意味解析システム ccg2lambda におけるドメイン知識導入手法の試案

村上 夏輝（指導教員：戸次大介）

1 はじめに

医療分野には画像所見や電子カルテといった電子化されたテキストが多く存在している。これらのテキストを診療支援や臨床研究等で使用するには、自然言語処理が必須である [2]。

そこで、症例テキストを有効活用する方法として、時系列を考慮した症例検索システムの構築が試みられている [10]。この症例検索システムでは、事象が発生した順番といった時間関係を考慮し、該当する文を検索結果として出力することが期待される。このシステムを実現するためには、時間関係や数量表現を考慮した高度な意味解析が必要となる。

時間関係や数量表現を考慮した高度な意味解析を実現することに対し、症例テキストが検索文を含意するかどうか、という含意関係認識の問題として解くことが提案されている。また、含意関係認識の問題を解くために、意味解析システム ccg2lambda [3] が用いられている。本研究ではこの ccg2lambda において、知識を導入するモジュールを追加し、ドメイン知識を利用した論理推論を行うことを提案する。

2 ccg2lambda

ccg2lambda とは、組み合わせ範疇文法 (Combinatory Categorial Grammar, CCG)[8, 4] に基づく高度な構文解析と、高階論理に基づく自動推論システムを組み合わせた意味解析・推論システムである。

ccg2lambda に、前提文と仮説文を入力する。すると、統語解析が行われ、構文木が出力される。この構文木に対し、意味解析を行うことで論理式が導出される。そして、2 文の論理式の含意関係について、定理証明器 Coq [5] を用いて、論理推論を行う。この論理推論では、前提文が仮説文を含意しているときは yes という結果が返り、矛盾している時は no、含意も矛盾もしていないときは unknown という結果が返る。

論理推論を行う上で、語彙知識が必要となる場合がある。たとえば、前提文として、「検査の結果、太郎に肺癌が見つかった」という文と、仮説文として「検査の結果、太郎に肺悪性腫瘍が見つかった」という文を入力したとする。肺悪性腫瘍は肺癌を表しているため、結果としては、yes が出力されるのが正解である。しかし、論理推論を行う際に、肺癌が肺悪性腫瘍を含意するという知識が利用できないために、unknown が結果として出力されてしまう。

意味解析システム ccg2lambda に語彙知識を補完する先行研究として、外園ら [7] と竹内 [9] がある。外園らでは金融テキストに関する知識を補完し、竹内では同義・反義知識を補完して論理推論を行なっている。これらの研究では、論理推論に用いられる知識を、公理としてあらかじめ定理証明器に追加しておく手法をとっている。

しかし、本研究では、ccg2lambda において、意味解析後に、必要な語彙知識のみを辞書等から導入し、論理推論を行う手法を提案する。

必要な語彙知識のみを辞書等から導入する手法をとることにより、新しい知識が追加された場合に、辞書が更新され続けていれば、ccg2lambda の定理証明器の公理を更新する必要がなくなる。また、辞書を差し替えることで、論理推論に導入する知識を柔軟に変更することが可能となる。

3 提案手法

ccg2lambda において、証明できずに残った帰結中の項は型とともに出力される。語彙知識を含む辞書を用いて、未証明項について検索し、辞書から得た知識を公理として定理証明器に追加する。そして、新しく公理を追加した定理証明器を用いて、再度証明を行うという手法をとる。この手法を含めた ccg2lambda の全体像を図 1（次項）に示す。

3.1 公理を追加するモジュール

下記の流れに則ったモジュールを ccg2lambda に組み込むことで上記の提案手法を実現する。

証明できずに残った論理式について、統語解析を行った結果を参照する。下記に統語解析結果の一部を示す。統語解析結果には、各トークンの品詞の情報が含まれている。品詞が一般名詞であるトークンを抜き出すことで、病名になり得るトークンのみを抜き出したこととなる。抜き出したトークンの表層形（属性 surf の値）を辞書で検索し、辞書に検索結果があった表層形のみ、公理を定理証明器に追加する。

ソースコード 1: 統語解析結果

```
1 <token reading="タロー" base="太郎"
      inflectionForm="*" inflectionType="*"
      pos3="名" pos2="人名" pos1="固有名詞"
      pos="名詞" surf="太郎" id="s0_4"/>
2 <token reading="ニ" base="に" inflectionForm
      ="*" inflectionType="*" pos3="*" pos2="
      一般" pos1="格助詞" pos="助詞" surf="に"
       id="s0_5"/>
3 <token reading="ハイガン" base="肺癌"
      inflectionForm="*" inflectionType="*"
      pos3="*" pos2="*" pos1="一般" pos="名詞"
       surf="肺癌" id="s0_6"/>
```

3.2 万病辞書

本研究では、語彙知識を補完するために医学用語辞典である万病辞書 [6] を使用する。万病辞書とは、医療従事者が記載した経過記録や退院サマリから、症状や病名に関連する用語を広く抽出したデータである。病名の正式名称だけでなく、略記や英語名を含めた 362,866 件の病名用語が収録されている。

3.3 Coq の公理

補完する知識が「肺癌 ⇒ 肺悪性腫瘍」である際には、Coq に追加する公理を下記のように書く。

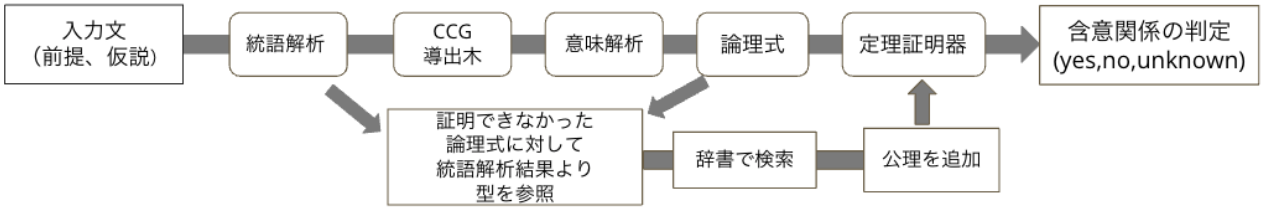

図 1: 提案手法を含めた cccg2lambda の全体像

```
Axiom a1 : for all x,
                _肺癌 x -> _肺悪性腫瘍 x.
Hint resolve a1.
```

他の語彙知識において、含意関係があるという内容を表す際には、肺癌のところに、万病辞書に記載されている標準病名の内容を、肺悪性腫瘍のところには症例テキストでの出現形を書くことで、含意関係を表すことができる。この公理を定理証明器に追加することで、論理推論で含意関係があるという推論結果を導出可能となる。

4 データセット

評価実験に使用するデータセットを作成した。J-STAGE で公開されている症例報告論文 PDF から生成された症例報告コーパスである J-Med-Std-CR[1]224 件を使用する。症例に含まれる文の中から、万病辞書に記載のある病名が一番最初に出てきた文、もしくはまとめの文を 1 文使用した。文から複雑な情報を抜き、文を短くした上で、該当する病名の欄を標準病名に置き換えた文を前提文、出現形に置き換えた文を仮説文とし、入力文のペアを 224 ペア作成した。構築したシステムに対して、このデータセットを用い、評価実験を行う。

5 現状の課題

現状の ccg2lambda では、トークンが一つであるはずの単語が、複数のトークンに分かれてしまっているという課題がある。以下に例を示す。

(1)
```
Parameter _肺悪性腫瘍 :
                        Entity -> Prop.
```

(2)
```
Parameter _悪性 : Entity -> Prop.
Parameter _肺  : Entity -> Prop.
Parameter _腫瘍 : Entity -> Prop.
```

例えば、肺悪性腫瘍という単語があった場合、本来は肺悪性腫瘍が一つの項として (1) のような述語であることが期待される。しかし、現状の ccg2lambda では (2) のように肺と悪性と腫瘍の三つのトークンに分かれた述語として出力されてしまう。

この課題に対して、石田ら [1] に倣い、構文解析結果から分かれてしまった複合語を取り出し、石田らで使用されている複合語解析モジュールを応用し、複合語箇所が修正された構文木を得る手法をとる。

[1]https://sociocom.naist.jp/medtxt/cr/

6 おわりに

本稿では、症例検索システムの構築に向け、ccg2lambda にドメイン知識を導入する手法を提案した。手法は、含意関係認識の問題において、語彙知識が必要な問題を解く際に、意味解析後に、必要な知識のみを辞書から導入する流れである。

システムを構築したのちに、作成したデータセットを用いて評価実験を行う予定である。

謝辞 本研究の一部は、政策科学総合研究事業（臨床研究等 ICT 基盤構築・人工知能実装研究事業）21AC5001 の支援を受けたものである。

参考文献

[1] Mana Ishida, Hitomi Yanaka, and Daisuke Bekki. Compositional semantics for compound words in medical case retrieval. *Proceedings of the 18th International Workshop on Logic and Engineering of Natural Language Semantics (LENLS18)*, 11 2021.

[2] Yasushi Matsumura, Kentaro Torisawa, Emiko Shinohara, Takahiro Suzuki, and Eiji Aramaki. (dream) frontier and promising application of natural language processing. *Japan Journal of Medical Informatics*, Vol. 38, No. 1, pp. 41–46, 2018.

[3] Koji Mineshima, Pascual Martínez-Gómez, Yusuke Miyao, and Daisuke Bekki. Higher-order logical inference with compositional semantics. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2055–2061, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[4] Mark Steedman. *The Syntactic Process*. MIT Press, 2000.

[5] The Coq Development Team. *The Coq Proof Assistant: Reference Manual: Version 8.9.0*. INRIA, 2019.

[6] 荒牧英治, 若宮翔子, 河添悦昌. カルテ情報の自動構造化システムと疾患数理モデルの逐次的構築, 及び, 自動構造化機能を有した入力機構の開発. 医療情報学, Vol. 40, No. 1, pp. 12–13, 2020.

[7] 外園康智, 長谷川貴博, 渡邉知樹, 馬目華奈, 簗有紀子, 谷中瞳, 田中リベカ, 峯島宏次, 戸次大介ほか. 意味解析システム ccg2lambda による金融ドキュメント処理. 人工知能学会全国大会論文集 第 32 回全国大会 (2018), 3G1-05. 一般社団法人 人工知能学会, 2018.

[8] 戸次大介. 日本語文法の形式理論. くろしお出版, 東京, 2010.

[9] 竹内至生. 同義反義知識と高階論理推論に基づく因果関係グラフの精錬. 学士論文, 京都大学, 2019.

[10] 石田真捺, 谷中瞳, 馬目華奈, 戸次大介. 論理推論による症例検索に向けた日本語症例テキストの複合語解析の試案. 人工知能学会全国大会論文集 第 35 回全国大会 (2021), 4J3GS6f05. 一般社団法人 人工知能学会, 2021.