づいて知識グラフは簡単に構築できる。トリプル抽出は、テキストから実体を抽出する固有表現抽出（Named Entity Recognition; NER）と二つの実体の間の関係を抽出する関係抽出（Relation Extraction; RE）の二つのタスクに分解され得る[10]。近年、深層学習技術を活用して固有表現抽出と関係抽出を行うことによって非構造化データからトリプルを抽出し、知識グラフを構築することがメインストリームとなっている。

## 2.2 質疑応答システム

質疑応答システムは膨大な情報から迅速かつ正確に特定の質問に答えることによって情報検索の時間を大幅に削減し、知識のアクセシビリティを高めることができるため、本研究では、肝細胞がん診療ガイドラインからの情報獲得の効率性と利便性を向上させる方法として質疑応答システムの構築を提案する。本節では、質疑応答システムの概要について説明する。

質疑応答システムとは、人間の自然言語による質問に対して自動的に回答を生成するシステムである。図 4 に段落の索引作成とランキング、候補回答の抽出と最終回答の出力の三つのモジュールから構成されるウィキペディアを利用した質疑応答システムを示している。ユーザーからの質問を受け取ると、この質疑応答システムではまず、段落の索引作成とランキングモジュールがウィキペディアデータベースの中から質問と関係がありそうな文書を検索し、その中から質問に関連する上位の段落をいくつか選択する。次に、候補回答抽出モジュールが与えられたテキストから複数の候補回答を抽出し、最後に最終回答出力モジュールが最も正しそうな予測を回答として生成する。例えば、上記の一連の処理を経て「Who was the next British Prime Minister after Arthur Balfour?」という質問に対する答えとして「Henry Campbell-Bannerman」が出力される。

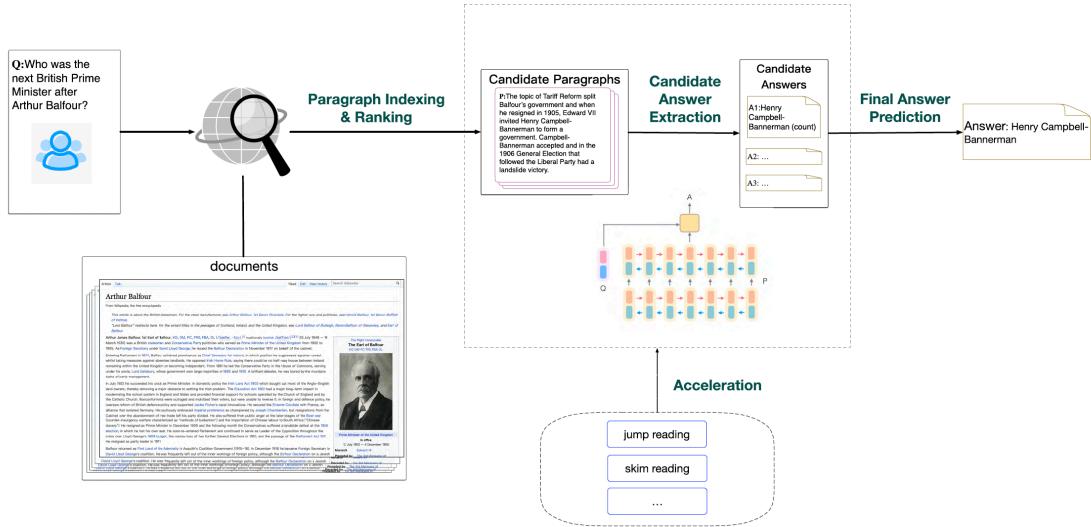


図 4 ウィキペディアの知識を用いた質疑応答システムの例[11]

質疑応答システムの急速な発展は主にここ 20 年に達成されたが、その概念が最初に現れたのは人工知能の黎明期、すなわち有名なチューリングテストにさかのぼることができる。初期の質疑応答システムは、限られた計算資源により、1961 年に開発された「Baseball」や 1977 年に開発された「Lunar」のように手動で設計された構文規則を主に使用し、単純な回答しか得られなかつた。2000 年頃から、TREC や CLEF、NTCIR などの質疑応答に関する国際コンテストが開催され、質疑応答システムの発展を大きく促進した。

2011 年に、IBM 社が開発した質疑応答システム「Watson」が「Jeopardy!」というクイズ番組で人間を打ち負かして優勝し、大きな注目を集めた[12]。このシステムはまずキーワードクエリに基づいて関連文書を特定し、その後固有表現認識および関係抽出技術を使用して回答を抽出する。しかし、このアプローチにはキーワードクエリの品質と固有表現認識の精度に大きく依存し、回答の質が不安定であるという弱点があることを後に指摘された[13]。

近年、深層学習技術が大きく発展し、SQuAD や CoQA などの質疑応答に関する大規模ベンチマークデータセットの公開を契機に、質疑応答システムの構築において深層学習技術を取り入れる流れが進んでいる。従来の単語頻度-逆文書頻度 (Term Frequency-Inverse Document Frequency ; TF-IDF) や潜在意味解析 (Latent Semantic Analysis ; LSA) などの情報検索技術 (Information Retrieval) に基づいた質疑応答システムと比較し、深層学習技術を用いた質疑応答システムは、SQuAD や CoQA を含む多数のベンチマークデータセットにより優れた性能を示しており、質疑応答システムの構築においては既に主流となっている[13]。本研究でも、深層学習技術を用いて質疑応答システムの構築を試みることにする。

## 2.3 基礎的な深層学習技術

2.1 節と 2.2 節に述べたように、深層学習技術は近年大きな進展を遂げており、知識グラフと質疑応答システムの主流的な構築方法も徐々に深層学習に基づくアプローチへと移行している。本研究では、肝細胞がん診療ガイドラインに関する知識グラフと質疑応答システムを実現するにあたって、自然言語処理分野の諸タスクで高い性能を示した深層学習モデルをベースに利用する。本節ではその基礎的な部分について説明する。深層学習とは、多層化したニューラルネットワーク (Deep Neural Network) を用いた機械学習のことであり、その基本となるものは、人間の脳の神経細胞の仕組みを再現した人工ニューラルネットワーク (Artificial neural network) というモデルである。人工ニューラルネットワークは、神経細胞 (ニューロン) を数理モデル化した人工ニューロン[14]を多層に連結させた構造を持ち、学習を通じて与えられたタスクを模倣することが可能である。

### 2.3.1 単純パーセプトロン

はじめに人工ニューラルネットワークの基本ユニットである人工ニューロンについて説明する。この人工ニューロンは単純パーセプトロンとも呼ばれ、生物の神経細胞を数学的にモデル化したものである。図 5 に示すように、単純パーセプトロンは複数のシグナル  $x$  を入力として受け取り、单一のシグナル  $\hat{y}$  を出力する。単純パーセプトロンは数式で表すと次のように定義される。

$$x = (1, x_1, x_2, \dots, x_m) \quad (1)$$

$$w = (w_0, w_1, w_2, \dots, w_m) \quad (2)$$

$$z = w_0 + \sum_{i=1}^m w_i x_i \quad (3)$$

$$g(z) = \begin{cases} 1 & (z \geq 0) \\ 0 & (z < 0) \end{cases} \quad (4)$$

$$\hat{y} = g(z) \quad (5)$$

ここで $x$ は単純パーセプトロンの入力シグナルを表し、 $w$ は各入力に対する重み成分である。 $w_0$ は1という入力に対応する重み成分であるため、単純パーセプトロンの持つバイアス成分とみなすことができる。各入力 $x$ に対して $w$ をかけて合計 $z$ を計算し、活性化関数の $g$ に代入する。この活性化関数は、単位ステップ関数が使用されており、 $z$ を受け取って $z \geq 0$ ならば1を出力し、そうでなければ0を出力する。この処理は「発火する」と表現される。単純パーセプトロンの最終的な出力 $\hat{y}$ は、活性化関数の出力 $g(z)$ で決められる。この出力 $\hat{y}$ と本当のラベルである $y$ の誤差を計算し、誤差が減少するように $w$ を調整することにより、パーセプトロンは線形分離可能な二値分類問題を解くことが可能である。

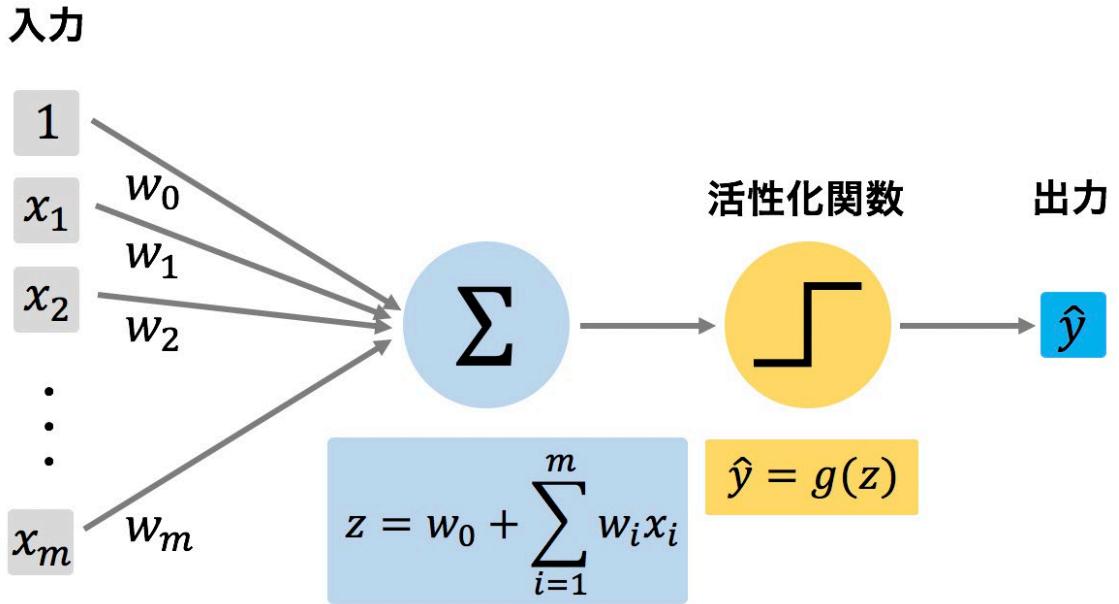


図5 単純パーセプトロンのモデル図

### 2.3.2 多層パーセプトロン

単純パーセプトロンにデータを線形的にしか分類できないという欠点がある。しかし、複数のパーセプトロンを適当に連結させ、パーセプトロンを多層化することで、線形分離が不可能なタスクも扱えるようになる。このような構造は多層パーセプトロン[15]と呼ばれる。図6は3層パーセプトロンのモデル構造を示すものである。図に示すように多層パーセプトロンは順伝播型ネットワークであり、このモデルではノードはモデルへの入力や

パーセプトロンの出力といった値を表し、隣接の二つのノードを結ぶリンクは重みを表す。

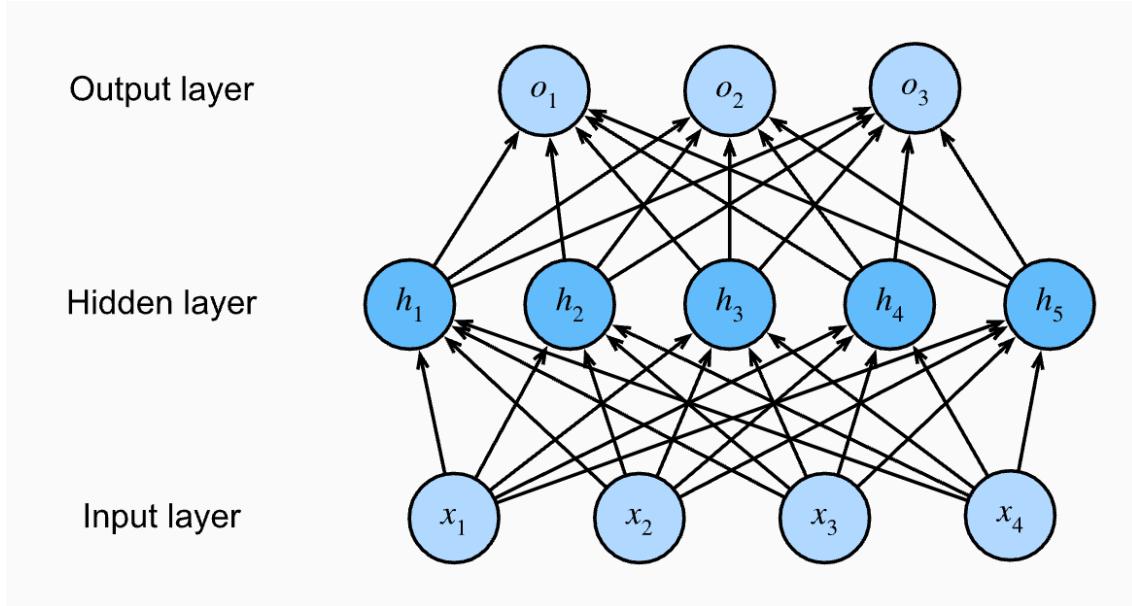


図 6 3層パーセプトロンの構造[16]

図 6 にも示しているように、多層パーセプトロンでは、入力を受け取るパーセプトロン層を入力層、最終的な出力を行うパーセプトロン層を出力層、それ以外の中間処理を行う層をまとめて隠れ層と呼ぶ。各パーセプトロン層は複数の人工ニューロンで構成され、各ニューロンは前の層の各パーセプトロンから出力を受け取り、個別に重みとバイアスを持つ。単純パーセプトロンと異なり、多層パーセプトロンではより多くの種類の活性化関数が使用される。最も基本的なのが、下記のように定義されるシグモイド関数 (Sigmoid Function) である。

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (6)$$

シグモイド関数は実数全体 $(-\infty, \infty)$ を定義域に持ち、 $(0, 1)$ を値域とし、入力の絶対値が大きくなると出力が飽和し一定値となることが特徴である。これとは対照的で、入力がマイナスでない限り入力が大きくなるにつれて出力も大きくなるという特徴を持つReLU 関数 (Rectified Linear Unit) も多層パーセプトロンの活性化関数としてよく使用される。ReLU 関数は式 (7) で定義される。また、ReLU の拡張がさまざまに行われており、代表的な例として LeakyReLU や GELU (Gaussian Error Linear Unit) が挙げられ、それぞれ式 (8) と式 (9) のように定義される。式 (9) にある  $\phi$  は標準正規分布の累積分布関数を表す。

$$ReLU(x) = \max(x, 0) \quad (7)$$

$$LeakyReLU(x) = \max(x, 0.01 * x) \quad (8)$$

$$GELU(x) = x * \phi(x) \quad (9)$$

多層パーセプトロンを使用して線形分離不可能なタスクを解決するには、各パーセプトロンのパラメータである重み  $w$ （バイアスも重みの一種とみなせる）を学習によって適切な値に調整する必要がある。多層パーセプトロンの各重みの学習に誤差逆伝播法（Back Propagation）と呼ばれる学習アルゴリズムが使用される。この過程では、多層パーセプトロンの出力と真の出力との間の誤差を最小限に抑えるように、各パーセプトロンの重みが更新される。

### 2.3.3 誤差逆伝播法

ニューラルネットワークの学習は、入力データに対するネットワークの出力と正解との間の、損失関数で測った損失に基づいて各層のパラメータを調整するプロセスである。ニューラルネットワークの学習を行うには、各パーセプトロンのパラメータである重みとバイアスについて損失関数の偏微分を計算する必要がある。しかし、この偏微分の計算は、特に入力に近い深い層のパラメータになると、微分の連鎖則を何度も繰り返す必要があり計算にかかる手間も多くなる。各層の重みとバイアスについての偏微分計算を効率よく実行する方法として、誤差逆伝播法[17]というアルゴリズムが一般的に使用される。誤差逆伝播法ではまず、訓練データをニューラルネットワークに入力した際のネットワークの出力と正解データとの間の差に基づいて損失を計算し、その損失を出力層から逆向きに伝播させることで各層の損失を算出する。算出した損失から損失が減少していく方向を求め、各層の重みとバイアスを更新することによって、訓練データに基づいて正しい出力を生成するようにニューラルネットワークを徐々に学習する。このパラメータの更新は、下記の式 (10) で定義される。この式では、 $w_{ji}^{(l)}$  は第  $l$  層の  $j$  番目のニューロンと第  $l - 1$  層の  $i$  番目のニューロンの結合重みであり、 $\eta$  はあらかじめ設定された学習率であり、 $\partial E / \partial w_{ji}^{(l)}$  は損失  $E$  の各層  $l$  のパラメータについての微分である。

$$w_{ji}^{(l)} = w_{ji}^{(l)} - \eta \frac{\partial E}{\partial w_{ji}^{(l)}} \quad (10)$$

ここで問題となるのは、右辺の  $\partial E / \partial w_{ji}^{(l)}$  をどのように求めるかということである。 $\partial E / \partial w_{ji}^{(l)}$  を求めるために、まずそれを式 (11) のように展開できる。

$$\frac{\partial E}{\partial w_{ji}^{(l)}} = \frac{\partial E}{\partial u_j^{(l)}} \frac{\partial u_j^{(l)}}{\partial w_{ji}^{(l)}} \quad (11)$$

ここで  $u_j^{(l)}$  は第  $l$  層の  $j$  番目のニューロンの入力である。多層パーセプトロンの原理より、入力  $u_j^{(l)}$  の変動が損失  $E$  に与える影響は、第  $l$  層  $j$  番目のニューロンからの出力  $z_j^{(l)}$  を通じて、第  $l+1$  層の各ニューロンの総入力を変化させることによってのみ生じることがわかる。したがって、第  $l+1$  層の各ニューロンへの入力を  $u_k^{(l+1)}$  として各  $u_k^{(l+1)}$  を経由した微分の連鎖により式 (11) の右辺の  $\partial E / \partial u_j^{(l)}$  を式 (12) のように分解できる。

$$\frac{\partial E}{\partial u_j^{(l)}} = \sum_k \frac{\partial E}{\partial u_k^{(l+1)}} \frac{\partial u_k^{(l+1)}}{\partial u_j^{(l)}} \quad (12)$$

この式の両辺に第  $l$  層と第  $l+1$  層の入力に関する微分が現れているので、 $\delta_j^{(l)} \equiv \frac{\partial E}{\partial u_j^{(l)}}$  のように置くことができる。 $\delta_j^{(l)}$  は各層  $l$  の各ニューロン  $j$  に対して定義される量である。式 (13) 、 (14) に示す関係より、 $\delta_j^{(l)}$  を使用して式 (12) を式 (15) のように書き換えることができる。

$$u_k^{(l+1)} = \sum_j w_{kj}^{(l+1)} z_j^{(l)} = \sum_j w_{kj}^{(l+1)} f(u_j^{(l)}) \quad (13)$$

$$\frac{\partial u_k^{(l+1)}}{\partial u_j^{(l)}} = w_{kj}^{(l+1)} f'(u_j^{(l)}) \quad (14)$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)} \left( w_{kj}^{(l+1)} f'(u_j^{(l)}) \right) \quad (15)$$

この  $f$  は、ニューラルネットワークの使用する活性化関数である。式 (15) は、 $\delta_j^{(l)}$  が  $\delta_k^{(l+1)}$  ( $k = 1, 2, 3 \dots$ ) から計算できることを意味する。つまり、上位の  $l+1$  層のニューロンの  $\delta_k^{(l+1)}$  が与えられれば、下位の  $l$  層の  $\delta_j^{(l)}$  は式 (15) に従って簡単に計算できるということである。式 (15) は中間層のどの層についても成立するので、中間層のどの  $l$  層の  $\delta_j^{(l)}$  でも上位層の  $\delta_k^{(l+1)}$  を使えば計算できるはずである。最初に出力層の各ニューロンの  $\delta$  が求まつていれば、式 (15) を繰り返して適用することで下位層の任意のニューロンにおける損失  $E$  のニューロン入力に関する微分が逆向きに計算できることになる。

また、式 (11) の第二項の  $\partial u_j^{(l)} / \partial w_{ji}^{(l)}$  は、式 (16) の関係から式 (17) のように計算できる。

$$u_j^{(l)} = \sum_i w_{ji}^{(l)} z_i^{(l-1)} \quad (16)$$

$$\frac{\partial u_j^{(l)}}{\partial w_{ji}^{(l)}} = z_i^{(l-1)} \quad (17)$$

ここでの  $z_i^{(l-1)}$  は、第  $l-1$  層のニューロン  $i$  の出力を表す。式 (12) 、 (15) 、 (17) を使用して式 (11) は以下のように表現することができる。

$$\frac{\partial E}{\partial w_{ji}^{(l)}} = \delta_j^{(l)} z_i^{(l-1)} \quad (18)$$

式 (16) のように、第  $l-1$  層のニューロン  $i$  と第  $l$  層のニューロン  $j$  を繋ぐ結合重み  $w_{ji}^{(l)}$  に関する微分は、ニューロン  $j$  に関する  $\delta_j^{(l)}$  とニューロン  $i$  からの出力  $z_i^{(l-1)}$  との積で与えられる。この関係を使用して式 (10) のようにニューラルネットワークのパラメータを比較的に簡便に更新できる。

上記の説明をまとめると、訓練サンプル  $x$  および正解データ  $d$  を入力とし、損失  $E$  の各層  $l$  のパラメータについての微分  $\partial E / \partial w_{ji}^{(l)}$  を出力とする時、誤差逆伝播法は以下の手順によって行われる。

1.  $z^1 = x$  とし、各層  $l$  ( $l = 2, 3, \dots, L$ ) のニューロンの入力  $u^{(l)}$  および出力  $z^{(l)}$  を順に計算する。 (順伝播)
2. 出力層  $L$  で  $\delta_j^{(L)}$  を求める。
3. 各中間層  $l$  ( $l = L-1, L-2, \dots, 2$ ) での  $\delta_j^{(l)}$  を、この順に式 (15) にしたがって計算する。 (逆伝播)
4. 各層  $l$  ( $l = 2, 3, \dots, L$ ) のパラメータ  $w_{ji}^{(l)}$  に関する微分を式 (18) にしたがって計算する。

上記の手順により計算された  $\partial E / \partial w_{ji}^{(l)}$  に事前に設定した学習率をかけてことで、パラメータが更新され、ニューラルネットワークの学習が実現される。

### 2.3.4 回帰型ニューラルネットワーク

深層学習の分野では、テキストやシグナルといった系列データを対象とするタスクがよくある。本節では系列データに関連するタスクを扱う代表的な構造として回帰型ニューラルネットワーク (Recurrent Neural Network) について紹介する。回帰型ニューラルネットワークとは、内部に有向閉路を持つニューラルネットワークの総称で、テキストや時間信号のような系列データのパターン認識に特化したニューラルネットワークである。

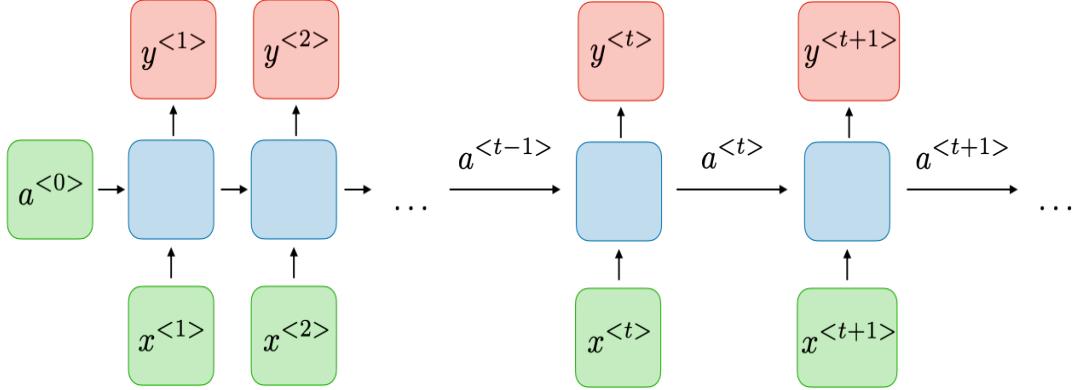


図 7 回帰型ニューラルネットワークのモデル図[18]

図 7 に典型的な回帰型ニューラルネットワークのモデル構造を示す。図 7 からわかるように、多層パーセプトロンと同様に回帰型ニューラルネットワークも順伝播型のネットワークである。ただし、従来の多層パーセプトロンでは全ての入力データが同じ時刻にネットワークに入力されるのに対し、回帰型ニューラルネットワークでは主な処理対象が系列データであるので、入力データとなるシーケンスを複数のタイムステップに分けて各タイムステップに一つだけ入力するという構造になっている。また、ネットワークの中に中間層の出力が自分自身に戻される「帰還路」が作られるので、回帰型ニューラルネットワークにおける中間層（＝隠れ層）が受け取る入力は、現時刻の前層からの出力と、自分自身に戻される一つ前の時刻の中間層出力という二つのものである。これにより、ネットワークの各時刻における最終的な出力は、過去に受け取った全ての入力から一定の影響を受けることになり、回帰型ニューラルネットワーク自体も過去のすべての入力から一つの出力への写像を表すとみなすことができる。このような構造により、回帰型ニューラルネットワークが情報を一時的に記憶し、系列データの中に存在する「文脈」を捉えることができようになり、一般的な多層パーセプトロンよりも系列データの処理に強いとされている。

回帰型ニューラルネットワークの一時刻における処理に着目し、その具体的な構造を図 8 に示す。図 8 からわかるように、中間層は実は、重みやバイアスといったパラメータを持つ複数のパーセプトロンによって構成されており、入力  $x^{<t>}$  はこれらのパーセプトロンの処理を経て中間層の出力  $a^{<t>}$ 、さらに最終的な出力  $y^{<t>}$  に変換される。各時刻における中間層出力  $a^{<t>}$  および各時刻における最終出力  $y^{<t>}$  は以下のように定義される。

$$a^{<t>} = g_1(W_{aa}a^{<t-1>} + W_{ax}x^{<t>} + b_a) \quad (19)$$

$$y^{<t>} = g_2(W_{ya}a^{<t>} + b_y) \quad (20)$$

ここで  $W_{aa}$ 、 $W_{ax}$ 、 $W_{ya}$  が重みであり、 $b_a$  と  $b_y$  がバイアス項である。重みとバイアスはすべての時刻  $t$  において共有され、学習によって更新されるものの順伝播計算中は定数であ

る。 $g_1$ と $g_2$ は活性化関数であり、回帰型ニューラルネットワークの活性化関数として伝統的にハイパボリックタンジェント関数が多く使われるが、ReLU関数も使われる。

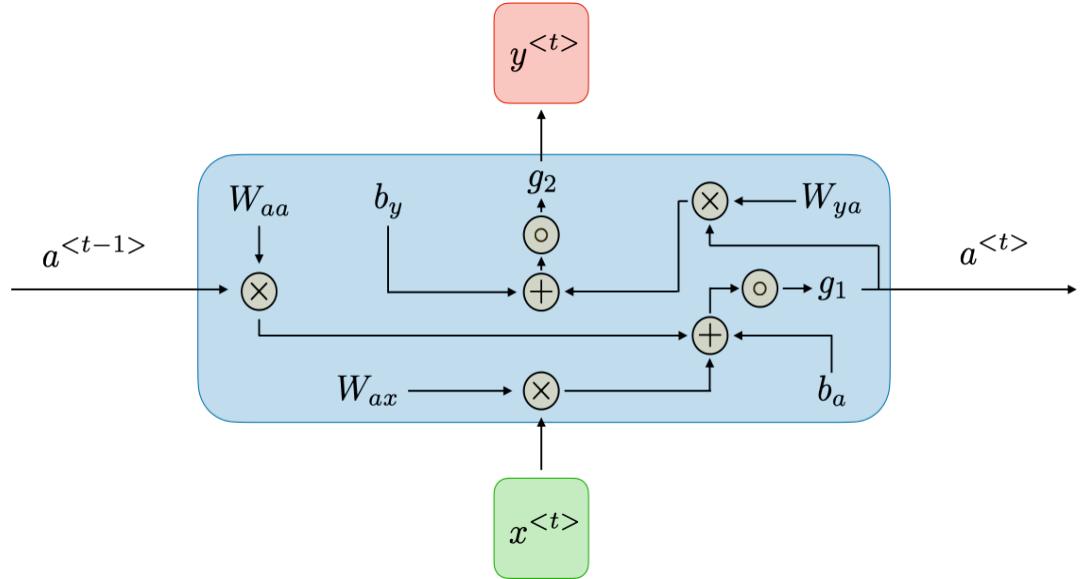


図8 回帰型ニューラルネットワークの中間層の内部構造[18]

また、タスクによって系列データが一括して与えられ、データに存在するパターンの認識に実時間性を必要としない場合もある。例えば、自然言語理解において一文の中の個々の単語は前の単語のみならず後ろの単語からも影響を受けることがある。そのようなタスクに特化したモデルとして、双方向性回帰型ニューラルネットワーク (Bidirectional Recurrent Neural Network) が提案される。双方向性回帰型ニューラルネットワークとは、順向きのシーケンスを入力にとる回帰型ニューラルネットワークと、逆向きのシーケンスを入力にとる回帰型ニューラルネットワークという二つのネットワークを置き、両者の出力層を統合することによって、同じシーケンスを二つの方向から処理することを実現させるモデルである。図9に示すように、双方向性回帰型ニューラルネットワークは二つの単方向の回帰型ニューラルネットワークを組み合わせたものであり、双方向性回帰型ニューラルネットワークにおける中間層の構造は单方向性のものと同じである。各時刻 $t$ における双方向性回帰型ニューラルネットワークの出力 $\hat{y}^{<t>}$ は、順向きと逆向きのネットワークのその時刻における出力を連結させて得られる。一般的に下記のように定義される。

$$\hat{y}^{<t>} = g(W_{ya}[a_{\leq t}^{<t>}, a_{\geq t}^{<t>}] + b_y) \quad (21)$$

ここで $a_{\rightarrow}^{<t>}$ と $a_{\leftarrow}^{<t>}$ はそれぞれ順向きと逆向きネットワークの中間層出力であり、 $g$ は活性化関数である。ただし、二つの中間層出力を連結させたものに重みとバイアスをかけず、活性化もせずにそのまま時刻の出力とする場合もある。

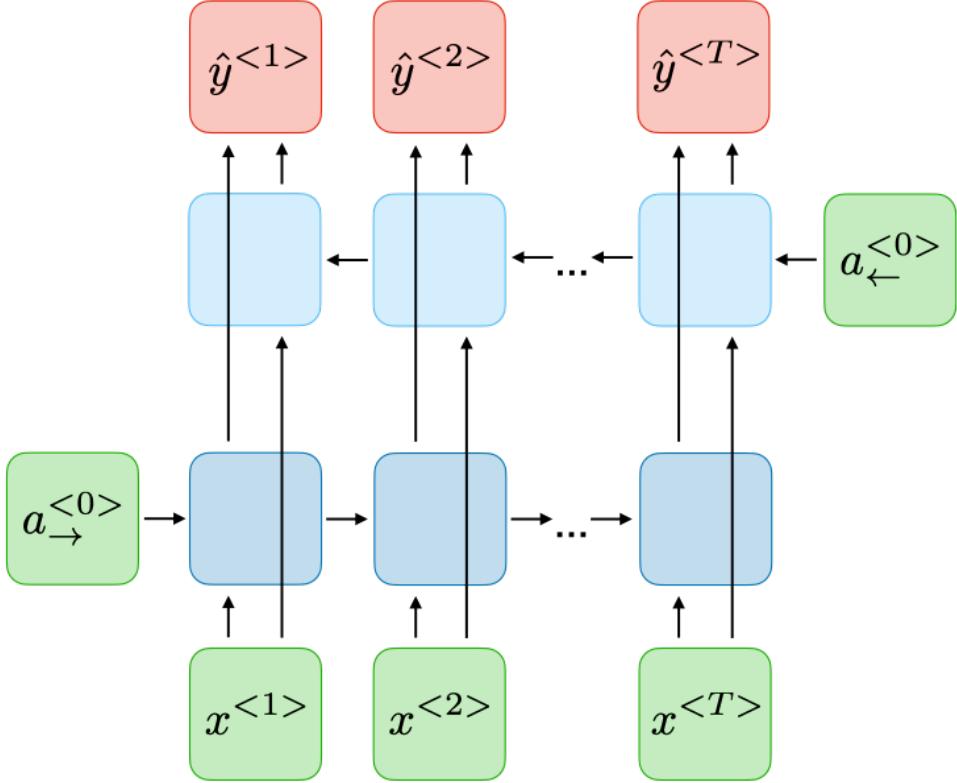


図9 双方向性回帰型ニューラルネットワークのモデル図[18]

### 2.3.5 長・短期記憶

2.3.4節に述べたように、回帰型ニューラルネットワークは過去の入力の履歴を最新の出力の計算に反映させることができる。理論的に、各時刻における出力に過去のすべての入力が関わるはずだが、実際に最新の出力に反映されたのは、せいぜい過去10時刻程度であると言われている[19]。この問題は、長期依存性 (Long-term dependency) 問題と呼ばれ、ニューラルネットワークの勾配消失問題に起因するものである。勾配消失問題は、層数の多い深いネットワークにおいて誤差逆伝播法で勾配を計算するとき、出力から層を逆向きに辿るために勾配の値が指数的に大きくなるか小さくなるという問題である。一時刻を一層とみなすと回帰型ニューラルネットワークもかなり深いネットワークであり、勾配消失の危険が高まるので、入力を短期的に記憶しておくことができても、それを長期にわたって記憶して出力の計算に利用することは難しい。

上記の問題を解決し、長期にわたる記憶を可能にするために考案されたのが長・短期記憶ユニット (Long Short-Term Memory Unit) である[20]。長・短期記憶ユニットは、三

種のゲート機構 (Gate Mechanism) を伝統的な回帰型ニューラルネットワークの中間層の中に追加したユニットであり、重要な情報を長期間記憶し不要な情報を忘却するという記憶の長期伝達を可能にする。追加された三種のゲート機構はそれぞれ、ユニットの保持している情報からどのくらい情報を廃棄するかを決める忘却ゲート、新しい情報をどのくらいにユニットに記憶させるかを決める入力ゲート、ユニットの保持している情報をどのように出力するかを決める出力ゲートである。図 10 に長・短期記憶ユニットの内部構造を示す。

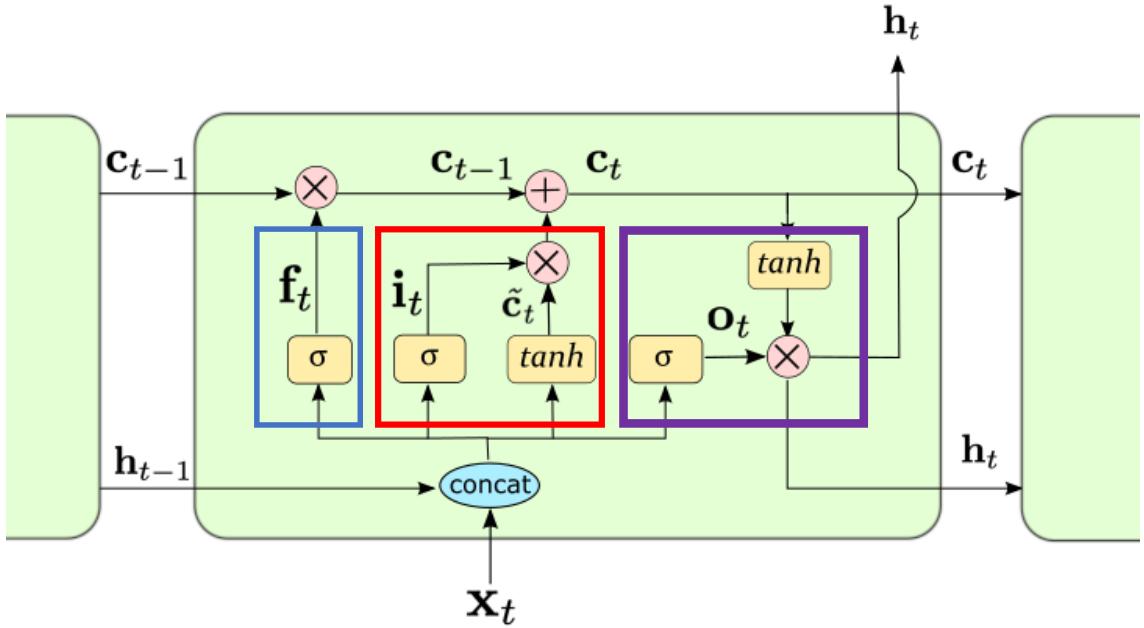


図 10 長・短期記憶ユニットの内部構造[21]

図 10 に示すように、長・短期記憶ユニットに、青色のブロックで囲まれた忘却ゲート、赤色のブロックで囲まれた入力ゲート、紫色のブロックで囲まれた出力ゲートの三つのゲート機構が順に配置される。各ゲート機構への入力は、各時刻の入力データ  $x_t$  と短期記憶  $h_{t-1}$  を連結させたものであり、最初の忘却ゲートだけ長期記憶  $c_{t-1}$  も入力として与えられる。忘却ゲートの処理は以下のように定義される。

$$f_t = \sigma(w_f[x_t, h_{t-1}] + b_f) \quad (22)$$

ここで  $w_f$  と  $b_f$  は重みとバイアスであり、 $\sigma$  は活性化関数のシグモイド関数を表す。 $f_t$  は忘却ゲートの出力である。次に、入力ゲートの処理は以下のように定義される。

$$i_t = \sigma(w_i[x_t, h_{t-1}] + b_i) \quad (23)$$

$$\tilde{c}_t = \tanh(w_h[x_t, h_{t-1}]) \quad (24)$$

ユニットの現在時刻の長期記憶  $c_t$  は、忘却ゲートの出力と入力ゲートの出力を用いてその重み付け和として次のように更新する。

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t \quad (25)$$

最後に、出力ゲートの処理によって現在時刻のユニット出力  $h_t$  が次式のように決定される。

$$o_t = \sigma(w_o[x_t, h_{t-1}] + b_o) \quad (26)$$

$$h_t = o_t \cdot \tanh(c_t) \quad (27)$$

このような長・短期記憶ユニットを中間層として基本的な回帰型ニューラルネットワークの構造に取り入れたニューラルネットワークは長・短期記憶ネットワーク (Long Short-Term Memory Network) と呼ばれる。長・短期記憶ネットワークは基本的な回帰型ニューラルネットワークの短期間の記憶しか実現できないという限界を緩和することができる。最も単純な場合、忘却ゲートを 1 (オープン) にし、入力ゲートを 0 (クローズ) にし続ければ、長・短期記憶ユニットの状態は永遠に記憶されることになる。もちろん、同じ状態を長く保持するのではなく適切なタイミングでこれらのゲート機構を開閉させる必要がある。ゲートの開閉をネットワークの学習を通して調整することにより、長い文脈を捉えてより高い性能を達成することが可能になる。また、双方向性回帰型ニューラルネットワークの中間層を長・短期記憶ユニットで置き換えることで、図 11 に示すように同じシーケンスを順向きと逆向きの二つの方向から処理する双方向性長・短期記憶ネットワーク (Bidirectional Long Short-Term Memory Network) もある。

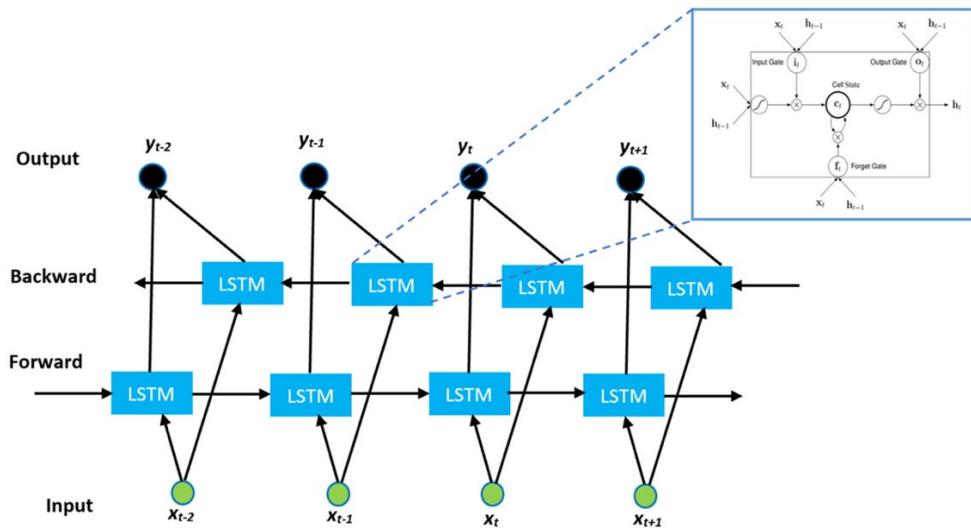


図 11 双方向性長・短期記憶ネットワークのモデル図[21]

## 2.4 自然言語処理における深層学習技術

2.3 節では、深層学習技術の基礎的な部分について説明した。近年、多層パーセプトロンや回帰型ニューラルネットワークのような基礎的な深層学習技術をさらに発展させる流れが進んでおり、自然言語処理に特化した深層学習技術も多数開発されている。自然言語処理は、計算機が人間の言語を理解し処理することを目的とし、自然言語の多様性や複雑性から難しい分野とされている。しかし、深層学習技術の隆盛とともに自然言語処理分野が大きく発展し、自然言語処理に関連するタスクにおいて人間に匹敵するほどの高度な能力を持つ深層学習モデルも現れた。本節では、自然言語処理に特化した深層学習技術の中の代表的なものについて紹介する。自然言語処理に特化した深層学習技術とこれまでの基礎的な深層学習技術との大きな違いとして、まずニューラルネットワークの構造が挙げられる。昔の深層学習モデルはせいぜい数十層の程度のニューラルネットワークで構築されてパラメータ数も数百万程度しか持っていないのに対し、近年ではそれを数百や数千層まで拡張してモデルに含まれるパラメータ数を億単位で増やすというモデルの大規模化が進んでいる。また、近年新しいニューラルネットワークの構造も考案されており、特に自然言語処理分野では、セフレアテンションという注意機構をベースとしてモデルを構築する動きがすでに一般的となっている。さらに、深層学習モデルの学習方法にも大きな変革が起きている。これまでのモデル学習では、特定のタスクに対応した教師データを用いて訓練したモデルをそのまま対応したタスクの処理に充てるというプロセスだったが、現在自然言語処理分野においては、モデルの学習プロセスは通常、特定のタスクに無関係な事前学習と、特定のタスクにモデルを適応させるファインチューニングの二段階で行われる。上記の進展により、モデル自身の表現能力も大幅に拡充され、自然言語理解 (Natural Language Understanding) や自然言語生成 (Natural Language Generation) といった高度で複雑なタスクに対しても深層学習モデルは高い性能を達成できるようになる。

### 2.4.1 Word2Vec

前述したように、ニューラルネットワークは多数のパーセプトロンから構成されるものであり、その内部にベクトルや行列演算が行われている。一方で、自然言語処理分野の主な処理対象は、単語や文字といった記号や複数の記号からなる文または文章であり、そういったものをそのままニューラルネットワークに入力することができない。自然言語をニューラルネットワークで処理するためには、単語や文字といった記号を実数値ベクトルに変換する必要がある。このような単語に対応する数値ベクトルは、単語埋め込み (Word Embedding/Word Representation) と呼ばれる。単語をニューラルネットワークの扱える単語埋め込みに変換する技術として、2013年にWord2vec (Word to Vector) という技術 [22]が提案された。

Word2vecは深層学習技術を用いて文章に含まれる単語をその意味を反映する数値ベクトルに変換する技術である。Word2vec技術にCBOWとskip-gramという2種類のモデルがあり、その構造を図12に示す。図12にあるように、Word2vecのどのモデルもただ2層の二

ューラルネットワークのみで構成されており、構造的にシンプルであり計算量も現実的なレベルに抑えられる。

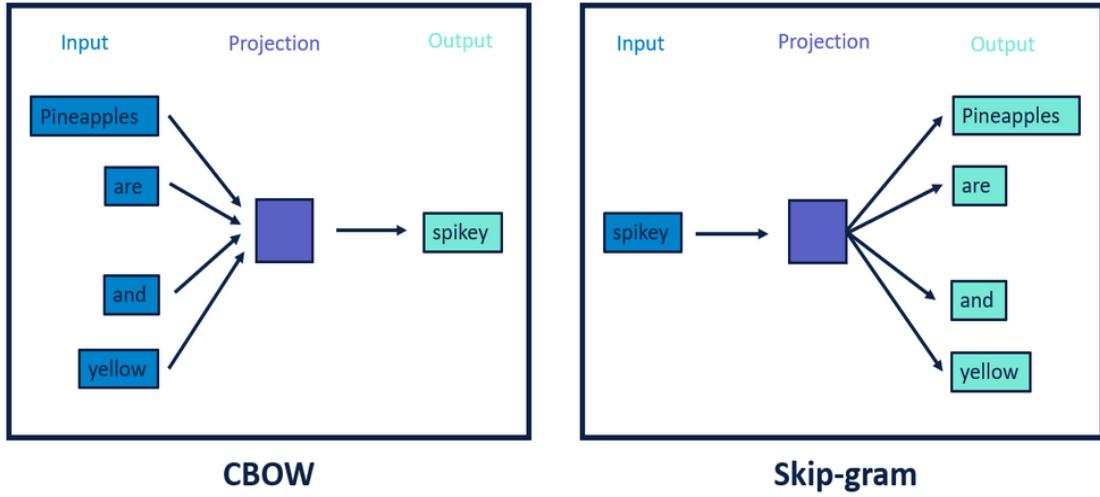


図 12 Word2vec モデルの構造 (CBOW および skip-gram) [23]

CBOW と skip-gram という二つのモデルは、与えられた文脈からターゲットとなる単語の生起確率を推定することを共通点として持つ。しかし、CBOW では予測のターゲットとなるのは一単語のみであり、その単語の前の文と後ろの文からなる単語リストを入力として受け取ってその単語がどの単語なのかを予測するのに対し、skip-gram では入力は一単語のみで、その単語の前の文と後ろの文としてどの単語が出現し得るかを予測する。つまり、CBOW モデルでは周辺語から中心語を予測することがモデル学習のタスクであり、図 12 に示すように「Pineapples」、「are」、「and」、「yellow」といった入力から、文の中心に現れる「spikey」という単語を予測する。skip-gram モデルでは中心語から周辺語を予測することがタスクであり、図 12 の例では与えられた「spikey」から「Pineapples」、「are」、「and」、「yellow」といった周辺語を予測する。

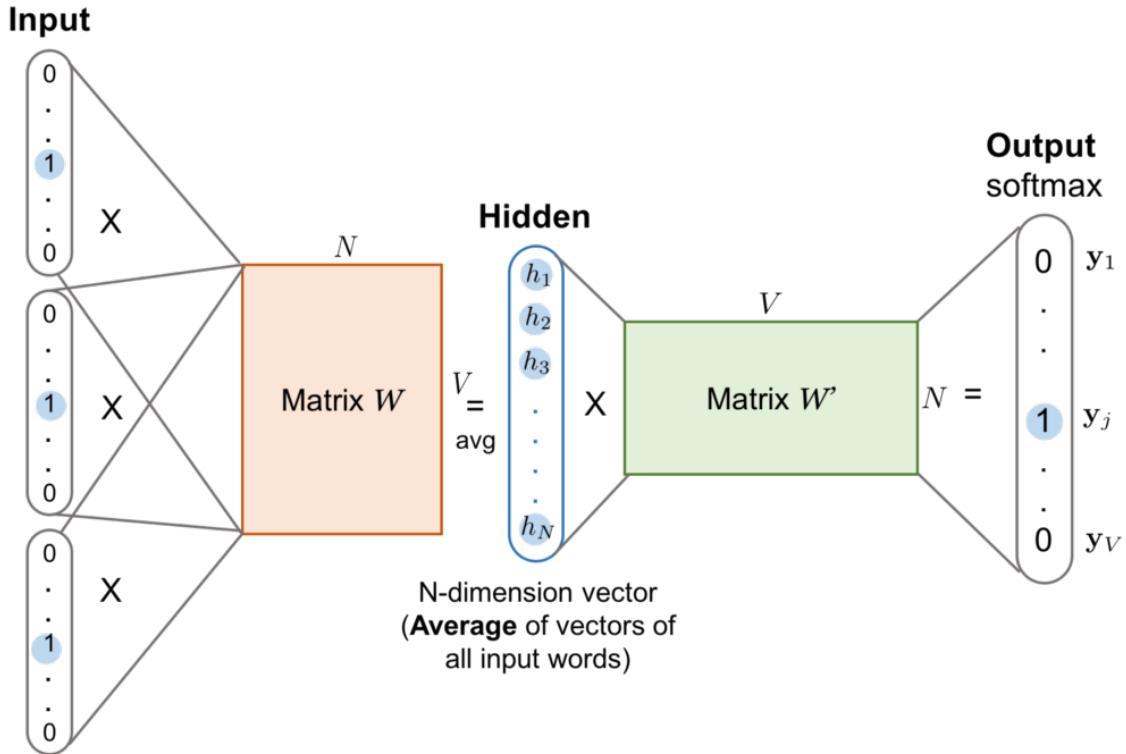


図 13 CBOW のモデル構造[24]

CBOW モデルを用いて周辺語から中心語を予測するという学習タスクを行うことにより、図 13 に示すように  $W$  という行列を単語埋め込みとして生成できる。この行列は  $V \times N$  の行列であり、 $V$  は語彙数、 $N$  は単語埋め込みの次元数を表し、各行ベクトルは語彙中の一つの単語を表現する。一方で、skip-gram モデルを用いて単語埋め込みを生成する方法は図 14 に示す。学習タスクを行うことにより生成される単語埋め込みは  $W$  であり、同様に  $V \times N$  の行列で各行に一つの単語の埋め込みベクトルが格納される。

Word2vec 技術によって生成された単語埋め込みの強さは、単語間の多様な意味関係をベクトルの演算によって捉えることができるという点にある。例えば、「マドリード」という単語の埋め込みベクトルから「スペイン」の埋め込みベクトルを引いた結果に「フランス」の埋め込みベクトルを足すと、できたベクトルは「パリ」の埋め込みベクトルに近いところに来る。これは、単語埋め込みは「フランス」と「パリ」に存在する首都という関係を捉えることができるからだ。

Word2vec 技術は言語に拘らず共通して使える単語埋め込み変換技術である。Word2vec 技術の登場により、大規模なコーパスによる単語埋め込み学習を現実的な計算量で行えるようになり、文章要約や機械翻訳といった自然言語処理に関するタスクにおいて深層学習技術を取り入れることも可能になった。

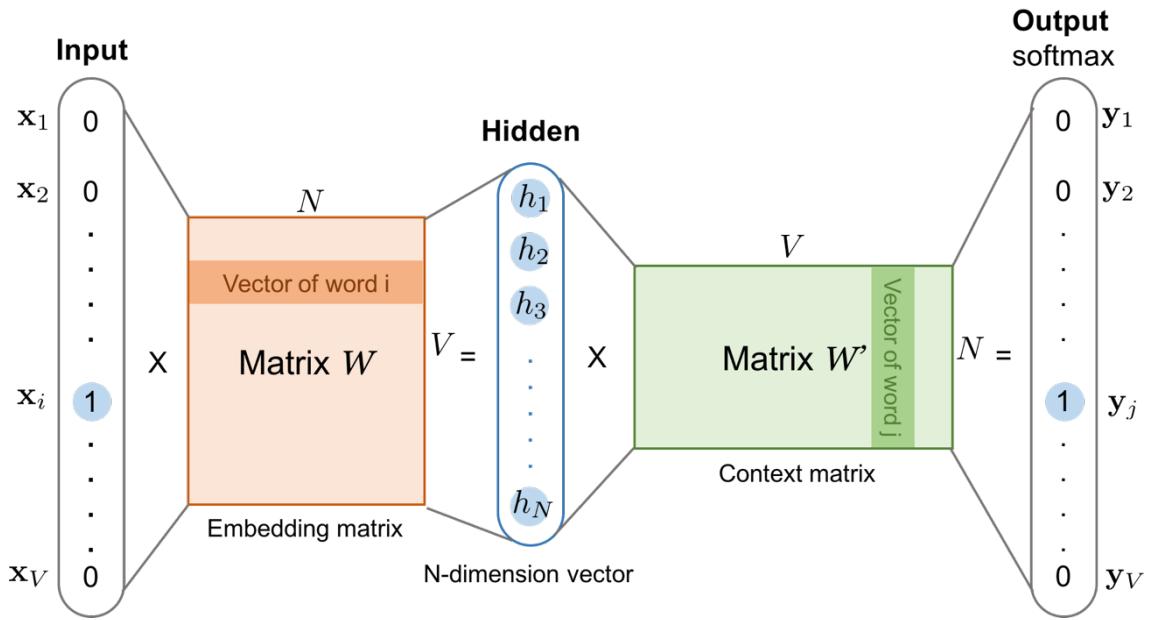


図 14 skip-gram のモデル構造[24]

#### 2.4.2 トランスフォーマー

回帰型ニューラルネットワークに、逐次的に単語を処理するが故に学習時に並列処理ができず時間がかかるという大きな欠点がある。この問題を解決するために、2017年にトランスフォーマー[25]というニューラルネットワーク構造が Google Brain によって提案され、自然言語処理分野に大きな変革をもたらした。トランスフォーマー (Transformer) とは、マルチヘッド注意 (Multi-head Attention) を主要部品として使用し、機械翻訳や構文解析といった系列変換タスク (Sequence-to-Sequence task) に特化した深層学習モデルである。計算効率性や高性能性、スケーラビリティといった長所を同時に備え、トランスフォーマーは従来の系列データ処理向けの回帰型ニューラルネットワークや長・短期記憶ネットワークに代わるものとして注目を集めた。マルチヘッド注意による並列処理を実現したトランスフォーマーでは、一回に一単語 (=1 トークン) しか入力できないという回帰型ニューラルネットワークの制限を乗り越え、シーケンス全体 (=数百トークン) を同時にモデルに入力するが可能となる。モデルの出力は、出力シーケンスを 1 トークンずつ予測していくという仕組みとなっている。

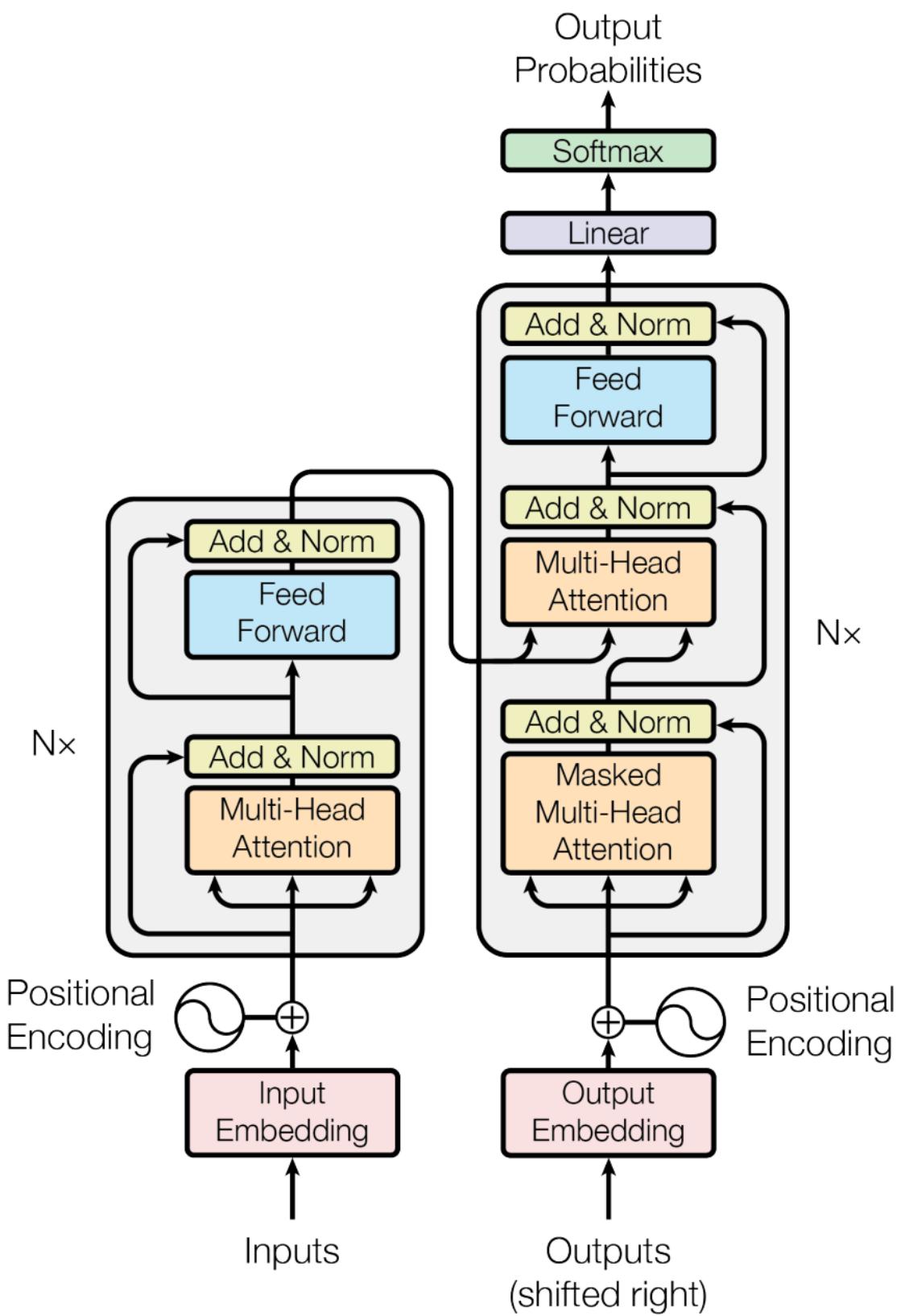


図 15 トランスフォーマーのモデル構造[25]

トランسفォーマーの最も特徴的なのは、マルチヘッド注意という注意機構だけ使用することで入力と出力の文章同士の広範囲な依存関係を捉え、系列変換タスクにおいて高性能を達成できるということである。図 15 に示すように、トランسفォーマーは、左に配置されたエンコーダブロックと、右に配置されたデコーダブロックをそれぞれ複数回（原論文では 6 回）積み重ねて構成したエンコーダ・デコーダモデル（Encoder-Decoder Model）である。エンコーダブロックとデコーダブロックはどちらもマルチヘッド注意を主部品として使用するが、エンコーダブロックでは、一つのマルチヘッド注意と一つのフィードフォワードネットワーク（Feed-Forward Network）を組み合わせたという構造になっているのに対し、デコーダブロックではフィードフォワードネットワークの前に二つのマルチヘッド注意が配置される。デコーダブロック中の一つ目のマルチヘッド注意はマスクされたマルチヘッド注意である。二つ目のマルチヘッド注意は前のマルチヘッド注意の出力のみならずエンコーダからの出力も入力として受け取る。このようなマルチヘッド注意は相互注意（Cross Attention）とも呼ばれ、入力シーケンスと出力シーケンスという二つのシーケンスにおける各トークン間の依存関係を捉えることができる。それ以外の二つのマルチヘッド注意は、同じシーケンスのみを入力として受け取るので自己注意（Self-Attention）と呼ばれ、同じシーケンス内の各トークン間の依存関係を捉えることができる。

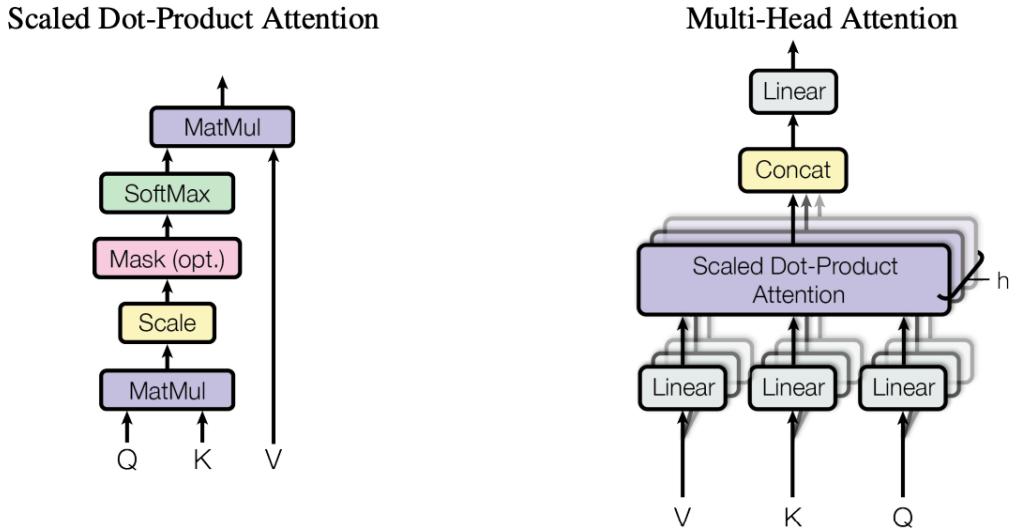


図 16 マルチヘッド注意の構造[25]

マルチヘッド注意の内部構造は図 16 に示す。マルチヘッド注意は入力として  $Q$ 、 $K$ 、 $V$  の三つの行列を受け取り、それぞれクエリ（Query）、キー（Key）とバリュー（Value）を意味する。この概念は key-value 型のデータベースに由来する。このことから、マルチヘッド注意はクエリ行列とキー行列の照合しながら、両者の関連度（= 注意）

に基づいてバリュー行列から必要な情報を取得するという操作を行っていることがわかる。 $Q$ 、 $K$ 、 $V$  の三つの入力行列が全部同一であり、すなわち同じ文が入力された場合に計算した関連度の例を図 17 に示す。この例では、文に含まれる異なる単語 (= トークン) 間のアテンションが計算される。

	I	<b>have</b>	<b>cats</b>
I	0.88	0.10	0.02
<b>have</b>	0.08	0.80	0.12
<b>cats</b>	0.03	0.14	0.83

図 17 関連度の例 [26]

三つの入力行列が全部同一のものである時、マルチヘッド注意は自己対自己の関連度に基づいて自分の中から必要な情報を取り出すことになる。入力が互いに異なるものである時、マルチヘッド注意は異なるクエリ行列とキー行列の関連度に基づいてキー行列から必要な情報を取り出すことになる。このマルチヘッド注意の計算は具体的に以下のように定義される。

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (28)$$

式 (28) に従い図 16 の中のスケール化内積注意 (Scaled Dot-Product Attention) を計算することができる。ここで  $d_k$  は行列  $Q$  の次元数である。 $\text{softmax}$  はソフトマックス活性化関数である [27]。マルチヘッド注意機構の最終出力は以下の式で与えられる。

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (29)$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^o \quad (30)$$

ここで、 $W$  はパラメータ行列であり、入力の  $Q$ 、 $K$ 、 $V$  それぞれにパラメータ行列をかけることで各ヘッド  $\text{head}_i$  の持つ注意が計算される。各ヘッドの注意を連結させたものにパラメータ  $W^o$  をかけた結果はマルチヘッド注意機構の最終出力である。

エンコーダブロックとデコーダブロックを構成するもう一つの部品は、フィードフォワードネットワークである。フィードフォワードネットワークは二つの線形変換と一つの活性化関数から構成される。その処理は具体的に次式のように定義される。

$$FFN(x) = \max(0, xW_1 + b_1) W_2 + b_2 \quad (31)$$

トランスフォーマーは最初、長・短期記憶をベースにした伝統的な系列変換モデルを差し替えるものとして機械翻訳向けに提案された深層学習モデルだが、その後の研究で文章分類や類似度評価など、多くの自然言語処理タスクにおいて優れた性能が確認された[28]。トランスフォーマーが提案されて以来、このアーキテクチャを基に自然言語処理に特化した深層学習モデルが多数考案されている。トランスフォーマーの構造に基づいて開発された言語モデルはトランスフォーマー系モデルと呼ばれる。トランスフォーマー系モデルは大きく自己回帰型モデル (Autoregressive Model)、オートエンコーディングモデル (Autoencoding Model)、系列変換モデル (Sequence-to-Sequence Model) の三種に分かれる。2.4.3、2.4.4、2.4.5 の各節では、それぞれのカテゴリーに属する代表的なモデルについて記述する。

#### 2.4.3 GPT

トランスフォーマーの構造にベースした自己回帰型モデルの代表的なものとして、2018年にOpenAIによって提案されたGPT (Generative Pre-Training / Generative Pre-trained Transformer) [29]というモデルがある。図18に示すように、GPTはトランスフォーマーのデコーダ部分のみ使用し、12層のデコーダブロックを積み重ねて構築されたモデルである。図18からわかるようにGPTの構造はトランスフォーマーとかなり似ているが、GPTの使用したデコーダブロックは各デコーダブロックにマルチヘッド注意機構が一つしか配置されていないという点でオリジナルのトランスフォーマーのデコーダと微妙に異なる。

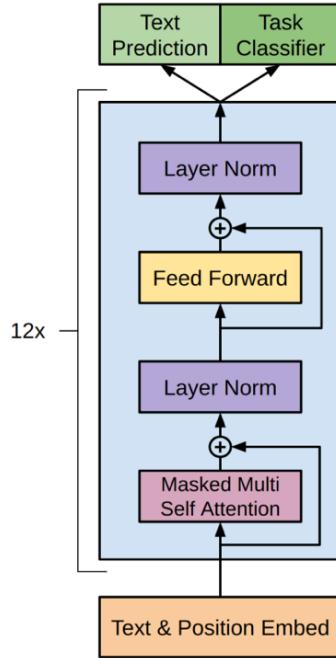


図 18 GPT のモデル構造[29]

GPT とトランスフォーマーとの最も大きな差異は、モデルの学習方式である。トランスフォーマーはこれまでの深層学習モデルと同じで大規模テキストで学習された単語埋め込みを使用し特定のタスクを学習するという方式を採用した。それに対し、GPT の学習手続きは教師なし事前学習（Unsupervised Pre-training）と教師ありファインチューニング（Supervised Fine-tuning）の二段階に分かれる。

教師なし事前学習とは、大量のラベルなしのコーパスを使用し、深層学習モデルに汎用的な言語ルールを学習させることであり、自己教師あり事前学習（Self-supervised Pre-training）と呼ばれることもある。この学習は非常に時間がかかるが、一回学習すれば学習済みのモデルをファインチューニングで様々な下流タスク（Downstream Task）に適用することができる。ファインチューニングとは、事前学習済みモデルをベースに、個別タスクに合わせて行う追加学習のことである。事前学習ではモデルは汎用的な言語ルールをしか学習しておらず、特定のタスクで高性能を出すにはこの第二段階の学習が必要である。ゼロから学習する場合と比べ、事前学習済みモデルは必要になるラベル付きの学習データと学習時間が格段に少なくなる。

GPT の教師なし事前学習は、標準的な言語モデル[30]タスクを事前学習タスクとして使用し、以下の対数尤度を最大化させることを目的とする。

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \dots, u_{i-1}; \Theta) \quad (31)$$

ここで  $\mathcal{U} = \{u_1, \dots, u_n\}$  は与えられたラベルなしのコーパスであり、その中に個々のトークンが入っている。 $k$  は次の単語の出現確率を計算する際に前の単語をどれだけ考慮するかというコンテキストウィンドウ (Context Window) である。 $\Theta$  はニューラルネットワークのパラメータを表す。教師なし事前学習において、GPT モデルの処理プロセスは以下のように定義される。

$$h_0 = UW_e + W_p \quad (32)$$

$$h_l = DecoderBlock(h_{l-1}) \forall i \in [1, n] \quad (33)$$

$$P(u) = softmax(h_n W_e^T) \quad (34)$$

ここで  $h_l$  は、各デコーダブロックの出力である。 $U$  は入力されたトークンからなるベクトルであり、 $W_e$  は単語埋め込み行列であり、 $W_p$  は語順に関する情報をトークンに付与する位置埋め込み行列である。モデルは複数個のトークンからなる入力シーケンスを 12 層のデコーダブロックを通して処理し、次に出現する単語の確率分布を予測する。教師なし事前学習が終わったモデルは学習済みモデル (Pre-trained Model) と呼ばれ、次の教師ありファインチューニングの段階に入る。教師ありファインチューニングにおいて、GPT モデルの挙動は以下のように定義される。

$$P(y|x^1, \dots, x^m) = softmax(h_l^m W_y) \quad (35)$$

ここで  $x^i$  は与えられた学習データセット  $\mathcal{C}$  のに含まる各シーケンスを構成するトークンである。 $y$  は各入力シーケンスのラベルである。 $h_l^m$  は最後のデコーダブロックの出力である。この出力に対して線形変換を行って入力シーケンスに対するラベルが予測される。ファインチューニングの目標は、下記の対数尤度を最大化することである。

$$L_2(\mathcal{C}) = \sum_{(x,y)} log P(y|x^1, \dots, x^m) \quad (36)$$

上記のように教師なし事前学習と教師ありファインチューニングの二段階で学習される GPT モデルは、事前学習済みモデルを特定のタスクに合わせて構造を改造し、さらにそのタスクの学習データで追加学習することによって様々な自然言語処理タスクに適用できる。図 19 に GPT モデルを下流タスクに適用する際に、下流タスクに合わせるためのモデル改造を示す。図 19 の中の「Transformer」は事前学習済みの GPT モデルを指し、学習済み GPT の後ろに線形層 (=多層ペーセptron) を追加で配置することで図に示した全ての下流タスクに適用することができる。

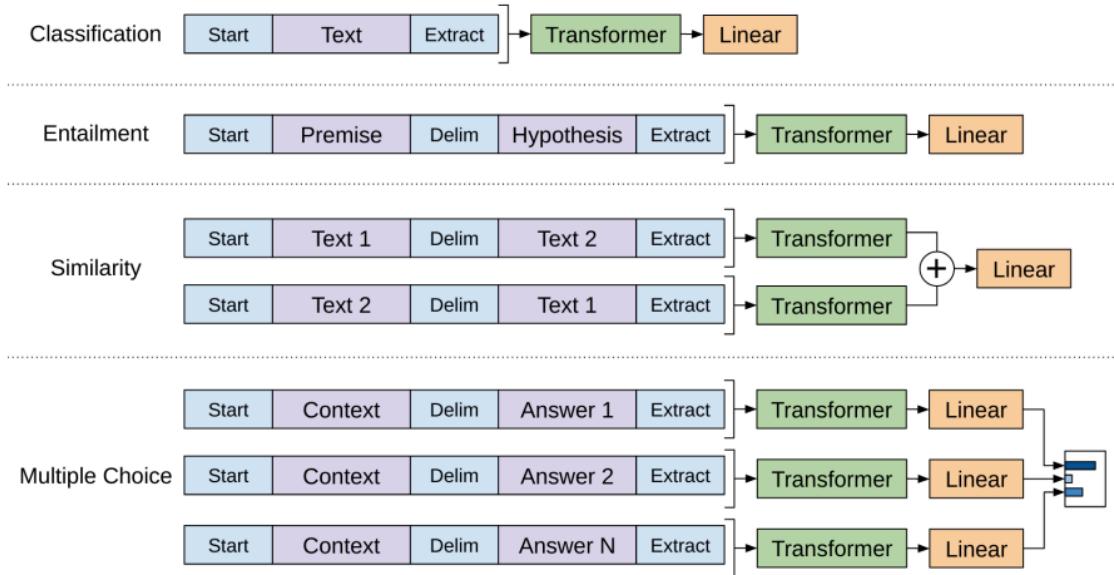


図 19 GPT の扱える下流タスク例[29]

しかし、多数の自然言語処理に関連するタスクで SOTA (State-of-the-art) 性能を叩き出した GPT モデルにも問題点がある。GPT は前出した単語から次に出現する単語を予測するという言語モデルタスクを事前学習タスクとしているので、文脈の後ろが参照できることになる。この問題点は、GPT の文脈全体を利用して意味を理解する能力にマイナスの影響を与える可能性がある。

#### 2.4.4 BERT

トランスフォーマーの構造に基づいて開発されたオートエンコーディングモデルとして最も代表的なのは、2018 年に提案された BERT (Bidirectional Encoder Representations from Transformers) [31] というモデルである。BERT はトランスフォーマーのエンコーダ部分のみによって構築されたモデルであるという点でトランスフォーマーのデコーダ部分のみ使用する GPT とは対照的であるが、GPT と同じように学習手続きを教師なし事前学習と教師ありファインチューニングの二段階に分けることによって精度向上を図った。

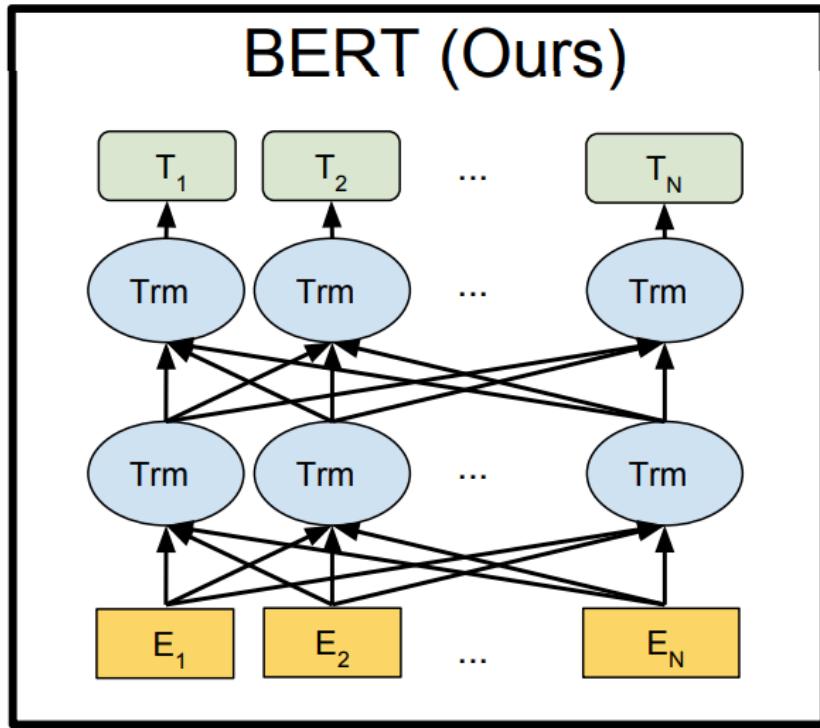


図 20 BERT のモデル構造[31]

図 20 に示すように、BERT は複数個のトランスフォーマーのエンコーダブロックを積み重ねてできたモデルである。BERT のエンコーダブロックは一つのマルチヘッド注意機構と一つのフィードフォワードネットワークから構成され、オリジナルのトランスフォーマーのエンコーダブロックと同じ構造を採用する。BERT は base モデルと large モデルの二種類のモデルが用意され、両者の主な差異はモデルの大きさにある。base モデルは 12 層のエンコーダブロックを持ち、隠れ層の次元数が 768 でマルチヘッド注意機構のヘッド数が 12 であるのに対し、large モデルではエンコーダブロックの層数が 2 倍の 24 層に拡張され、隠れ層の次元数とマルチヘッド注意機構のヘッド数もそれぞれ 1024 と 16 に増える。総パラメータ数で見ると base モデルは約 1 億個のパラメータがあり、large モデルは約 3 億個のパラメータがある。base モデルと比べて large モデルはかなり規模が増加したとともに多くのタスクでより高い性能を示した。

図 21 に示すように BERT の教師なし事前学習は具体的にマスク言語モデル (Masked Language Model) と次文予測 (Next Sentence Prediction) の二つの事前学習タスクが使用される。マスク言語モデルは、ランダムに選んだ単語を「MASK」というトークンに置き換え、周りの単語からこの単語を推測する問題、いわゆる「穴埋め問題」であり、この問題を解くことによってモデルが文法的・意味的知識を学習することができる。次文予測は、渡された二つの文が連続した文であるかどうかを判定する問題である。この問題を解くことによってモデルが二つの文の意味と二文間の関係を理解する必要があり、モデルの

論理力を鍛えることができる。この二つの事前学習タスクは GPT の場合と異なり双方向の事前学習であり、これによりモデルが前後の文脈を考慮することができるようになり、テキスト分類や質疑解答といった文章理解が重要なタスクにおいて GPT よりも高い性能を示した。一般にウィキペディアのような大規模なコーパスを使ってこの二つのタスクでモデルを事前学習し、データ量が膨大であるので事前学習は数十日程度かかる。

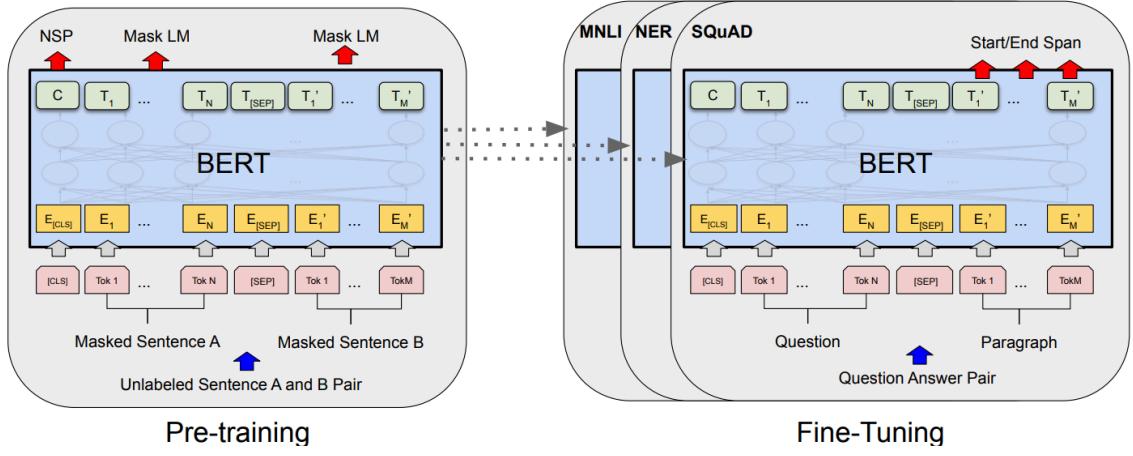


図 21 BERT の事前学習とファインチューニング[31]

BERT も二段階の学習手続きを採用したモデルであり、下流タスクに適用する際はタスクの学習データでモデルをファインチューニングする必要がある。事前学習済みの BERT モデルのファインチューニングは、図 21 に示すようにモデルの構造を大きく改造する必要がなく比較的に容易にできる。BERT の登場により、深層学習モデルは画像認識分野のみならず、自然言語処理分野においても人間を超えるような性能を達成できるようになる。

#### 2.4.5 BART

GPT のような自己回帰型モデルに前文しか参照できず後ろが参照できないという問題点がある一方、BERT を代表とするオートエンコーディングモデルは言語生成に関連するタスクに不向きであるという欠点が存在する。この二つの問題点を同時に解決するために、トランسفォーマーをベースに新しい世代の系列変換モデルが多数提案されており、その中の代表的なものとして、2019 年に Facebook によって提案された BART (Bidirectional and Auto-Regressive Transformers) [32] という深層学習モデルが挙げられる。トランسفォーマー構造の中の一部だけ採用する自己回帰型モデルとオートエンコーディングモデルとは対照的で、BART では図 15 に示す標準的なトランسفォーマーのエンコーダ・デコーダ構造に変更を加えずそのまま採用する。このような構造により、図 22 に示すように BART は双方向に文脈を考慮することができるのみならず、機械翻訳や文章要約といった生成式タスクにも対応可能になる。

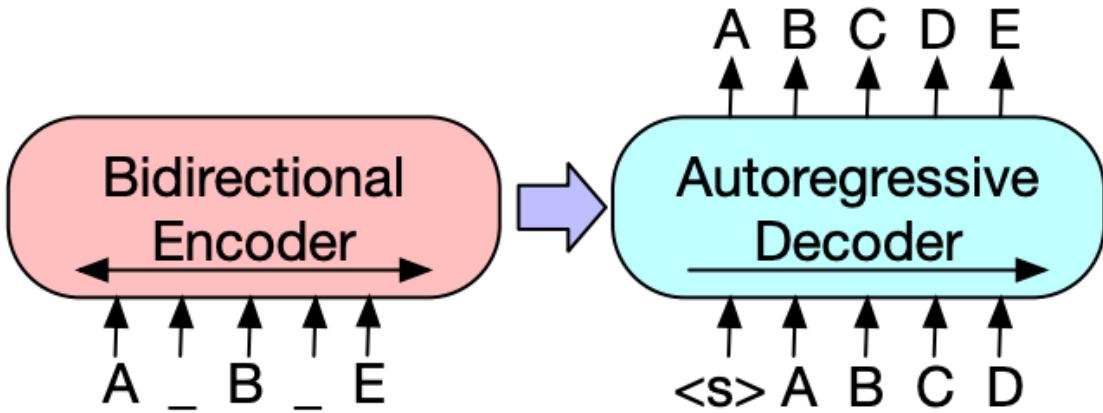


図 22 BART のモデル構造[32]

図 22 は BART の構造を示すものである。トランスフォーマーと同様に BART もエンコーダとデコーダという二つの部分の連結から構成される。エンコーダは、BERT と似ていて複数個のエンコーダブロックの積み重なりであり、入力シーケンス全体を考慮して各トークンの分散表現を生成しデコーダへ渡す。デコーダは、GPT に似た自己回帰型のデコーダブロックを複数個積み重ねた構造となっており、エンコーダから渡された入力に基づいて出力シーケンスを 1 トークンずつ生成する。しかし、エンコーダとデコーダの中の活性化関数を元々の ReLU 関数ではなく GELU 関数 (Gaussian Error Linear Unit) [33] 変更したという点で BART とオリジナルのトランスフォーマーと微妙に異なる。前述した BERT と同じように、BART も base モデルと large モデルの二種類のモデルがあり、base モデルではエンコーダとデコーダはそれぞれ 6 層のブロックで構成され、large モデルではエンコーダとデコーダはそれぞれ 12 層のブロックで構成される。BART の base モデルでは約 1.4 億のパラメータがあり、large モデルでは約 4 億のパラメータがある。

BART も事前学習とファインチューニングからなる二段階の学習方法を採用する。BART の事前学習タスクは、文章理解と生成能力を同時に向上させることを目的とし、その核心は様々な方法で入力となるテキストにノイズを導入し、ノイズを除去し元のクリーンなテキストを復元するようにモデルを学習することにある。BART の採用した事前学習タスクは図 23 に示すように、「Token Masking」、「Sentence Permutation」、「Document Rotation」、「Token Deletion」、「Text Infilling」の 5 種類がある。

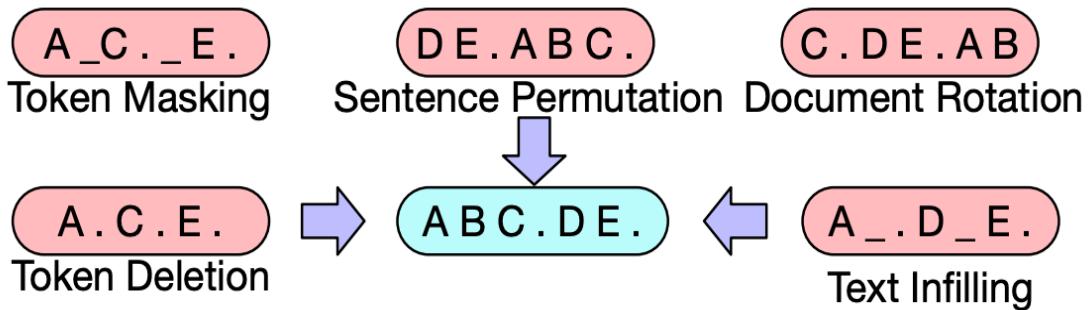


図 23 BART の事前学習タスク [32]

「Token Masking」は、BERT で使用したタスクと同じでランダムに単語を「MASK」トークンで置き換える部分を予測するという穴埋め問題である。「Token Deletion」は、ランダムに単語を削除し、その単語を埋めた文章を生成するというタスクであり、どの単語が削除されているかモデルがわからないので「Token Masking」と比べて難易度が上がると考えられる。「Sentence Permutation」は、複数の文の順番をシャッフルしてモデルに正しい順番を予測させるタスクである。「Document Rotation」は、文から単語一つ選び、その単語が最初に位置するように文を回転させ、モデルにどの単語が最初の単語かを学習させる。「Text Infilling」は、「Token Masking」と同じで穴埋め問題であり、ランダムに複数の単語の並びを一つの「MASK」トークンで置き換え、元の単語の並びを予測するというタスクである。大量の実験で検証したところ、他のタスクと比べて「Text Infilling」という事前学習タスクが多くのタスクにおいてより良い性能を達成できることが確認される。

BART のファインチューニングは、オリジナルのトランスフォーマーと同じように比較的に簡単に行われる。図 24 の左に示すように、テキスト分類タスクに対して、対象となるテキストをエンコーダとデコーダの両方に入力し、デコーダの最終層の最後の出力トークンをモデルによる予測として使用してモデルを学習する。生成式タスク向けのファインチューニングでは、大半のタスクは元のテキストをエンコーダに入力し、デコーダからの出力を結果として使用しモデルを学習するという方式を採用するが、機械翻訳タスクでは単純に外国語をエンコーダの入力とするのではなく、図 24 の右に示すように事前学習済みエンコーダの埋め込み層をもう一つのランダムに初期化されるエンコーダに変える必要があり、このエンコーダの役割は外国語を英語にマッピングすることである。

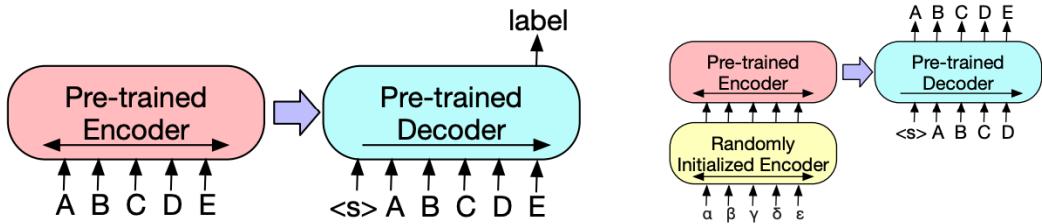


図 24 BART のファインチューニング方法[32]

BERT と GPT の二つのモデルの強さを融合したものとして、BART はテキスト理解がメインであるタスクにおいて BERT に匹敵するほどの性能を示しながら、生成式タスクにおいて GPT を上回る性能を達成している。

#### 2.4.6 大規模言語モデル

前節まで説明したように、トランスフォーマーの構造をベースにした深層学習モデルは自然言語処理分野において著しい進歩を遂げている。2020 年に深層学習言語モデルに関する Scaling Law が OpenAI によって提案され、トランスフォーマーの性能はパラメータ数、データセットサイズ、計算予算の三つを変数としたべき乗則に従うことが証明された[34]。これを受けた近年、トランスフォーマーのデコーダ構造を大量に積み重ねることによって深層学習モデルのパラメータ数を拡張し、さらに大量のデータを学習に使用してモデル自身の表現能力を強化する傾向が見られる。中でも特にパラメータ数が多いモデルは、大規模言語モデル（large Language Model）と呼ばれる。大規模言語モデルに厳密な定義がなされていないが、数十億以上のパラメータを持つ自然言語処理に特化した生成式深層学習モデルは一般的に大規模言語モデルとして認知される。大規模言語モデルは、ウェブページや書籍、論文に含まれる大量のデータを学習することによって様々な自然言語処理タスクを汎用的に解くことを目的とし、特定タスクのデータに依存せずに各種タスクの精度改善を見せ続けている。代表的大規模言語モデルとして、OpenAI が開発した GPT シリーズや Meta によって公開された LLaMA シリーズ、Google が開発した LaMDA などもある。本節では主に、本研究で使用する GPT シリーズについて記述する。

OpenAI によって開発された一連の大規模言語モデルである GPT シリーズに、GPT-3 や InstructGPT、GPT-3.5、GPT-4 といったモデルが含まれる。GPT-3[35]は 2020 年に提案された大規模言語モデルであり、2.4.3 節で説明した GPT の基本的な構造を踏襲しデコーダブロックの数を大幅に拡張した構造を持つ。当時の深層学習言語モデルの平均的なサイズを十数倍以上上回った約 1750 億のパラメータを持つ GPT-3 は、学習時に約 45 テラバイトのデータが使用され、これによりファインチューニングせずに少量の学習データさえ参照すれば特定の自然言語処理タスクを実行できるというフューショット学習（Few-shot Learning）を初めて実現した。

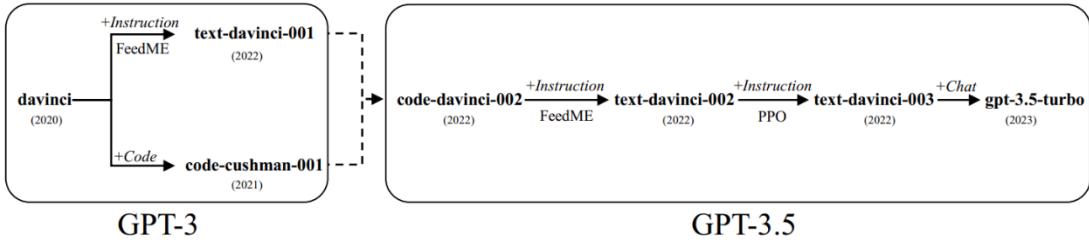


図 25 GPT-3 から GPT-3.5 への進化過程[36]

GPT-3 の後継モデルは GPT-3.5 である。厳密な定義は存在しないが、GPT-3.5 は一般的に 2022 年に OpenAI によって開発された「code-davinci-002」をベースモデルとする一連の大規模言語モデルを指し、モデルのサイズがさらに拡充され GPT-3 よりも多くのパラメータを持つとされている。GPT-3 から GPT-3.5 への進化の過程は図 25 に示す。図 25 にあらわすように GPT-3.5 の学習に、人間が作成した高品質なテキストサンプルに基づいてモデルのファインチューニングを行うという「FeedME」と強化学習アルゴリズムである「PPO」という二つの学習ストラテジーが使用された。GPT-3 に比べて GPT-3.5 に属するモデルはコード生成や対話といったタスクにおいて性能が強化され、のち大ヒットした対話プロダクトである ChatGPT のベースともなった。

Benchmark	GPT-4	GPT-3.5	Contamination	GPT-4 (non-contaminated)	Degradation
MMLU	86.4%	70.0%	~0.6%	-	-
GSM-8K	92.0%	57.1%	~1%	-	-
HellaSwag	95.3%	85.5%	~*	-	-
AI2	96.3%	85.2%	~3.4%	-	-
WinoGrande	87.5%	81.6%	~0.9%	-	-
HumanEval	67.0%	48.1%	25%	65.58%	-2.12%
DROP (F1)	80.9	64.1	~21%	82.8* (subsample)	0

図 26 GPT-4 の性能[37]

GPT-3.5 からさらに進歩したのは GPT-4 である。GPT-4[37]は 2023 年に OpenAI によって開発された最新世代の大規模言語モデルであり、テキストのみならず画像の入力と処理も可能になったマルチモーダルの大規模言語モデルでもある。GPT-4 はこれまでの GPT シリーズモデルと同じようにトランスフォーマーのデコーダ構造をベースとして採用しながら、大量のコーパスによる教師なし事前学習、対話タスク向けの教師あり学習と人間フィードバックによる強化学習 (Reinforcement Learning from Human Feedback) からなる独創的な三段階の学習手続きを採用したとされている。GPT-4 モデルの入力できるトークン

数も最大 32,768 トークンに届き、GPT-3.5 の約 4000 トークンから大幅に増加した。こういった工夫により、アメリカの司法試験や大学入試テストにおいて優秀な成績を達成し、図 26 にあるように高度な自然言語処理能力を必要とする複数のベンチマークタスクにおいて極めて高い性能を示した。しかし、その先行モデルである GPT-3 や GPT-3.5 と同様に、GPT-4 はオープンソースプラットフォームとして公開されておらず、GPT-4 の利用は OpenAI が提供する API を介してのみ可能であり、直接的なアクセスやカスタマイズの選択肢が提供されていない。

上記で述べたように、大規模言語モデルは自然言語処理分野において多大な進歩をもたらしているが、と同時に重要な問題点も存在する。まず挙げられるのがハルシネーション (hallucination) という問題である。ハルシネーションはモデルが実際に存在しない情報や文脈と矛盾した内容を生成する現象を指す。また、大規模言語モデルは特定のドメインに関する深い知識や専門的な理解を欠くことがある。例えば、医学や法律などの専門的な分野に関して、適切な学習データが不足しているため正確な情報提供が難しい場合がある。さらに、大規模言語モデルの訓練には膨大な計算リソースと時間が必要であり、これが学習コストの増大につながる。これらの問題に対処するためには、より効率的な学習手法の開発やドメイン固有の知識を統合するアプローチの探求が求められる。

## 2.5 知識グラフ構築への深層学習技術の応用と課題

2.4 節では、自然言語処理分野における深層学習技術および近年までの進展について説明した。知識グラフの構築も深層学習技術なしでは語れないのが現状である。本節では、深層学習技術を知識グラフ構築に応用した先行研究を概観し、深層学習技術を用いた肝細胞がん診療ガイドラインに関する知識グラフの構築に存在しうる課題について検討する。

### 2.5.1 知識グラフ構築への深層学習技術の応用

2.1 節に説明したように、知識グラフの基本的な構成単位は二つの実体とこの二つの実体間の関係からなる関係トリプル (Relation Triplet) であり、非構造化データから知識グラフを構築する上で中核的なタスクはこの関係トリプルの作成である。非構造化データからの関係トリプルの作成に関して様々な応用研究が行われている。関係トリプルの作成は大きく固有表現認識と関係抽出の二つのタスクに分けられる[38]。

固有表現認識とは、テキストから固有名詞や日付、時間表現などの固有表現を自動的に抽出し、あらかじめ定義されたクラスに分類するタスクである。固有表現認識は一般的に、系列データ中の各トークンに対してラベルを付与する系列ラベリング (Sequence Labeling) のタスクとして扱われ、図 27 にそのイメージを示す。

## Albert is going to United States of America.



図 27 固有表現認識の例[39]

固有表現認識では、単語の系列に対して BIO (Begin、Inside、Outside) 形式で固有表現のラベルが付与され、「B」が固有表現の開始位置であること、「I」が固有表現に含まれること、「O」が固有表現に含まれないことを表す。B と I をラベルとして使用する際は、該当するタイプ名を連結して「B-LOC」や「I-LOC」のような形式にする。例えば、「United States」という表現に「B-LOC」と「I-LOC」の二つのラベルが付与され、「Albert」のような短い固有表現に「B-PER」の一つのラベルだけ付与される。系列の中の各トークンにどのラベルを付与するかを判定することがタスクの目標となる。

固有表現認識は主に、ルールベースの認識、統計モデルベースの認識、深層学習技術による認識の三種類の方法が存在する[40]。

ルールベースの認識とは、テキストから固有表現の出現パターンを探りテンプレートを作成し、テキストに対して作成したテンプレートを用いてパターンマッチングをしながら固有表現を認識する方法である。ルールベースの手法を用いた先行研究として、Riaz らはルールベースのウルドゥー語の固有表現認識のアルゴリズムを開発し、統計的機械学習に基づいたモデルを上回る精度を達成した[41]。日本語の固有表現認識に関する先行研究として、Iwakura らはラベルなしのデータから収集した固有表現の出現パターンを使用した固有認識方法を開発した[42]。統計モデルベースの認識とは、ある程度ラベリングされたデータを用いて統計モデルを学習し、ラベルなしのデータに適用して固有表現を認識する方法である。統計モデルベースの手法を用いた先行研究として、Finkel らは隠れマルコフモデル (Hidden Markov Model) とモンテカルロ法 (Monte Carlo Method) に基づいた固有表現認識システムを提案し[43]、Liao らは条件付き確率場 (Conditional Random Field) に基づいた半教師あり学習アルゴリズムをトルコ語の固有表現認識に適用した[44]。しかし、いずれの研究も今から 10 年以上経過しており、ルールベースと統計モデルベースの認識方法はすでに固有表現認識のメインストリームではないと言える。

近年、自然言語処理分野において深層学習技術が大きく進展し、深層学習技術に基づいた固有表現認識の方法も多数提案されている。深層学習技術による認識は主に大量の BIO 形式でラベリングされたテキストを用いて深層学習モデルを学習し、学習済みモデルを使用して固有表現認識を行うという教師あり学習の方式を採用する。固有表現認識へ深層学習技術を応用する例として、Lample らによって提案された深層学習技術と伝統的な条件付き確率場を融合した LSTM-CRF モデル[45]や、Ma らによって提案された LSTM-CNNs-CRF モ

デル[46]がある。これらの方は、単語と単語の文法的関係のみならず文脈を捉える能力も持つため、従来のルールベースや統計モデルベースの手法よりも高い性能を示した。

2019年に自然言語理解に関連するタスクにおいて高い性能を示したBERTモデルが提案されて以来、深層学習技術による固有表現認識手法にBERTを取り入れる応用研究が行われている。SouzaらはBERTとCRFを結合させたBERT-CRF構造を用いてポルトガル語の固有表現認識ベンチマークにおいて優れた性能を達成し、データが少ない言語でもBERTのファインチューニングが有効であることを証明した[47]。日本語の固有表現認識にBERTを応用した先行研究として、柴田らはBERTに基づいたJoint NER-REモデルを用いて固有表現認識による診療テキストからの情報抽出を成功させた[48]。

一方、関係トリプル抽出のもう一つのサブタスクである関係抽出もほぼ同様の発展を辿ってきた。関係抽出とは、テキストの中から与えられた複数の実体間の関係を抽出するというタスクである。関係抽出は固有表現認識とも密接に関わり、一般的に固有表現認識の次のステージとして位置付けられている。深層学習技術による関係抽出手法が発達する前に、テンプレートを用いたルールベースの関係抽出手法が採用されていた。例として、坂地らによって提案された手がかり表現を用いて因果関係を表す表現を抽出する手法[49]や、佐藤らによって提案された格フレームを用いる手法[50]がある。近年、関係抽出の主流的な手法は深層学習技術による抽出手法に移行し、Zhouらによって提案されたAtt-BLSTM[51]やJiらによって提案されたAPCNNs[52]など、関係抽出向けの深層学習モデルが多数開発されており、従来の手法を凌ぐ性能を見せていている。日本語の関係抽出タスクに深層学習技術を取り入れる応用研究もある。Sugimotoらは、画像診断レポートからの情報抽出を目的とし、BiLSTM-CRFモデルとBERTを融合した二段階の深層学習によるアプローチを提案し、クリニカル情報の抽出に成功した[53]。

固有表現認識と関係抽出を組み合わせて一つのタスクとして扱い、与えられたテキストから関係トリプルを直接に抽出する応用研究が近年行われている。代表的な例として、Cabotらによって提案されたREBEL (Relation Extraction By End-to-end Language generation) [54]というエンドツーエンドの抽出手法が挙げられる。REBELは2.4.5節で説明したBARTという系列変換モデルに基づいたモデルであり、関係トリプル抽出を言語生成式タスクとして扱うことにより英語文書の関係トリプル抽出タスクにおいて高い性能を示した。2023年にその後継モデルであるmREBEL[55]が初めてのエンドツーエンド型多言語関係抽出モデルとして提案され、複数の言語の関係トリプル抽出タスクに対応できるようになる。

### 2.5.2 知識グラフ構築に存在する課題

肝細胞がん診療ガイドラインに関する知識グラフを構築する上で存在する課題として、まず挙げられるのがタスクの複雑性である。この複雑性は、関係トリプルの抽出が固有表現認識と関係抽出という二つのサブタスクから構成されることに由来するのみならず、対象となるテキストが日本語で書かれているということもタスクの複雑性を増加させる。深

層学習技術による固有表現認識・関係抽出の方法は多くは英語を対象に開発されており、日本語に適用すると親和性が低いことがある。例えば、REBEL の多言語バージョンである mREBEL モデルを SRED<sup>FM</sup> という英語以外の言語がメインなデータセットに適用したところ精度が下がることが確認された[55]。また、日本語に分かち書きされていないことや語順が自由で構文規則が難しいこと、文節の省略が多いことといった英語にはない言語学的な現象が存在するので言語処理の複雑性が増加し、深層学習モデルを用いて日本語テキストを処理する際に高度な自然言語理解能力が求められる。

第二の課題として、深層学習技術はデータ・ドリブンの技術であり、関連先行研究でも示したように固有表現認識と関係抽出の両タスクを深層学習のアプローチで行うには、大量のデータを用いた教師あり学習を必要とする。それに、本研究では構築の対象となるのは医学的知識グラフであり、医学的ガイドラインから関係トリプルを抽出することが必要である。そのため、汎用的な日本語固有表現認識・関係抽出データセットが存在するにもかかわらず本研究に不向きであり、医学分野で質の高いアノテーション付きデータの確保が求められる。しかし、日本語のアノテーション付き医療言語処理データセットで一般向けに公開されているものは少なく、深層学習モデルの学習に使用できる教師データが限られているのが現状である。アノテーション付きデータの不足を克服し、肝細胞がん診療ガイドラインのテキストから関係トリプルを作成することも、肝細胞がん診療ガイドラインに関する知識グラフを構築する上で一つの課題となる。

## 2.6 質疑応答システム構築に関する先行研究と課題

2.2 節に述べたように、深層学習技術を活用した質疑応答システムが大きな成功を収めている。本節では、深層学習技術を質疑応答システムの構築に活用した応用研究を通観し、質疑応答システムを構築する際の課題について説明する。

### 2.6.1 質疑応答システム構築への深層学習技術の応用

質疑応答システムは大まかに抽出式質疑応答システム (Extractive-based Question Answering System) と生成式質疑応答システム (Generative-based Question Answering System) の二種類に分類される[11]。抽出式質疑応答システムはデータベースや文書から特定のテキストの範囲や幾つかの単語をそのまま抽出したものを回答として返す質疑応答システムであり、このようなシステムにオートエンコーディング系の深層学習モデルが使用されることが多い。深層学習技術を抽出式質疑応答システムに用いた先行研究として、Alloatti らはテキスト分類に特化した BERT モデルを用いて与えられたデジタル請求書や帳票に関する質問的回答するイタリア語の質疑応答システムを開発した[56]。Alzubi らは、DistilBERT[57]を活用して選ばれた文書から短い回答を抽出し、新型コロナウイルス (COVID-19) に関する情報を提供する質疑応答システムを構築した[58]。

一方で生成式質疑応答システムはテキストの抽出ではなくテキストの生成を通して回答を作成するので、より多様性のある回答を提供できる。このようなシステムはオートエンコーディング系モデル以外に自己回帰型深層学習モデルや系列変換モデルも使用される。例えば、Izacard らは文章抽出を扱う BERT と回答の生成を扱う系列変換モデルを組み合わせた質疑応答システムを開発し、生成式系列変換モデルに入力する文章数の増加につれ質疑応答システム性能が向上することを証明した[59]。Zheng らは系列変換モデル、注意機構と敵対的生成ネットワークを用いて回答を生成する質疑回答システム NLQA を提案した[60]。近年大ヒットしている大規模言語モデルも生成式質疑応答システムの一種として数えられ、大規模言語モデルを駆使して地理情報に関する質疑回答システムを構築した先行研究が存在する[61]。

### 2.6.2 質疑応答システム構築における課題

肝細胞がん診療ガイドラインに関する質疑応答システムを構築する際には下記のように幾つかの課題が存在する。まず専門性の付与という課題が挙げられる。本研究では肝細胞がん診療という分野に特化した質疑応答システムの構築を目標としており、汎用的なオープンドメイン質疑応答システムと比べ肝細胞がん診療における専門性が求められる。構築した肝細胞がん診療ガイドラインに関する知識グラフを質疑応答システムと組み合わせることにより質疑応答システムに専門知識を与えることが可能であるが、質疑応答システムの構築に専門性の高い知識グラフをうまく組み込むには適切な方法を要する。

質の高い回答を提供するためには、質疑応答システムは利用者からの質問の意図を正しく理解することが前提である。質疑応答システムの出力する回答も、利用者が理解できない不自然な回答であれば意味がない。そのため、構築した質疑応答システムは高度な専門性のみならず、十分な自然言語理解能力、特に日本語の文章を理解する能力と作成する能力を持つ必要がある。構築した質疑応答システムに基礎的な言語力が備わる必要があるという点も課題である。

最後に、構築した質疑応答システムをどのようなデータで評価するかという課題は無視できない。肝細胞がん診療に関する質問の回答に特化した質疑応答システムであるので、汎用的な日本語質疑回答データセットを用いて性能を評価するのは適切ではない。このようなドメイン特化型質疑応答システムを評価するには、その分野と関連性の高い評価データを使用する必要がある。

### 3 肝細胞がん診療ガイドラインに関する知識グラフの構築

肝細胞がん診療ガイドラインに含まれる知識の構造化および可視化を実現するために、本章では、知識グラフの構築に必要な関係トリプル抽出方法を提案し肝細胞がん診療ガイドラインに関する知識グラフを構築する。2.5.2で説明した知識グラフの構築にあたってのタスク複雑性とデータ不足という二つ課題を念頭に置いて対処しつつ、3.1節で日本語事前学習済み BART モデルを用いて REBEL を構築し、関係トリプル抽出を行いその性能を検証する。3.2節で複数の深層学習モデルを組み合わせた多段階関係トリプル抽出アルゴリズムを提案し知識グラフを構築する。3.2節で提案した手法を用いて構築した知識グラフの評価を3.3節で行う。

#### 3.1 REBEL による関係トリプル抽出の試み

##### 3.1.1 REBEL の概要

2.5節で説明したように、REBEL は系列変換モデルである BART をベースとして使用し、与えられたテキストから直接に関係トリプルを抽出するエンドツーエンド型の深層学習モデルである。図 28 に示すように REBEL は、関係トリプルの抽出を系列変換タスクとして扱い、入力されたテキストをシーケンスの形として表現された関係トリプルに変換することで関係トリプルの抽出を実現する。この系列変換タスクは次式のように定義される。

$$p_{REBEL}(y|x) = \prod_{i=1}^{len(y)} p_{BART}(y_i|y_{<i}, x) \quad (37)$$

ここで  $x$  は入力シーケンスを表し、 $y$  は  $x$  の中に含まれる実体と関係を線形化された形として表現する出力シーケンスを表す。REBEL の扱う系列変換タスクは式(37)のように  $x$ に基づいて  $y$  を自己回帰的に生成するタスクとして定義できる。しかし、関係トリプルは元々シーケンスではないので、それをシーケンスとして線形化する必要があり、REBEL では  $\langle \text{triplet} \rangle$ 、 $\langle \text{subj} \rangle$  と  $\langle \text{obj} \rangle$  の三つの特殊トークンを使用して関係トリプルを図 28 のようにシーケンスとして線形化する方法が提案されている。

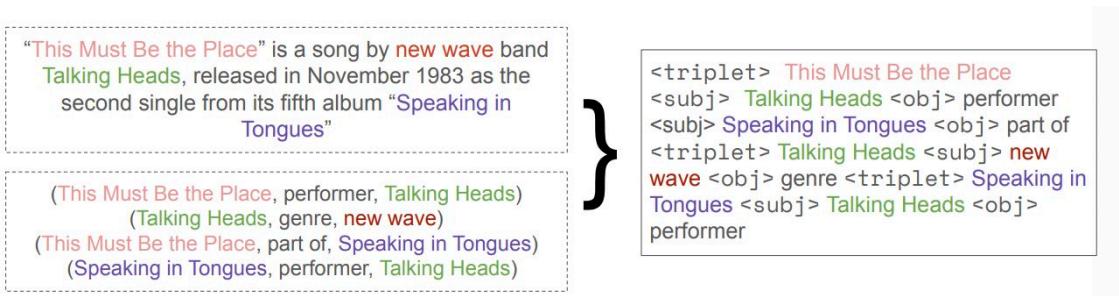


図 28 REBEL による関係トリプル抽出のイメージ[54]

REBEL はエンドツーエンドで関係トリプルを直接に抽出する能力があり手間を省くことができるが、しかし元々英語を対象として開発された手法であり、公開されている REBEL モデルも日本語に対応しているものはない。そのため、REBEL のように日本語テキストから関係トリプルを抽出するために、日本語の BART モデルと関係抽出データセットを用いて日本語に対応した REBEL を構築する必要がある。

### 3.1.2 REBEL の構築に用いたベースモデルとデータセット

日本語対応 REBEL のベースモデルとして、愛媛大学人工知能研究室が公開した日本語事前学習 BART モデルである AcademicBART[62]を使用する。この事前学習モデルを[53]にある学習方法に従い追加の関係抽出データセットを用いてファインチューニングすることで関係トリプル抽出というタスクに対応させ、REBEL を構築する。AcademicBART のファインチューニングに iCorpus[63]という関係抽出データセットを使用する。

iCorpus とは、東京大学大学院医学系研究科医療 AI・デジタルツイン開発学講座によって構築され一般向けに公開されている日本語の医療分野の関係抽出データセットである。医療分野の関係抽出に関するデータセットであるため、本研究との親和性が高いと考え REBEL の構築に使用した。このデータセットはテキストとテキストの中の疾患名や薬品名といった固有表現と固有表現の間の関係を表すアノテーションという二つの部分から構成される。テキストは、厚生労働省の指定難病を対象として JStage で検索して得られた結果から選んだ本文が公開されている 179 件の症例報告のテキストである。アノテーションは、七十種類の固有表現と三十五種類の関係について文字単位で作られている。図 29 に iCorpus データセットのデータ例を示す。下段はテキストであり、上段にアノテーションされた実体のカテゴリーと複数の実体間の関係が示されている。図 29 からわかるように、iCorpus では短いテキストでも実体が細かくアノテーションされており、実体間の関係も多数付与されている。また、iCorpus データセット中のアノテーションはシーケンスの形ではなく独立した文字列であるので、<triplet>、<subj> と <obj> の三つ特殊トークンを用いてアノテーション文字列を連結させ、REBEL が扱えるシーケンス形式に変換する必要がある。例えば、図 29 にあるシーケンスに付与されたアノテーションは、<

`triplet > 開始 < subj > SGLT2 < obj > executed < subj > 判断し < obj > reason < triplet >`  
`判断し < subj > 糖尿病 2 型 < obj > value-of` のようにシーケンスとして線形化できる。

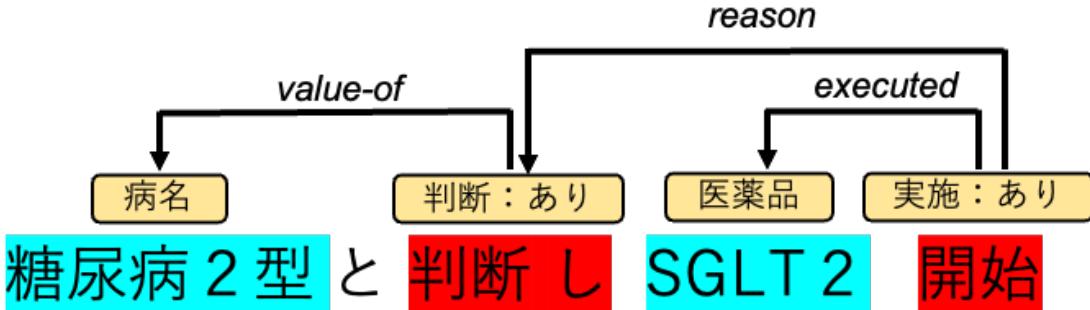


図 29 iCorpus のデータ例 [63]

### 3.1.3 REBEL による抽出性能の検証

iCorpus データセットに合計 2004 個のアノテーションされたシーケンスがある。これを 9 対 1 の割合で学習用データセットと検証用データセットに分割し、それぞれ 1803 と 201 個のシーケンスが入っている。この 1803 個の学習データを REBEL に入力し、各パラメータを調整することでファインチューニング学習を行い、201 個の検証データをモデルに入力しパラメータを調整せずに精度の評価を行う。

学習時のハイパープラメータは表 1 のように設定した。バッチサイズは学習時に一回入力するデータの総数を表し、12 個のシーケンスを同時に入力するように設定した。バッチごとに勾配が計算され誤差逆伝播法に従い REBEL の重みパラメータが更新され、1803 個の学習データすべてにパラメータの更新が行われる。全ての 1803 個の学習データに対して更新を行うことは 1 エポック (Epoch) と呼ばれ、このエポックを繰り返すことで誤差関数が小さくなるようにパラメータが更新される。学習時のエポック数を 100 エポックに設定した。各エポックにおける学習データの入力順番はランダムであり、エポックごとにシャッフルされる。また、REBEL の学習は以下の実験環境のもとで行われた。使用した計算機はオペレーティングシステムが Ubuntu 18.04 LTS であり、CPU として Core i9 10940X を搭載し、GPU として 24 ギガバイトの NVIDIA TITAN RTX を 2 枚搭載する。使用した Python バージョンは Python 3.10.13 であり、CUDA バージョンは CUDA 11.8 であり、深層学習フレームワークは PyTorch 2.0.1 である。モデル学習時の乱数シードは 42 に設定される。本研究における実験は、特に説明がない限り全てこの環境下で行われた。

表 1 REBEL の学習条件

学習パラメータ	数値
エポック数（ステップ数）	100 (15, 100)
バッチサイズ	12
バッチ数	151
学習率	5e-5
ラベル平滑化率	0.1
学習データ数	1,803

学習時の最適化アルゴリズムとして AdamW[64]を使用する。抽出精度の評価にあたりシーケンスの正解率という評価基準を利用し、モデルの出力したシーケンスが正解シーケンスと完全に一致する場合だけ正解とする。上記の学習条件のもと学習した REBEL に対して 201 個のシーケンスを検証データとしてモデルの関係トリプル抽出性能を評価した。その結果、100 エポック学習したにも関わらず REBEL の正解率は 0.1045 にとどまり、関係トリプルの抽出精度が極めて低いことがわかる。また、REBEL による抽出結果について表 2 に例を示す。表 2 からもわかるように「急性」や「高危険群 | が'度」のような入力文にない単語も実体として誤って抽出されており、REBEL による高精度な関係トリプル抽出は実現できなかった。

表 2 REBEL によるトリプル抽出例

入力文	肝細胞癌は高危険群の設定が容易な癌である		
Subject 実体	急性	癌	高危険群   が'度
Object 実体	癌	肝細胞	病変
関係	value_of	site	unit

REBEL によるエンドツーエンド型の関係トリプル抽出が失敗したことに学習データの少なさが主な原因として考えられる。与えられたテキストから生成的に関係トリプルを抽出するというタスクが複雑であり、医学的文書に存在する専門性も複雑性を増加させる。このような複雑性の高いタスクにモデルを適用させるには数千規模の学習データでは足りず、より多くのデータが必要となる。しかし、医学分野のアノテーション付き関係抽出データが大量に確保できないことが現状であり、REBEL による高精度な関係トリプル抽出の実現が難しいと考える。

## 3.2 多段階関係トリプル抽出アルゴリズムによる知識グラフ構築

前節では REBEL によるエンドツーエンドの関係トリプル抽出方法を提案し性能を検証したところ実用性がないことがわかる。その原因として学習データの少なさおよびタスクの複雑性が考えられる。こういった課題を解決するために、大量の学習データがなくても高性能を達成できるより強力な深層学習モデルを使用することや、関係トリプル抽出というタスクを複数のステップに分割することが有望な解決案として挙げられる。これを踏まえ、本節では、BERT や ELECTRA、大規模言語モデルである GPT-4 といった深層学習技術を組み合わせた多段階の関係トリプル抽出アルゴリズムを提案し知識グラフを構築する。

### 3.2.1 提案手法

本研究で提案する多段階関係トリプル抽出アルゴリズムは図 30 に示すように固有表現認識モデルの構築、固有表現認識によるキーコンセプトの抽出、構文解析による対象文の選別、大規模言語モデルによる属性と属性値の抽出、の 4 段階に分けることができる。まず、日本語の医療文書固有表現認識データセットを用いて複数の深層学習モデルをファインチューニングし固有表現認識モデルを構築する。次に、学習した固有表現認識に特化した深層学習モデルで最も性能の良い BERT-MLP モデルを採用し、肝細胞がん診療ガイドラインのテキストから重要な固有表現（キーコンセプト）を抽出する。その後、オートエンコーディング型深層学習モデル ELECTRA にベースした日本語自然言語処理ライブラリ GiNZA を使用して構文解析を行い、関係トリプル抽出の対象文を選別する。最後に最新世代の大規模言語モデルである GPT-4 を用いて関係トリプル抽出の対象となる文から、文に現れたキーコンセプトの持ちうる属性およびその属性値の抽出を行う。以下の各節でこのアルゴリズムの詳細について説明する。

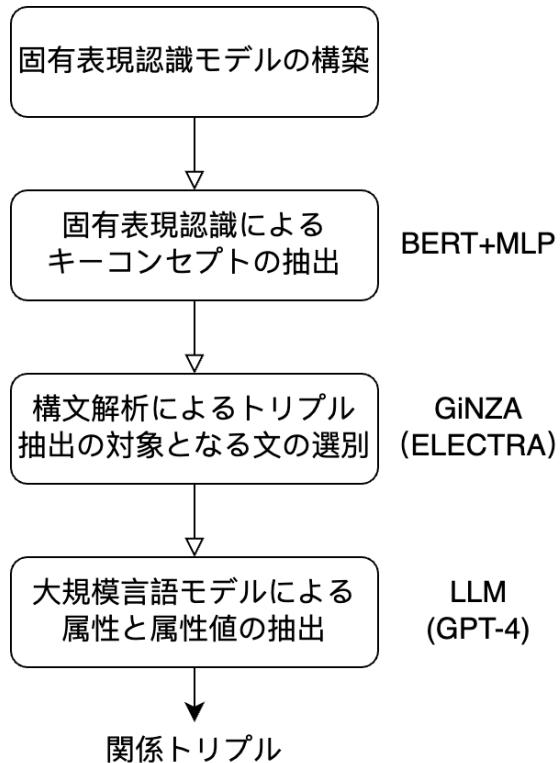


図 30 多段階関係トリプル抽出アルゴリズム

### 3.2.2 固有表現認識モデルの構築

関係トリプルに実体が不可欠であり多段階的に関係トリプル抽出を行うために実体を抽出する固有表現認識というタスクがなくてはならない。本研究で提案する多段階関係トリプル抽出アルゴリズムにおいても、固有表現認識モデルの構築が最初のステップとして位置付けられる。方法としては、日本語の医療分野固有表現認識データセットである MedTxt (Medical NLP Standard Text dataset) [65]を用いて大量の日本語コーパスで学習済みの自然言語処理向け深層学習モデルをファインチューニングすることにより医療分野に特化した固有表現認識モデルを構築する。

#### i ) データセット

モデルのファインチューニング学習に使用したデータセットについて説明する。MedTxt は奈良先端科学技術大学院大学ソーシャル・コンピューティング研究室によって構築され、一般向けに公開されている固有表現認識データセットである。MedTxt は MedTxt-CR と MedTxt-RR の二種類のコーパスから構成され、前者は J-Stage でオープンアクセス公開されている症例報告論文から抽出したテキストをベースに病名や部位名といった固有表現がアノテーションされているコーパスであり、後者は肺がん CT 画像を対象として複数名の読影医が作成した読影レポートをベースに固有表現がアノテーションされているコーパス

である。固有表現認識モデルの構築にあたり学習データをより多く確保することが重要であるため MedTxt の二種類のコーパスを全部使用することにする。

図 31 に MedTxt データセットの中の一文書をサンプル例として示す。図 31 に示すようにこの症例報告のテキストは多数の固有表現が複数のカテゴリーに分かれてアノテーションされており、テキストからラベリングされた個々の固有表現を正しく検出するのがモデルのタスクである。固有表現が分類されるカテゴリーとして、病名、部位名、特徴、変化、時間表現、検査名、検査項目、検査値、薬品名、薬品値、医療処置、クリニカルテキストの 12 クラスがある。データの前処理として、MedTxt のテキストを文字レベルで分割し、2.5.1 節で述べたようにそれぞれの文字に BIO 形式のラベルを付与し深層学習モデルの扱いやすい形に変換した。

症例は [時(AGE) 70 歳]、男性。

[回血疾] を主訴に当院紹介となり、[Tt(+)] C T・気管支鏡検査 の結果、[D(+)] 非小細胞肺癌 と診断された。

[時(DATE) その後]、[R(+)] 放射線化学療法 (S-1 + CDDP 療法) が導入され、[時(CC) 2 コース施行後] の [R(+)] 維持療法 として [R(+)] デュルバルマブ療法 (抗 PD-L1 抗体療法) を行ったところ、  
[D(+)] 発熱と炎症反応上昇 を認めたため [Mk(+)] 抗 PD-L1 抗体 [時(CC) 投与 5 日後] から [Mk(+)] LVFX 投与を開始した。

[Mk(+)] LVFX [時(CC) 投与 3 口目] の [Tt(+)] 採血 で [D(+)] 肝機能障害 と [D(+)] 炎症反応 の [回更なる上昇] を認めたため [Tt(+)] 造影 CT を施行したところ、[D(+)] 肺癌 に対する [R(+)] 放射線治療 の [回照射範囲に含まれる肝臓領域] に [F 多発] する [D(+)] 造影効果が乏しい小腫瘍 を認めた。

経過から [D(?) 多発肝臓癌または肺癌の多発転移] を疑い、[Tt(+)] 肝腫瘍生検 を施行したところ、病理組織所見は [D(+)] 凝固壊死と思われる肝細胞壊死 を [F 散見] し、[F 明らか] な [D(-)] 肺癌の肝浸潤所見 は認められなかった。

[Tt(+)] 血液培養 も施行したが [D(-)] 菌 は検出されなかった。

図 31 MedTxt データセットのサンプル例[65]

## ii) モデル

固有表現認識モデルのベースとして大量の日本語コーパスで事前学習を行なった学習済み言語モデルが必要であり、本研究では東北大学自然言語処理研究グループが公開した TohokuBERT[66]、早稲田大学河原研究室が公開した RoBERTa[67]、愛媛大学が公開した AcademicBART の三種類の事前学習モデルを使用する。この中 TohokuBERT と RoBERTa はトランスフォーマーのエンコーダを主要部品としたオートエンコーディング型モデルであり、AcademicBART はトランスフォーマーのエンコーダ・デコーダ構造をそのまま採用した系列変換モデルである。また、上記のモデルの事前学習タスクは固有表現認識ではないので固有表現認識というタスクに対応させるための改造を加える必要がある。TohokuBERT と RoBERTa の両モデルはエンコーダがメインな構造となっており BIO タグを解読するデコーダがないため、条件付き確率場と多層ペーセptron の二種類のデコーダをモデルの後にそれぞれ追加するという改造を行なった。AcademicBART はそもそもデコーダを持つ構造となっているが、元のデコーダの上にさらにモジュールを追加して同様の改造を行うことができる。最後に固有表現認識モデルの候補モデルとして、「TohokuBERT<sub>base</sub>-MLP」、

「TohokuBERT<sub>base</sub>-CRF」、「TohokuBERT<sub>large</sub>-MLP」、「TohokuBERT<sub>large</sub>-CRF」、「RoBERTa<sub>base</sub>-MLP」、「RoBERTa<sub>base</sub>-CRF」、「RoBERTa<sub>large</sub>-MLP」、「RoBERTa<sub>large</sub>-CRF」、「AcademicBART-MLP」、「AcademicBART-CRF」の10種類のモデルを準備した。

### iii) 固有表現認識性能の評価

MedTxt データセットに合計 3082 個のアノテーションされたシーケンスがある。これを 85% 対 15% の割合で学習用データセットと検証用データセットに分割し、学習データと検証データはそれぞれ 2619 と 463 個のシーケンスがある。この 2619 個の学習データを用いて事前に準備した 10 種類の候補モデルをファインチューニングし、463 個の検証データを用いて学習したモデルの性能を評価する。実験環境は 3.1.3 節の時と同じであり、学習条件は表 3 に示す。表 3 に示すように学習条件においてオートエンコーディング型の TohokuBERT 系、RoBERTa 系モデルと系列変換型の AcademicBART 系モデルは少し相違があり特に学習のエポック数は AcademicBART 系モデルに対してより多めに設定した。それは AcademicBART 系モデルの学習を 50 エポックだけ行えば精度が低く学習が不十分からである。

表 3 固有表現認識モデルの学習条件

学習パラメータ (TohokuBERT、RoBERTa 系)	
エポック数 (ステップ数)	50 (5, 500)
バッチサイズ	24
バッチ数	110
学習率	1e-5
学習データ数	2, 619
ウォームアップエポック数	5
学習パラメータ (AcademicBART 系)	
エポック数 (ステップ数)	100 (11, 000)
バッチサイズ	24
バッチ数	110
学習率	1e-5
学習データ数	2, 619
ウォームアップエポック数	5

上記の学習条件のもと 10 個のモデルを学習し精度の評価を行なった。評価時に、一般的な固有表現認識評価指標である F1 スコアを利用する。固有表現認識では固有表現に付与された BIO 形式のラベルを全部認識し、かつ正しいクラスに分類した場合のみ、その固有表現の認識に関してモデルが正解したとみなす。モデルが認識した固有表現で正しい固有表現の数を真陽性数 (True Positive; TP) と呼び、モデルが認識した固有表現の中誤

ったものの数を偽陽性数 (False Positive; FP) と呼ぶ。モデルが認識できなかった正しい固有表現の数を偽陰性数 (False Negative; FN) と呼ぶ。この三つの値を使用し、下記のように適合率 (Precision) と再現率 (Recall) を計算することができる。さらに、適合率 (Precision) と再現率 (Recall) の調和平均として F1 スコアを式 (40) ように計算できる。

$$recall = \frac{TP}{TP + FN} \quad (38)$$

$$precision = \frac{TP}{TP + FP} \quad (39)$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (40)$$

F1 スコアに基づき 463 個の評価データを用いて学習したモデルの固有表現認識精度を評価した結果を表 4 に示す。

表 4 評価データに対する各モデルの固有表現認識精度

モデル	F1 スコア
TohokuBERT <sub>base</sub> -MLP	0.9051
TohokuBERT <sub>base</sub> -CRF	0.9026
TohokuBERT <sub>large</sub> -MLP	0.9159
TohokuBERT <sub>large</sub> -CRF	0.9092
RoBERTa <sub>base</sub> -MLP	0.9056
RoBERTa <sub>base</sub> -CRF	0.9031
RoBERTa <sub>large</sub> -MLP	<b>0.9162</b>
RoBERTa <sub>large</sub> -CRF	0.9063
AcademicBART-MLP	0.8880
AcademicBART-CRF	0.8828

表 4 に示すように、RoBERTa<sub>large</sub>-MLP が評価データで最も良い精度を実現した。また、多層ペーセプトロン (MLP) をデコーダとして使用したモデルが条件付き確率場をデコーダとしたモデルよりも精度が高い傾向が見られる。BERT や RoBERTa といったオートエンコーディング型モデルをベースとしたモデルが自己回帰型の BART 系モデルよりも精度が高く、固有表現認識のような自然言語理解タスクでは自己回帰型モデルがオートエンコーディング型モデルに見劣りすることが確認される。10 個のモデルいずれも 9 割前後の精度を達成しており、特にベストモデルは 0.9162 という高い精度を実現したことから、医療分野に特化した固有表現認識モデルの構築が成功したと言える。

### 3.2.3 固有表現認識によるワードクラウドの作成およびキーコンセプトの抽出

前節では医療分野の固有表現認識に特化した深層学習モデルを構築した。その結果を踏まえ、性能が最も良い RoBERTa<sub>large</sub>-MLP モデルを固有表現認識モデルとして採用し、多段階関係トリプル抽出アルゴリズムの第二段階で使用する。第二段階では、固有表現認識による各章に関するワードクラウドの作成およびキーコンセプトの抽出を行なった。

まず、肝細胞がん診療ガイドラインの全文に対して固有表現認識モデルを適用し、テキスト中の病名、検査名、薬品名と処置名の四種類の重要な固有表現に絞り、テキストから大量の固有表現を抽出した。ここで病名、検査名、薬品名と処置名の四種類だけに着目したのは、この四種類の固有名詞は多数の属性と属性値を持ちうるのでそこから関係トリプルを作成することができると考えるからである。各章のテキストから抽出した四種類の固有表現に基づいて各固有表現の出現頻度を反映する各章の内容に関するワードクラウドを9個それぞれ作成した。作成した各章のワードクラウドは図32に示す。

## 第一章 診断およびサーベイランス



図 32(a) 第一章に関するワードクラウド

## 第二章 治療アルゴリズム



図 32(b) 第二章に関するワードクラウド

第三章 予防



図 32(c) 第三章に関するワードクラウド

## 第四章 手術

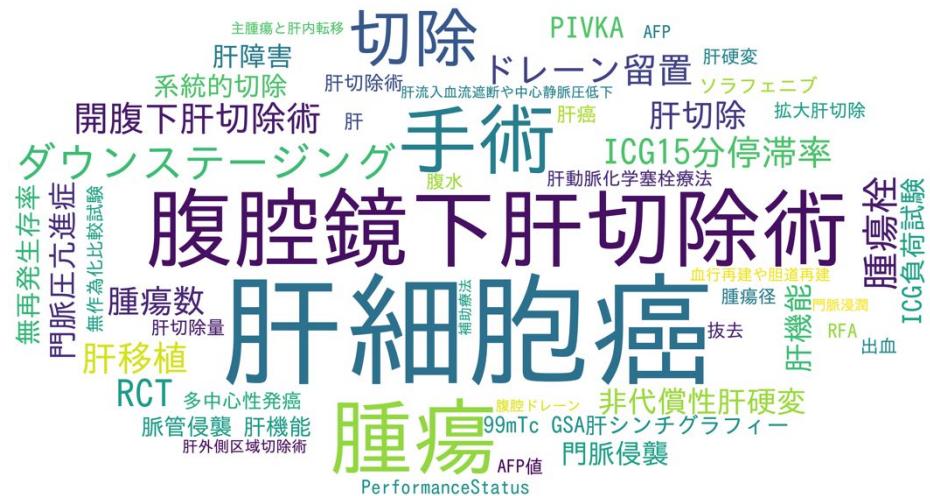


図 32(d) 第四章に関するワードクラウド

## 第五章 穿刺局所療法



図 32(e) 第五章に関するワードクラウド

## 第六章 肝動脈（化学）塞栓療法TA (C) E



図 32(f) 第六章に関するワードクラウド

## 第七章 薬物療法



図 32(g) 第七章に関するワードクラウド

## 第八章 放射線治療



図 32(h) 第八章に関するワードクラウド

第九章 治療後のサーベイランス



図 32(i) 第九章に関するワードクラウド

図32のように、肝細胞がん診療ガイドラインから抽出した重要な固有表現を章ごとにワードクラウドとして可視化した。これらの図から、各章のテーマおよび大まかな内容について把握できる。ワードクラウドから読み取れる特徴として、「肝細胞がん」という固有表現はほぼ全ての章に高頻度で出現しており、肝細胞がん診療ガイドラインにおける最重要的コンセプトと言っても過言ではない。また、第一章の「腫瘍マーカー」や第五章の「穿刺局所療法」のような他の章にあまり現れず各章独自の固有表現も多数存在することがわかる。

次に、固有表現認識モデルにより抽出された全て固有表現からキーコンセプトの選定を行う。方法としてはシンプルで抽出された各固有表現の出現頻度を集計して出現頻度が上位のものを全体から洗い出し関係トリプル抽出にあたってのキーコンセプトとする。具体的に固有表現として抽出された回数（＝固有表現の出現頻度）が20回以上の固有表現をキーコンセプトとする。ここで作られたキーコンセプトは後の段階で使用する。抽出されたキーコンセプトに基づいて第三段階で肝細胞がん診療ガイドライン中のどの文に対して関係トリプルを抽出するかという抽出対象文の選別を行い、第四段階で関係トリプルの抽出対象となる文の中から各キーコンセプトの持つ属性とその属性値を抽出し、キーコンセプトを属性および属性値と連結させて「キーコンセプト-属性-属性値」のように関係トリプルを作成する。

### 3.2.4 構文解析による対象文の選別

第三段階では肝細胞がん診療ガイドラインの全ての文から関係トリプルの抽出対象文となる文の選別を行う。肝細胞がん診療ガイドラインに大量の文が格納されているが、その全てが関係トリプルの抽出対象として適切であるというわけではない。キーコンセプトが全く出現していない文であればその文を抽出の対象として関係トリプルを抽出することができないのである。キーコンセプトが現れた文であってもキーコンセプトを含む文節がその文にとって重要な成分でなければ、キーコンセプトが文の中に属性と属性値を持たない可能性が高く関係トリプル抽出の対象とする必要がない。本研究では、一文の主語か目的語に任意のキーコンセプトが現れることを基準として肝細胞がん診療ガイドラインの中の基準を満たした文のみ関係トリプルの抽出対象とするように選別を行った。文の主語と目的語を特定するために日本語自然言語処理ライブラリである GiNZA の提供する構文解析機能を利用する。

GiNZA[68]とはリクルート社と国立国語研究所によって共同開発されたオープンソースの日本語自然言語処理ライブラリである。GiNZA は自然言語処理ライブラリ spaCy をフレームワークとして利用し内部に自然言語処理向けの深層学習モデルを組み込むことで形態素解析や品詞タグ付け、構文解析といった高度な自然言語処理タスクを効率的に実行することができる。本研究で利用するのはバージョン5の GiNZA v5.0.0 であり内部に ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) [69] という深層学習モデルが組み込まれている。ELECTRA とは、敵対的生成

ネットワークのアイデアを取り込み BERT の事前学習方法を改良したオートエンコーディング型モデルである。GiNZA に学習済みの ELECTRA が搭載されるので利用の際は特別な学習を行う必要はない。

与えられた文の主語と目的語にキーコンセプトが存在するかどうかを判定する方法として、文に対して GiNZA を利用して構文解析を行ない各形態素に予測された依存関係ラベルに基づいて主語と目的語を抽出しキーコンセプトの有無をチェックする。図 33 に示すように、構文解析を行なった結果、文の中の各形態素に依存関係ラベルと係り受けが判定される。依存関係ラベルは、与えられた文において単語の持つ文法的機能を表す。依存関係ラベルに注目することで文のどの部分が主語・目的語であるかを知ることができる。主語の場合、「nsubj」、「csubj」、「dislocated」のラベルが予測された形態素に着目し、その形態素の修飾する全て形態素、すなわち全ての係り先を再帰的に辿り連結させて主語とする。目的語に関しても同様の方法で抽出する。目的語の場合「obj」、「iobj」、「obl」、「advcl」、「acl」、「nmod」、「amod」、「advmod」のラベルが予測された形態素に着目し、該当する形態素の係り先を再帰的に辿り連結させたものを目的語とする。主語は通常一つしか抽出されないので対して目的語として抽出された単語のひとまとめりは複数個ありうる。実はこのように抽出された単語のひとまとめりは厳密に言語学における目的語ではないが本研究では便宜上「目的語」と呼ぶ。図 33 の例では「肝細胞がん」が主語として「高危険群の設定が容易な」というひとまとめりが目的語として抽出され、キーコンセプトが存在することが容易にわかるのでこの文を関係トリプル抽出の対象文とする。肝細胞がん診療ガイドラインに含まれる全ての文に対してこのように抽出対象文の選別を行なった。

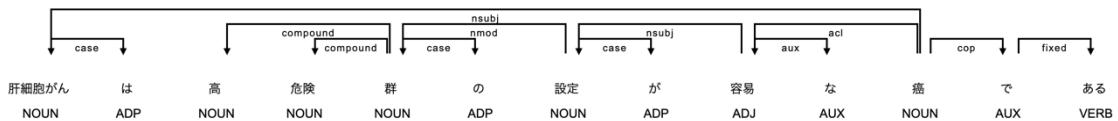


図 33 GiNZA による構文解析例

### 3.2.5 大規模言語モデルによる属性と属性値の抽出

多段階関係トリプル抽出アルゴリズムの第四段階は属性と属性値の抽出である。関係トリプルの抽出対象として選定された文からその中に存在するキーコンセプトの持つ属性と属性値を、OpenAI が開発した大規模言語モデル GPT-4 を利用して抽出する。具体的に、OpenAI の提供する Python ライブラリを利用して GPT-4 の Chat Completions API を利用し、表 5 に示すテンプレートのように system メッセージとしてタスクの内容を指定し、user メッセージとして関係トリプルの抽出対象となる文とその文に含まれるキーコンセプトを指定する。関係トリプルの抽出対象文に複数個のキーコンセプトが含まれる場合、それぞれのキーコンセプトに対して以下のテンプレートに従い属性と属性値の抽出を行う。毎回は一つのキーコンセプトだけ指定する。

表5 GPT-4への入力形式

Role	Content
system	Based on the given sentence, please extract only one property for the given entity. The extracted property should be the most appropriate property held by the given entity. Please provide the property's name and its value in your answer following the example given below. If there is no appropriate property for the given entity in the given sentence, please answer with "No property". Moreover, the contents that you answered should be in Japanese. You can perform this task just like this example. Sentence : 「肝細胞癌は高危険群の設定が容易な癌である」、Entity : 「肝細胞癌」 ->属性 : 特性、属性値 : 高危険群の設定が容易である
user	Sentence : 「(ここに対象文を入れる)」、Entity : 「(ここにキーコンセプトを入れる)」

表5に示すように、GPT-4に与えるプロンプトは英語で書かれているが対象文とキーコンセプトは日本語であり、出力も日本語で書くようにGPT-4に指示した。上記のsystemメッセージの内容を日本語に訳すとこのようになる：「以下の文に基づいて、指定されたエンティティについて、そのエンティティによって保持されているプロパティの中で最も適切なものを1つだけ抽出してください。抽出されたプロパティは、そのエンティティが持っている最も適切なプロパティであるべきです。例に従って、回答にはプロパティの名前とその値を含めてください。指定された文中に指定されたエンティティに適切なプロパティがない場合は、プロパティなしと回答してください。また、あなたが回答した内容は日本語であるべきです。以下の例のようにこのタスクを実行できます。Sentence : 「肝細胞癌は高危険群の設定が容易な癌である」、Entity : 「肝細胞癌」 ->属性 : 特性、属性値 : 高危険群の設定が容易である」。この日本語訳はChatGPTが翻訳したものである。この翻訳から分かるように、GPT-4へのプロンプトに実際の処理例が一つ含まれ、それは大規模言語モデルGPT-4にフューショット学習能力を発揮させるためであり、一種のプロンプトエンジニアリングである。

表 6 大規模言語モデルによる抽出結果

<b>抽出対象文</b>	DynamicCT/MR における肝細胞癌の典型的所見とは、動脈相で高吸収域として描出され（早期濃染），門脈・平衡相で周囲肝実質よりも相対的に低吸収域として描出される（washout）結節と定義されている。	CTAP および CTHA を含む血管造影は典型的肝細胞癌の診断において非常に有用な検査である。
<b>キーコンセプト</b>	肝細胞癌	血管造影
<b>抽出された属性</b>	典型的所見	用途
<b>抽出された属性値</b>	動脈相で高吸収域として描出され（早期濃染），門脈・平衡相で周囲肝実質よりも相対的に低吸収域として描出される（washout）結節	典型的肝細胞癌の診断

大規模言語モデルによる属性と属性値の抽出例を表 6 に示す。表から大規模言語モデルは対象文から属性と属性値を見事に抽出していることがわかる。抽出された属性および属性値をキーコンセプトと連結させて「キーコンセプト-属性-属性値」のように関係トリプルを作成することができる。これまで説明した多段階関係トリプル抽出アルゴリズムを利用することにより、知識グラフのもととなる関係トリプルを肝細胞癌診療ガイドラインのテキストから抽出することが実現される。抽出された大量の関係トリプルを知識グラフとして可視化するツールとして、グラフ型データベースである Neo4j を利用することにする。Neo4j とは Java で開発されたオープンソースのグラフ型データベースである。MySQL のような関係型データベースとは異なり、Neo4j はノード、エッジ、およびプロパティを使用してデータを格納しクエリすることに特化しており、保存されているデータをグラフのように可視化できるという大きな長所も備えている。Neo4j のクエリ言語である Cypher を使用して関係トリプルに基づいてノードとエッジを作ることで抽出された関係トリプルを Neo4j にインポートし、肝細胞がん診療ガイドラインに関する知識グラフを構築した。かなり大きな知識グラフとなっているため全体から一部を抜き取って図 34 にそのイメージを示す。

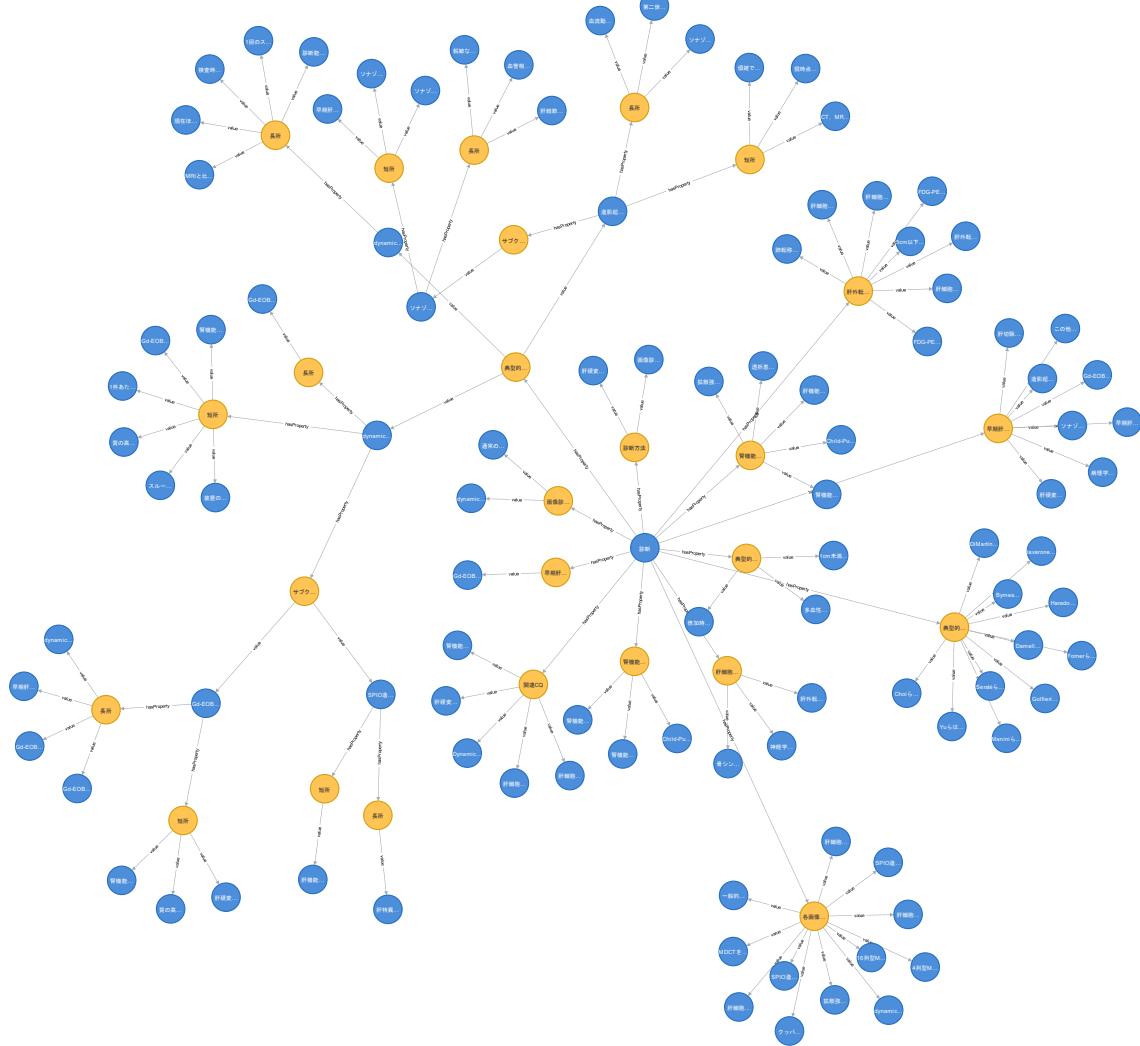


図 34 構築した知識グラフ（一部）

図 34 に示すように構築した知識グラフは青色と黄色の 2 色のノードがあり、黄色のノードは属性を表し、青色のノードはキーコンセプトと属性値の二種類のエンティティを表す。本来は属性をノードではなく二つのノードを結ぶエッジで表現すべきだが、用いたデータベース Neo4j ではエッジに属性の文字列を付与するのが機能上の制限により困難であるため、属性もノードで表すこととした。構築した知識グラフの中のエッジは全部有向エッジであり意味も単純で、キーコンセプトのノードと属性のノードを結ぶエッジは始点のキーコンセプトが終点の属性を持つことを意味し、属性のノードと属性値のノードを結ぶエッジは始点の属性が終点の属性値を持つことを意味する。

### 3.3 構築した知識グラフの評価

本研究では、構築した知識グラフに対する評価方法として直接的評価と間接的評価の二つのアプローチで評価を行う。直接的評価は、外部の専門家、特に肝臓専門医を招いて知識グラフの内容について直接的に評価してもらう方法である。間接的評価は、知識グラフに基づいて質疑応答システムを構築し、質疑応答システムの性能をもって知識グラフの質を評価する方法である。間接的評価については第五章で記述し、本節では主に直接的評価について説明する。

[70]によると、知識グラフの直接的評価において一般的な評価基準として正確性 (Accuracy) と完全性 (Integrity) の二つの基準が用いられる。正確性は関係トリプルの内容の正しさを意味し、キーコンセプト自体が医学的名詞として正しく、関係トリプルの中の属性が本当にキーコンセプトの持つ属性でありかつその属性値も正しい場合のみ関係トリプルが正しいと判定する。完全性は主に属性に着目する評価基準であり、キーコンセプトに対して抽出された属性が完全であるかどうか、それ以外に抽出されていない属性が存在するかどうかで完全性を判定する。ただし、あるキーコンセプトに対してどのぐらいの属性を抽出すべきかを定義するような正解データがなく専門家でももなく指摘するのが難しい時があり、さらに「抽出すべき」ということに主観性があり曖昧性も存在するので完全性という評価基準を定量化することが難しい。この点を踏まえ正確性という基準に対して定量化し、完全性に対して定性的に評価することにする。

直接的評価に必要な外部専門家として国立国際医療研究センターの肝臓専門医三名に評価を依頼した。評価の方法としては知識グラフからランダムに「サーベイランス」、「肝移植」と「診断」の三つのキーコンセプトを選び、三名の肝臓専門医にこの三つのキーコンセプトに関する関係トリプルを正確性と完全性の二つの観点から評価してもらった。

評価の結果として、選ばれた関係トリプルの正確性が 95% という高い数値に届きながらも完全性が低く肝細胞がん診療ガイドラインから抽出すべきだが実際に抽出されていない属性や属性値が一部存在する。例えば診断というキーコンセプトに対して方法という属性に血液検査という属性値が抽出すべきものだが抽出されなかった。肝移植の持つ適応という属性に対しても少なくとも二点の属性値の漏れがある。専門医によって評価してもらった結果として完全性に欠けるという不足点が存在しながらも知識グラフ自体の正確性が高いことから、本来の目的である肝細胞がん診療ガイドラインに含まれる知識の構造化および可視化を達成したと言える。

## 4 肝細胞がん診療ガイドラインを対象とした質疑応答システムの構築

第三章では、提案した多段階関係トリプル抽出アルゴリズムを用いて肝細胞がん診療ガイドラインに関する知識グラフを構築し、肝細胞がん診療ガイドラインに含まる知識の構造化および可視化を実現した。本章では、本研究の主要目的である肝細胞がん診療ガイドラインからの情報検索の効率性と利便性の向上を実現するために肝細胞癌診療ガイドラインを対象とした質疑応答システムの構築を行う。2.6.2節で説明したように肝細胞がん診療ガイドラインに関する質疑応答システムを構築するにあたって専門性の付与、基礎的言語力の必要性、適切なデータによる評価の三つの課題が存在する。そこで本研究では、質疑応答システムのベースとしてOpenAI社の公開した大規模言語モデルGPT-3.5を利用し第三章で構築した肝細胞がん診療ガイドラインに関する知識グラフの質疑応答システムへの組み込みを提案することにより基礎的言語力の習得と専門性の付与という二つ課題の克服を試みる。さらに、肝細胞がん診療ガイドラインの中のクリニカルクエスチョンという本タスクと関連性の高いデータを質疑応答システムの評価に使用することで適切なデータによる評価という課題の解決に取り組む。

### 4.1 構築手法

質疑応答システムの構築に存在する基礎的言語力の習得という課題を克服するために、本研究では基礎的な自然言語能力を備えた大規模言語モデルGPT-3.5-turboをベースに質疑応答システムを構築する。また、今回の構築目標はドメイン特化型の質疑応答システムであるため専門性の付与という課題もあり、これに対処するために肝細胞がん診療ガイドラインの内容を何らかの形で質疑応答システムに習得させることが必要となる。そこで二つの候補案として、肝細胞がん診療ガイドラインのテキストを直接に大規模言語モデルのファインチューニングに使用する案と肝細胞がん診療ガイドラインをもとに構築した知識グラフを用いて大規模言語モデルをファインチューニングする案を提案する。

大規模言語モデルGPT-3.5-turboは生成式の自己回帰型モデルであり、それをファインチューニングするには複数の質問と解答例からなる質問応答式データが必要である。つまり、モデルに質問を一個与えて回答を受け取り、モデルによる回答と事前に用意した正しい解答との差異を縮小させるように内部にある大量のパラメータを調整するという形でモデルのファインチューニングが行われる。そのため、ファインチューニングのデータとして利用する肝細胞がん診療ガイドラインのテキストと肝細胞がん診療ガイドラインに関する知識グラフを質問応答式のデータに変換する必要がある。

肝細胞がん診療ガイドラインの各章に図1に示すように前書きやクリニカルクエスチョンのような複数のセクションによって構成されている。このような構造を利用して肝細胞

がん診療ガイドラインのテキストを質問応答式データに変換する。例えば「肝細胞がん診療ガイドラインの第一章の前書きにどのような内容が書かれていますか」という質問を置き、対応する解答例として前書きに書かれた内容をそのまま利用する。各部分に何が書かれているのかを基本的な質問形式として肝細胞がん診療ガイドラインの本文を形式的な質問応答式データに変換した。

また、肝細胞がん診療ガイドラインに関する知識グラフに対して属性と属性値のペアを利用して関係トリプルを質問応答式のデータに変換した。質問として関係トリプルから「血管造影の用途は何ですか」のようにキーコンセプトの持つ属性の属性値について訊ねる質問を作り、応答として対応する属性値「典型的肝細胞癌の診断」をその質問に対する解答例とする。一つの属性に複数の属性値が存在することがあるがその場合複数の属性値を読点で前後に連結させることで一つの属性値としてまとめる。そうすると同様に質問応答式データに容易に変換できる。作成した二つの質問応答式データセットを用いてOpenAIの提供するAPIを介して大規模言語モデルGPT-3.5-turboをそれぞれ10エポックファインチューニングする。ファインチューニング時に使った大規模言語モデルへのプロンプトは表7に示す。

表7 ファインチューニング時のプロンプト

Role	Content
system	あなたは肝細胞がん診療に関する質問に回答できるAIチャットボットです。ユーザーからの肝細胞がん診療に関する質問に答えてください。
user	(ここに用意した質問データを入れる)

## 4.2 質疑応答システムの性能評価

上記のように肝細胞がん診療ガイドラインのテキストと肝細胞がん診療ガイドラインに関する知識グラフをそれぞれ用いて大規模言語モデルをファインチューニングすることで、大規模言語モデルGPT-3.5-turboをベースとした質疑応答システムを二つ構築した。構築した質疑応答システムの性能に対する評価方法として、2.6.2節で述べたように汎用的な日本語質疑応答データセットでは不十分で肝細胞がん診療と関連性の高いデータによる評価が必要であり、そこで肝細胞がん診療ガイドラインにあるクリニカルクエスチョンを利用して構築した質疑応答システムの性能を評価することにした。質疑応答システムの構築時に、評価用の質問と回答が学習データに明示的に現れないように注意したのでクリニカルクエスチョンによる評価は妥当であると考える。

## CQ3

肝細胞癌の診断に有用な腫瘍マーカーは何か？

### 推奨

肝細胞癌の補助診断に有用な腫瘍マーカーとして、 AFP, PIVKA-II, AFP-L3 分画が推奨される。**(強い推奨)**

図 35 クリニカルクエスチョンの例

図 35 に示すようにクリニカルクエスチョンは質問文と、それに回答する推奨文から構成される。この推奨文はクリニカルクエスチョンに対する正解とみなすことができる。質問文と推奨文をそれぞれ質問と正解としてそのまま利用することで評価データを作成することができるが、正解は普通長い文であるためそれに基づいて質疑応答システムの出力した回答が正しいかどうかを判定するのが困難な場合がある。そこで、質疑応答システムによる回答の正確性評価を容易にするために、各クリニカルクエスチョンを事実確認式問題、選択式問題、穴埋め問題の三種類の質問に変換した。質問例は表 8 に示す。

表 8 評価データの質問例

問題の種類	質問	正解
選択式問題	以下の選択肢の中から、推奨される穿刺局所療法を選んでください。1: RFA; 2: MCT; 3: IRE	1
事実確認式問題	「肝切除前に補助療法を行う必要はある」という記述は正しいでしょうか。	正しくない
穴埋め問題	穿刺局所療法の効果判定に有用な画像診断は()である。空欄()に入る適切な語句を答えなさい。	dynamic CT/MRI

表 8 のように各クリニカルクエスチョンを質問文と推奨文から一部のみ抜き取り事実確認式問題、選択式問題、穴埋め問題のいずれかの形式に変換した。このように正解と質疑応答システムの出力がシンプルになり評価も容易になる。57 個の質問と正解を作成し評価データとして構築した質疑応答システムの評価を行なった。評価を行うにあたり、肝細胞がん診療ガイドラインのテキストをベースに構築した質疑応答システムを GuidelineQA-Text と呼び、肝細胞がん診療ガイドラインに関する知識グラフをベースに構築した質疑応答システムを GuidelineQA-KG と呼ぶ。また、構築した質疑応答システムの有効性を検証するための比較対象としてファインチューニングなしのオリジナルな GPT3.5-turbo モデルに対しても作成した評価データによる評価を行なった。

表9 質疑応答システムの評価結果

システム	正解率
Original GPT-3.5-turbo	54.39%
GuidelineQA-Text	75.44%
GuidelineQA-KG	<b>89.47%</b>

表9に示すように、ファインチューニングなしのオリジナルなGPT3.5-turboモデルの正解率は54.39%であると最も低く、肝細胞がん診療ガイドラインに関する知識グラフをベースに構築したGuidelineQA-KGは最も高い89.47%の正解率を実現した。GuidelineQA-Textは75.44%という2番目に高い正解率を達成しているがGuidelineQA-KGに及ばなかった。

## 5 考察

### 5.1 構築した知識グラフに関する考察

肝細胞がん診療ガイドラインに関する知識グラフの構築にあたりタスクの複雑性とデータの不足という二つの課題が存在する。これらの課題を解決するために、関係トリプルの抽出を複数の段階に分解することによりタスクの複雑性を克服し、限られたデータを有効活用する BERT と大量の学習データを必要としないフューショット能力の強い大規模言語モデル GPT-4 を提案した多段階抽出アルゴリズムを通して有機的に組み合わせることによりデータの不足に対処した。このように構築した知識グラフを三名の肝臓専門医に直接的評価のアプローチで評価してもらった結果、知識グラフの内容に関して高い正確性が認められた。また、間接的評価として第四章で肝細胞がん診療ガイドラインに関する知識グラフをベースに質疑応答システム GuidelineQA-KG を構築し、GuidelineQA-KG を肝細胞がん診療ガイドラインのクリニカルクエスチョンをもとに作成した質疑応答データを用いて評価した結果、性能の評価基準として 89.47% の正解率を達成し、構築時に知識グラフを利用しなかった二つの質疑応答システム GuidelineQA-Text と Original GPT-3.5-turbo の正解率をそれぞれ 14%、35% 上回っている。評価に使用した質疑回答データは全部肝細胞がんの診療というドメインと密接に関わっており専門性を要する質問であるため、知識グラフの導入により質疑応答システムの専門性を大幅に向上させたと言える。以上のことから、第三章で構築した肝細胞がん診療ガイドラインに関する知識グラフが有効であることが明らかとなった。

しかし、構築した知識グラフは完璧とは言えず完全性に欠けるという不足点がまだ存在する。その原因としてまず固有表現認識モデルの精度による影響が挙げられる。構築した固有表現認識モデルは確かに評価データにおいて 91% の固有表現を正しく認識できているが全ての固有表現を完全に抽出することができず、肝細胞がん診療ガイドラインのテキストに対して固有表現認識を行う際も一部抽出できなかつた固有表現があるため知識グラフの完全性を下げてしまう可能性がある。固有表現認識モデルの学習に使用したデータは医療分野に関するものであるが肝細胞がん診療ガイドラインと微妙なコンテキストの不一致があるため学習のシーンと使用のシーンが完全に一致するとは言えず、この点も実際の固有表現認識性能に影響することがある。

また、本研究では大規模言語モデルを利用し対象文から属性と属性値を抽出する際に最も適切な属性と属性値を一組だけ抽出するようにモデルに指示した。これは大規模言語モデルのハルシネーションによる誤抽出を防ぐための工夫であるが抽出された属性の数を減らしてしまい本来抽出すべきものを抽出できなかつたことにより属性の不完全につながるという逆効果も同時にあると認めざるを得ない。

もう一つの原因として今回は文の中に存在する属性だけに着目し抽出を行った。しかしキー・コンセプトの属性は複数の文にまたがって存在する可能性もあるため一文の中だけ着

目すると document-level に存在する属性と属性値を無視してしまうことになる。この点も構築した知識グラフの完全性に負の影響を与える可能性がある。

構築した知識グラフの完全性不足の原因として主に以上の三点が考えられる。今後の研究でこの三点の課題を解決することでより完全性の高い知識グラフを構築することができると考える。

## 5.2 構築した質疑応答システムに関する考察

肝細胞がん診療ガイドラインを対象とした質疑応答システムを構築する上で専門性の付与、基礎的言語力の必要性、適切なデータによる評価の三つの課題が存在する。質疑応答システムに専門性を付与するために肝細胞がん診療ガイドラインに関する知識グラフの質疑応答システムへの組み込みを提案した。基礎的言語力の必要性という課題を解決するために大規模言語モデル GPT-3.5-turbo をベースとしてファインチューニングすることを提案した。適切なデータによる評価に関して、肝細胞がん診療ガイドラインにあるクリニカルクエスチョンをもとにドメイン関連性の高い評価データを作成し、構築した質疑応答システムを対象として評価を行なった。その結果、肝細胞がん診療ガイドラインに関する知識グラフに基づいて構築した質疑応答システム GuidelineQA-KG が三種の質疑応答システムの中で最も高い約 9 割の正解率を実現した。質疑応答システム GuidelineQA-KG の肝細胞がん診療に関する質問への高度な回答性能が証明されており、肝細胞がん診療に関する情報検索の利便性と効率性の向上という本研究の目的が達成されたと言える。

GuidelineQA-KG が最も良い性能を実現できた原因として、肝細胞がん診療ガイドラインによる専門性の付与が一つの要因として挙げられる。具体的な使用方法が異なるが肝細胞がん診療ガイドラインを学習に用いた GuidelineQA-Text と GuidelineQA-KG の二つ質疑応答システムは、使用しなかった Original GPT-3.5-turbo よりも 20%以上正解率が高かつた。このことから、肝細胞がん診療ガイドラインによる専門性の付与の重要性が窺える。

また、学習に使用したデータの質も性能に影響する要因である。肝細胞がん診療ガイドラインのテキストを使った GuidelineQA-Text は元のテキストを特別に処理せずにそのまま質疑応答形式に変換して学習に使用した。一方で知識グラフをベースとした GuidelineQA-KG はテキストから知識を抽出するという処理を通して肝細胞がん診療ガイドラインの中の重要な内容だけ残し不必要な内容を削除したとみなすことができる。知識グラフは肝細胞がん診療ガイドラインのテキストを深層学習モデルにとってより処理しやすい形に構造化した。これらのことから、GuidelineQA-KG はより良い学習データを使用したと考えられ、同じく肝細胞がん診療ガイドラインをベースに構築した質疑応答システム GuidelineQA-Text よりも高い精度を達成できた。

ただし、精度的にやや見劣りするが GuidelineQA-Text にも長所がある。例えば、「肝外転移の危険因子を有する肝細胞癌患者に対して胸部 CT、FDG-PET を施行することは推奨されますか。以下の選択肢から正しい答えを選びなさい。 1：推奨される、2：推奨され

ない」という質問に、GuidelineQA-Text と Original GPT-3.5-turbo の二つの質疑回答システムが正解し、GuidelineQA-KG が誤った。

また、評価データとして使わなかつたクリニカルクエスチョンに対する三つの質疑応答システムの回答は表 10 に示す。与えた質問は「肝移植後の再発に対する有効な治療法は何ですか」である。このクリニカルクエスチョンに対する肝細胞がん診療ガイドラインの態度は「肝移植後の再発に対しては可能であれば再発病巣の切除を、不可能であれば分子標的薬を考慮してもよいが十分なエビデンスはない」である。この意味では Original GPT-3.5-turbo の回答が誤りで GuidelineQA-Text と GuidelineQA-KG が同時に正解し、それぞれの回答に肝細胞がん診療ガイドラインの「十分なエビデンスはない」という態度も反映されている。GuidelineQA-KG に比べて GuidelineQA-Text の長所は回答がより詳細に書かれていることである。これは、GuidelineQA-Text の学習データとしてガイドライン本文をそのまま使用し長い段落が多数含まれているからだと考える。

表 10 質疑応答システムの回答例

システム	回答
Original GPT-3.5-turbo	肝移植後の再発に対する有効な治療法は、再発した疾患に応じて抗がん剤療法や放射線療法、免疫療法などがあります。また、定期的な検査やモニタリングが重要です。
GuidelineQA-Text	肝癌の肝移植後再発に対する効果的な治療法は確立されていないが、肝内再発や肺転移に対する切除や塞栓療法、局所照射療法で長期生存例が報告されている。再移植も有効な治療法の一つであるが、ドナー不明の場合は困難である。近年、再発肝癌に対してソラフェニブが有効であることが報告されており、ソラフェニブ投与を含む多施設試験が行われている。
GuidelineQA-KG	肝移植後の再発に対する有効な治療法は確立していないが、肝移植前と同様の治療が考慮される。局所療法、TACE、分子標的薬、免疫チェックポイント阻害薬などが考えられる。

一部の質問に対して GuidelineQA-KG が正しく答えなかつた原因として、学習に使用した肝細胞がん診療ガイドラインに関する知識グラフの質が主な原因として質疑応答システムの精度に影響している可能性がある。前節で述べたように本研究で構築した肝細胞がん診療ガイドラインに関する知識グラフには完全性の不足という問題点は依然として存在し、より良い知識グラフを構築し質疑応答システムの構築に使用することで質疑応答システムの性能を向上させることができると考える。

# 6 結論

## 6.1 研究内容のまとめ

本研究では、肝細胞がん診療ガイドラインに含まれる知識の構造化と可視化を実現し、肝細胞がん診療ガイドラインから情報検索の効率性と利便性を向上させるために、知識グラフと質疑応答システムの構築手法を提案し、深層学習技術を用いて肝細胞がん診療ガイドラインに関する知識グラフと質疑応答システムの構築および評価を行なった。

はじめに第二章では知識グラフと質疑応答システム構築のベースとなる深層学習技術について概観し、肝細胞がん診療ガイドラインに関する知識グラフおよび質疑応答システムを構築するにあたり存在する課題を説明した。第三章では REBEL によるエンドツーエンドの関係トリプル抽出性能について検証を行った。REBEL の性能が不十分だったこともあり第二章で提示したタスクの複雑性やデータの不足といった課題を解決するために、BERT、ELECTRA、GPT-4 の三つの深層学習モデルを組み合わせた知識グラフ構築のための多段階関係トリプル抽出アルゴリズムを提案し肝細胞がん診療ガイドラインに関する知識グラフを構築した。構築した知識グラフについて肝臓専門医による直接的評価を行なった。結果として構築した知識グラフに関して高い正確性が認められたが完全性に欠けるという不足点も指摘された。

第四章では、専門性の付与や基礎的言語力の必要性といった質疑応答システムを構築する上での課題に対処しつつ、第三章で構築した肝細胞がん診療ガイドラインに関する知識グラフと肝細胞がん診療ガイドラインのテキストをそれぞれ用いて大規模言語モデル GPT-3.5-turbo をファインチューニングすることで肝細胞がん診療ガイドラインを対象とした質疑応答システムを構築した。構築した質疑応答システムの評価を肝細胞がん診療ガイドラインの中のクリニカルクエスチョンを用いて行なった。肝細胞がん診療ガイドラインに関する知識グラフをベースに構築した質疑応答システム GuidelineQA-KG が最も良い性能を達成したことがわかった。

第五章では、構築した知識グラフの間接的評価を行い、知識グラフをベースに構築した質疑応答システム GuidelineQA-KG の高性能からその有効性を確認した。知識グラフに存在する完全性の不足という問題点についても原因を検討した。また、構築した質疑応答システムの高精度および不足について考察を加えた。

## 6.2 結論と今後の展望

本研究では、肝細胞がん診療ガイドラインに関する知識グラフと質疑応答システムを構築することにより、肝細胞がん診療ガイドラインに含まれる知識の構造化と可視化を実現し、肝細胞がん診療に関する情報検索の効率性と利便性の向上を実現した。

一方、本研究には構築した知識グラフの完全性の不足や質疑応答システムの性能向上といった課題が依然として残っており、今後の展望としては、より質の高い知識グラフを構築し質疑応答システムに組み込むことで質疑応答システムの性能向上を目指したい。

## 参考文献

- [1] 肝臓がん, がん情報サービス, 2023, <https://ganjoho.jp/public/cancer/liver/index.html>
- [2] がんの部位別統計, 日本対がん協会, 2023  
[https://www.jcancer.jp/about\\_cancer\\_and\\_knowledge/%E3%81%8C%E3%82%93%E3%81%AE%E9%83%A8%E4%BD%8D%E5%88%A5%E7%B5%B1%E8%A8%88#male\\_d](https://www.jcancer.jp/about_cancer_and_knowledge/%E3%81%8C%E3%82%93%E3%81%AE%E9%83%A8%E4%BD%8D%E5%88%A5%E7%B5%B1%E8%A8%88#male_d)
- [3] 肝細胞がん, 大阪赤十字病院がん診療センター, 2023, <https://www.osaka-med.jrc.or.jp/cancer2/each/cancer4.html>
- [4] 肝細胞がん診療ガイドライン, 日本肝臓学会, 2023,  
[https://www.jsh.or.jp/medical/guidelines/jsh\\_guidlines/medical/](https://www.jsh.or.jp/medical/guidelines/jsh_guidlines/medical/)
- [5] L. Ehrlinger, et al., Towards a Definition of Knowledge Graphs, International Conference on Semantic Systems, 2016, <https://ceur-ws.org/Vol-1695/paper4.pdf>
- [6] P. James, Knowledge Graphs, In Linguistic Instruments in Knowledge Engineering, pages 97-117, Elsevier Science Publishers B.V., 1992.
- [7] A. Singhal, Introducing the Knowledge Graph: Things, not Strings, 2012,  
<https://googleblog.blogspot.co.at/2012/05/introducing-knowledge-graph-things-not.html>
- [8] Google ナレッジグラフの仕組み, ジオコード, 2022,  
<https://gc-seo.jp/journal/knowledge-graph/>
- [9] K. Kawasaki, Information extraction in structuring non-structured data, 2020,  
<http://id.nii.ac.jp/1001/00208024/>
- [10] S. Ji, et al., A Survey on Knowledge Graphs: Representation, Acquisition, and Applications, in IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 2, pp. 494-514, 2022, doi:10.1109/TNNLS.2021.3070843.
- [11] Z. Huang et al., Recent Trends in Deep Learning Based Open-Domain Textual Question Answering Systems, in IEEE Access, vol. 8, pp. 94341-94356, 2020, doi:10.1109/ACCESS.2020.2988903.
- [12] Watson, IBM, 2023, <https://www.ibm.com/watson>
- [13] A. Pandey, et al., A Review on Textual Question Answering with Information Retrieval and Deep Learning Aspect, 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2023, pp. 224-229, doi:10.1109/ICICCS56967.2023.10142729.
- [14] W. S. McCulloch, et al., A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5(4), 115-133. 1943
- [15] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 65(6), 386, 1958
- [16] [https://d2l.ai/chapter\\_multilayer-perceptrons/mlp.html](https://d2l.ai/chapter_multilayer-perceptrons/mlp.html)
- [17] D. E. Rumelhart, Learning representations by back propagating errors. nature, 323(6088), 533 536, 1986

- [18] CS 230 - Deep Learning, Standford University,  
<https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks>
- [19] Graves, A. (2013). Generating Sequences with Recurrent Neural Networks. ArXiv, abs/1308.0850.
- [20] S. Hochreiter and J. Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (November 15, 1997), 1735–1780.  
<https://doi.org/10.1162/neco.1997.9.8.1735>
- [21] LSTM, CVML エキスパートガイド, 2021, <https://cvml-expertguide.net/terms/dl/rnn/lstm/>
- [22] Mikolov, T., Chen, K., Corrado, G.S., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. International Conference on Learning Representations.
- [23] SydneyF, Word2vec for the Alteryx Community, 2018,  
<https://community.alteryx.com/t5/Data-Science/Word2vec-for-the-Alteryx-Community/ba-p/305285>
- [24] L. Wang, Learning Word Embedding, 2017, <https://lilianweng.github.io/posts/2017-10-15-word-embedding/>
- [25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.
- [26] 深層学習界の大前提 Transformer の論文解説, Qiita, 2022,  
<https://qiita.com/omiita/items/07e69aef6c156d23c538>
- [27] John S. Bridle. 1989. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In Proceedings of the 2nd International Conference on Neural Information Processing Systems (NIPS'89). MIT Press, Cambridge, MA, USA, 211–217.
- [28] Karl, F., & Scherp, A. (2022). Transformers are Short Text Classifiers: A Study of Inductive Short Text Classifiers on Benchmarks and Real-world Datasets. arXiv preprint arXiv:2211.16878.
- [29] Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training.
- [30] Ahmed Alajrami and Nikolaos Aletras. 2022. How does the pre-training objective affect what large language models learn about linguistic properties, In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 131–147, Dublin, Ireland. Association for Computational Linguistics.
- [31] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

- [32] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.
- [33] Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (GELUs). arXiv preprint arXiv:1606.08415.
- [34] J. Kaplan, et al., (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.
- [35] Brown, Tom, et al. Language models are few-shot learners. Advances in neural information processing systems 33 (2020): 1877-1901.
- [36] Ye, Junjie, et al. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. arXiv preprint arXiv:2303.10420 (2023).
- [37] Achiam, Josh, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023).
- [38] Y. Xiao, Knowledge Graph: Concepts and Techniques, 2020
- [39] M. Maurya, Name Entity Recognition and various tagging scheme, 2023, <https://medium.com/@muskaan.maurya06/name-entity-recognition-and-various-tagging-schemes-533f2ac99f52>
- [40] P. Sun, X. Yang, X. Zhao and Z. Wang, An Overview of Named Entity Recognition, 2018 International Conference on Asian Language Processing (IALP), Bandung, Indonesia, 2018, pp. 273-278, doi: 10.1109/IALP.2018.8629225.
- [41] K. Riaz, Rule-based named entity recognition in urdu, Proceedings of the 2010 named entities workshop. Association for Computational Linguistics, pp. 126-135, 2010.
- [42] T. Iwakura, A named entity recognition method using rules acquired from unlabeled data, Proceedings of the International Conference Recent Advances in Natural Language Processing 2011, pp. 170-177, 2011.
- [43] Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling (Finkel et al., ACL 2005)
- [44] W. Liao and S. Veeramachaneni, "A simple semi-supervised algorithm for named entity recognition", Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing. Association for Computational Linguistics, pp. 58-65, 2009.
- [45] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 260–270, San Diego, California. Association for Computational Linguistics.

- [46] Xuezhe Ma and Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.
- [47] Souza, Fábio, Rodrigo Nogueira, and Roberto Lotufo. "Portuguese named entity recognition using BERT-CRF." arXiv preprint arXiv:1909.10649 (2019).
- [48] 柴田 大作, 河添 悅昌, 篠原 恵美子, 嶋本 公徳. 診療テキストの構造化に向けた症例報告コーパスからの情報抽出, 2022,  
[https://doi.org/10.11517/pjsai.JSAI2022.0\\_1J4OS13a03](https://doi.org/10.11517/pjsai.JSAI2022.0_1J4OS13a03)
- [49] 坂地 泰紀, 増山 繁, テキストマイニングによる因果関係抽出, 2012,  
[https://www.jstage.jst.go.jp/article/jacc/54/0/54\\_0\\_124/\\_article/-char/ja/](https://www.jstage.jst.go.jp/article/jacc/54/0/54_0_124/_article/-char/ja/)
- [50] 佐藤 岳文, 堀田 昌英, Web マイニングを用いた因果ネットワークの自動構築手法の開発, 2006, <https://doi.org/10.3392/sociotechnica.4.66>
- [51] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 207–212, Berlin, Germany. Association for Computational Linguistics.
- [52] Ji, Guoliang, et al. Distant supervision for relation extraction with sentence-level attention and entity descriptions. Proceedings of the AAAI conference on artificial intelligence. Vol. 31. No. 1. 2017. <https://ojs.aaai.org/index.php/AAAI/article/view/10953>
- [53] Sugimoto K, Wada S, Konishi S, Okada K, Manabe S, Matsumura Y, Takeda T. Extracting Clinical Information From Japanese Radiology Reports Using a 2-Stage Deep Learning Approach: Algorithm Development and Validation. JMIR Med Inform. 2023 Nov 14;11:e49041. doi: 10.2196/49041. PMID: 37991979; PMCID: PMC10686535.
- [54] Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation Extraction By End-to-end Language generation. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- [55] Pere-Lluís Huguet Cabot, Simone Tedeschi, Axel-Cyrille Ngonga Ngomo, and Roberto Navigli. 2023. REDFM: a Filtered and Multilingual Relation Extraction Dataset. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4326–4343, Toronto, Canada. Association for Computational Linguistics.
- [56] Real Life Application of a Question Answering System Using BERT Language Model (Alloatti et al., SIGDIAL 2019)
- [57] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv, abs/1910.01108.
- [58] Alzubi JA, Jain R, Singh A, Parwekar P, Gupta M. COBERT: COVID-19 Question Answering System Using BERT. Arab J Sci Eng. 2021 Jun 23:1-11. doi:

10.1007/s13369-021-05810-5. Epub ahead of print. PMID: 34178569; PMCID: PMC8220121.

- [59] Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering (Izacard & Grave, EACL 2021)
- [60] H. -T. Zheng, J. -Y. Chen, Z. -Y. Fu, Z. -H. Xu and C. -Z. Zhao, Automatically Answering Questions With Nature Languages, 2018 5th International Conference on Systems and Informatics (ICSAI), Nanjing, China, 2018, pp. 350-355, doi: 10.1109/ICSAI.2018.8599337.
- [61] L. Chen, A Large-Language-Model driven Geographic Information QA system, 2023, <http://hdl.handle.net/2261/0002008196>
- [62] 山内洋輝, 梶原智之, 桂井麻里衣, 大向一輝, 二宮崇. 学術ドメインに特化した日本語事前訓練モデルの構築. 言語処理学会第 29 回年次大会, pp.2842-2846, March 2023.
- [63] <https://ai-health.m.u-tokyo.ac.jp/home/research/corpus>
- [64] Loshchilov, I. and Hutter, F. (2019) Decoupled Weight Decay Regularization. 7th International Conference on Learning Representations, New Orleans, 6-9 May 2019. <https://dblp.org/rec/conf/iclr/LoshchilovH19.html>
- [65] Shuntaro Yada, Yuta Nakamura, Shoko Wakamiya, Eiji Aramaki: Real-MedNLP: Overview of REAL document-based MEDical Natural Language Processing Task, In Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-16), pp. 285-296, 2022
- [66] 日本語 BERT モデル公開, 東北大学自然言語処理研究グループ, <https://www.nlp.ecei.tohoku.ac.jp/news-release/3284/>
- [67] 早大 RoBERTa, <https://huggingface.co/nlp-waseda/roberta-base-japanese>
- [68] GiNZA - Japanese NLP Library, <https://megagonlabs.github.io/ginza/>
- [69] Clark, Kevin, et al. "Electra: Pre-training text encoders as discriminators rather than generators." arXiv preprint arXiv:2003.10555 (2020).
- [70] F. Wang, et al., Knowledge Graph: Method, Practice and Application, 2019.