Knowledge Injection for Disease Names in Logical Inference between Japanese Clinical Texts

Natsuki Murakami[1] **Mana Ishida**[1] **Yuta Takahashi**[1]
Hitomi Yanaka[2] **Daisuke Bekki**[1]
[1] Ochanomizu University
[2] the University of Tokyo
{murakami.natsuki,ishida.mana,takahashi.yuta,bekki}@is.ocha.ac.jp
hyanaka@is.s.u-tokyo.ac.jp

Abstract

In the medical field, there are many clinical texts such as electronic medical records, and research on Japanese natural language processing using these texts has been conducted. One such research involves Recognizing Textual Entailment (RTE) in clinical texts using a semantic analysis and logical inference system, ccg2lambda. However, it is difficult for existing inference systems to correctly determine the entailment relations, if the input sentence contains medical domain specific paraphrases such as disease names. In this study, we propose a method to supplement the equivalence relations of disease names as axioms by identifying candidates for paraphrases that lack in theorem proving. Candidates of paraphrases are identified by using a model for the NER task for disease names and a disease name dictionary. We also construct an inference test set that requires knowledge injection of disease names and evaluate our inference system. Experiments showed that our inference system was able to correctly infer for 106 out of 149 inference test sets.

1 Introduction

In the medical field, there are many electronic texts, such as image detections and electronic medical records, and using such texts becomes more active in research on natural language processing (NLP) in Japanese (Aramaki et al., 2018; Doi et al., 2011). However, many of these studies utilize machine learning approaches, and it is argued that the machine learning approaches have problems in dealing with challenging linguistic phenomena such as negation and quantification. Logic-based approaches have been proposed to perform systematic inferences involving these challenging linguistic phenomena, and one task of these inferences can be referred to as recognizing textual entailment (RTE). RTE is the task of determining whether a hypothesis sentence H can be inferred from a premise sentence T. For example, the following example illustrates a case where T entails H.

T : Some patients are given Loxoprofen.

H : There are patients who are given headache medicines.

One such effort at RTE between clinical texts is the logical inference system in the medical domain proposed by Ishida et al. (2022). Compounds words are often appeared in Japanese clinical texts, and methods to analyze linguistic phenomena in compound words were desirable. The system is an inference system based on *ccg2lambda* (Martínez-Gómez et al., 2016), a semantic analysis and logical inference system: the system extends ccg2lambda to enable an analysis of compound words that are frequently appeared in clinical texts. However, this system fails to perform inference when there are differences in the notation of disease names in medical texts. For example, the disease name "Deep vein thrombosis" has multiple paraphrases, such as "DVT" and "Homann's sign", and different clinical texts use different phrases that refer to the same disease name. The premise sentence is "The patient developed Homann's sign." and the hypothesis sentence is "The patient developed deep vein thrombosis.", then empirically the premise sentence implies the hypothesis sentence. To show this entailment relation, the knowledge that "Homann's sign means deep vein thrombosis" must be supplemented in the theorem prover.

In this study, we propose our logical inference system with knowledge injection in the medical field. We identify candidates of paraphrase of disease names that are necessary for theorem proving by a named entity recognition (NER) model for disease names and the Japanese disease name dictionary called J-MeDic (Ito et al., 2018). By generating axioms according to the combination of compound word semantic tags assigned to the identified

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 108–117

Surface form	ICD10	Standard disease name	Reliability	Frequency
sinbu-jomyaku-kessen-syo deep-vein-thrombosis 深部静脈血栓症	I802	deep vein thrombosis	S	85-90%
kasi-sinbu-jomyaku-kessen lower extremity-deep-vein-thrombosis 下肢深部静脈血栓症	I802	deep vein thrombosis	A	5-10%
DVT DVT	I802	deep vein thrombosis	C	90-95%
DVT-tyoukou DVT-sign DVT徴候	I802	deep vein thrombosis	C	60-65%
homanzu-tyoukou Homann's-sign ホーマンズ徴候	I802	deep vein thrombosis	C	25-30%

Table 1: An example of the J-MeDic with the standard disease name column listing "Deep Vein Thrombosis."

disease names, we inject paraphrase knowledge of disease names as axioms. We also evaluate the effectiveness of our inference system by constructing an inference test set that requires knowledge injection of disease names.

2 Background

2.1 Inference systems for clinical texts

There has been growing progress in research on neural network models for RTE with large-scale datasets using crowdsourcing such as SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018). However, neural network models for RTE generally require a large amount of training data, and inference is black-box, which makes it difficult to correct errors. Therefore, there is a need to develop RTE systems that are applicable to fields such as medical field, where it is difficult to prepare a large amount of training data and transparency is necessary.

The logical inference system ccg2lambda has the advantage of being able to judge entailment relations between sentences without using a large amount of training data and being easy to personalize and modify processing. However, the original ccg2lambda fails to analyze the semantic relations within compound words because they are treated as one word.

Ishida et al. (2022) addressed this problem by adding a compound word analysis module to ccg2lambda. This module extracts compound words from the Combinatory Categorial Grammar (CCG) (Steedman, 2000; Bekki, 2010) syntactic structures obtained by the CCG parser of ccg2lambda and assigns compound word semantic tags that represent semantic relations within compound words, using a compound word semantic tagger. Based on syntactic structures, semantic tags, and lambda calculus, the semantic representation was derived by taking into account the semantic relations within the compound words, and inference between clinical texts containing compound words was realized.

2.2 Related studies on axiom injection

As for related studies on axiom injection of logical inference systems, including ccg2lambda, Martínez-Gómez et al. (2017) proposed word axiom injection by using lexical knowledge. Hokazono et al. (2018) used this word abduction mechanism to inject word knowledge specific in the financial texts as the lexical knowledge. However, these previous studies were limited to handle word-to-word relations in natural deduction proofs. Yanaka et al. (2018) proposed a method for paraphrase detection by natural deduction proofs of semantic relations between sentence pairs to complement phrasal knowledge. In this study, we propose how to detect phrasal knowledge of disease names necessary for proving entailment relations between clinical texts and inject the knowledge into logical inference.

2.3 J-MeDic

In this study, we use J-MeDic to inject disease name knowledge into logical inference. J-MeDic is a Japanese dataset that extensively extracts

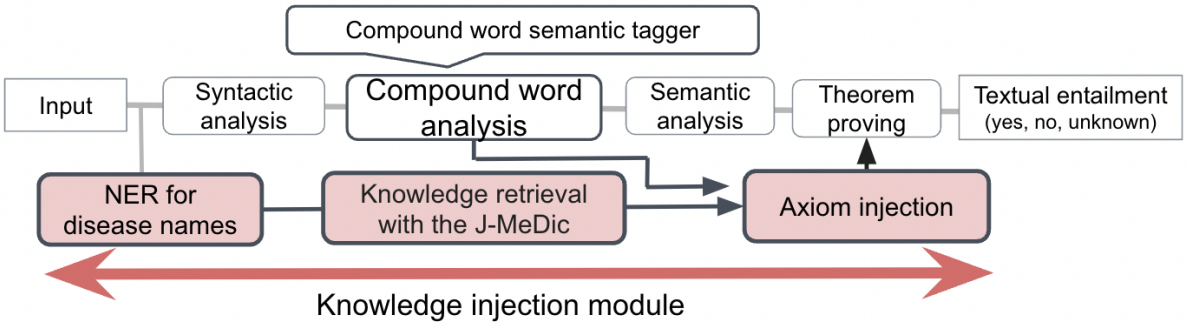

Figure 1: The overview of the proposed system.

words related to symptoms and disease names from progress records and discharge summaries recorded in electronic medical records by medical professionals. The dataset contains not only the formal names of diseases but also abbreviations and English names. The dataset covers 362,866 disease names in total. Table 1 provides an example where the standard disease name column in J-MeDic is "Deep vein thrombosis." J-MeDic records include the surface forms of the disease name, their pronunciation, ICD-10 codes, standard disease names, levels of reliability, and levels of frequency.

2.4 Related studies on NER in the medical domain

There are related studies on NER for disease names, such as the work by Goino and Hamagami (2021) and MedNER (Nishiyama et al., 2022). In Goino-Hamagami's work, NER and modality estimation of medical conditions such as disease names and symptoms along with their modalities (five types) were performed using BERT and CRF. Experiments were conducted using three BERT models released by Tohoku University (tohoku-charBERT, tohoku-BERT, tohoku-wwm-BERT) and UTH-BERT[1], a Japanese BERT model pretrained on medical documents by the University of Tokyo. MedNER[2] is a tool for the NER for disease names using word embeddings of BERT. MedNER follows J-MeDic and also performs modality estimation.

3 System Overview

The overview of our proposed system is shown in Figure 1. We provide our inference system by extending Ishida et al. (2022)'s inference system. The system consists of syntactic analysis, compound word analysis, semantic analysis, and theorem proving. In this study, we add a knowledge injection module in the medical domain to the previous system.

In the knowledge injection module, we first apply an NER model based on a pretrained language model to extract disease names from the input sentence. For the extracted disease names, we perform an exact match search for the surface form column of disease names in J-MeDic. If the disease name and the surface form match, we inject the knowledge about the disease name in J-MeDic as an axiom to the automated theorem prover Coq (Bertot and Castéran, 2013). Since the additional axioms vary depending on the semantic relations between morphemes, we check the semantic tags of compound words assigned by the compound word semantic tagger, derive the axiom using the semantic tag and the knowledge of J-MeDic, and then inject the knowledge. We describe the details to provide an NER model in Section 4 and Subsection 6.1 and the details of axiom injection in Section 5.

4 Building an NER dataset for disease names

To train the model for the NER task for disease names in our inference system, we constructed a dataset newly for NER using clinical texts. We use a corpus of case reports, J-MedStd-CR[3], which was extracted through OCR from case report papers in PDF format that are openly accessible on J-Stage. In this study, we manually annotated the appearances of disease names from J-MeDic in the 2,626 sentences of the J-MedStd-CR corpus.

[1]https://ai-health.m.u-tokyo.ac.jp/home/research/uth-bert
[2]https://github.com/sociocom/MedNER-J
[3]https://sociocom.naist.jp/medtxt/cr/

Tag	Type	Example
EN	Entity	DVT (Deep vein thrombosis)
M_EN	Modifying words	ryuukisei-byouhen elevated-lesion 隆起性 病変 (Elevated lesions)
PA	Body part	nou-kousoku brain-infarction 脳 梗塞 (Brain infarction)
GA	Nominative	kyousui-tyoryuu pleural effusion-retention 胸水 貯留 (Pleural effusion)
WO	Accusative	kotuzui-yokusei bone marrow-suppression 骨髄 抑制 (Myelosuppression)
NI	Dative	kotu-ten'i bone-metastasis 骨 転移 (Bone metastasis)
EV	Event	kyousui-tyoryuu pleural effusion-suppression 胸水 貯留 (Pleural effusion)

Table 2: Examples of semantic tags. Underlined parts correspond to each tags.

5 Axiom Injection

In the proposed method, after identifying candidate paraphrases of the disease names, the knowledge of paraphrases is injected as axioms. In axiom injection, it is necessary to generate the knowledge about paraphrases of disease names as axioms according to the semantic representations in ccg2lambda. For example, when the premise sentence is (1) and the hypothesis sentence is (2), the underlined disease name's semantic tag in (1) is "EN" because "PE" is one word and entity. The semantic representation in this case is (1a). Semantic tags for the disease name underlined in (2) are "PA EN" because "pulmonary" in " pulmenary embolism" is a body part and "embolism" is an entity. The semantic representation is (2a). $\mathsf{PartOf}(e_1, e_2)$ indicates $\mathsf{hai}(e_2)$ is a body part for $\mathsf{sokusensyo}(e_1)$. As in (1a) and (2a), semantic representations differ depending on the semantic relations between morphemes within the compound words. In this study, we realize axiom injection by defining axioms generated through the combination of semantic tag assigned to both the surface form and standard disease names.

Combination	Number	Example
EN	276	syuryuu 腫瘤 (Neoplasm)
M_EN EN	145	ryuukisei-byouhen elevated-lesion 隆起性 病変 (Elevated lesions)
PA EN	45	nou-kousoku brain-infarction (Brain infarction) 脳 梗塞
M_EN M_EN EN	22	genpatusei-tanzyusei- primary-biliary- kankouhen liver cirrhosis 原発性 胆汁性 肝硬変 (Primary biliary cholangitis)
M_EN PA EN	22	ten'isei-kan-syuyou metastatic-liver-neoplasm 転移性 肝 腫瘍 (Metastatic liver tumor)
Others	141	kyousui-choryuu pleural effusion-retention 転移性 肝 腫瘍 GA EV (Metastatic liver tumor)

Table 3: Combinations of semantic tags in disease names.

(1) PE ga zouaku-si-teita
PE NOM worsen-EUPH-PST
PEが増悪していた。
PE worsened.

a. $\exists e_1(\mathsf{PE}(e_1))$

(2) hai-sokusensyo ga zouaku-si-teita
pulmonary-embolism NOM worsen-EUPH-PST
肺塞栓症が増悪していた。
Pulmonary embolism worsend.

a. $\exists e_2(\mathsf{hai}(e_2) \wedge \exists e_1(\mathsf{sokusensyo}(e_1) \wedge \mathsf{PartOf}(e_1, e_2)))$

5.1 Trends in combinations of semantic tags in disease names

Table 2 shows some of the semantic tag assigned by Ishida et al. (2022)'s compound word semantic tagger. We investigated the composition of semantic tag to generate axioms based on the combination of semantic tag assigned to disease names. We applied the compound word semantic tagger to assign semantic tag to 651 disease names and their standard disease names extracted from randomly selected sentences containing disease names in clinical texts of the J-MedStd-CR corpus. The compound word semantic tagger is based on BiLSTM and BERT models. BERT model is the model released by

Axiom			Exapmles of disease name			Semantic tags of disease name		
Surface form	⇒	Disease name	Surface form	⇒	Disease name	Surface form	⇒	Disease name
∗	⇒	∗	hokou-hunou walking-impossibility 歩行 不能 (Walking difficulty)	⇒	hokou-konnan walking-difficulty 歩行 困難	EV EN	⇒	EV EN
M_EN ∗	⇒	∗	mansei-gisei-tyouheisokusuou chronic-pseudo-intestinal obstruction 慢性 偽性 腸閉塞症 (Pseudoileus)	⇒	gisei-ireusu pseudo-ileus 偽性 イレウス	M_EN M_EN EN	⇒	M_EN EN
PA ∗	⇒	∗	nou-kekkan'en brain-vasculitis 脳 血管炎 (Cerebral arteritis)	⇒	noudoumyakuen brain arteritis 脳動脈炎	PA EN	⇒	EN
PA ∗	⇒	M_EN ∗	ketsuryu-syougai bloodstream-disorder 血流 障害 (Peripheral circulatory disturbance)	⇒	massyojunkan-syogai peripheral circulation-disorder 末梢循環 障害	PA EN	⇒	M_EN EN
M_EN ∗	⇒	PA ∗	bimansei-shikiso-tintyaku diffuse-pigment-deposition びまん性 色素 沈着 (Skin pigmentation)	⇒	hihu-sikiso-tintyaku skin-pigment-deposition 皮膚 色素 沈着	M_EN GA EV	⇒	PA GA EV
GA EV	⇒	WO EV	kettin-kousin blood sedimentation-accentuation 血沈 亢進 (Erythrocyte sedimentation rate acceleration)	⇒	sekitin-kousin erythrocyte sedimentation-accentuation 赤沈 亢進	GA EV	⇒	WO EV
EN	⇒	M_EN+ EN	PBC (Primary biliary cholangitis)	⇒	genpatusei-tanzyuusei-kankouhen primary-biliary-liver cirrhosis 原発性 胆汁性 肝硬変	EN	⇒	M_EN M_EN EN
EN	⇒	PA EN	PE (Pulmonary embolism)	⇒	hai-sokusensyo lung-embolism 肺 塞栓症	EN	⇒	PA EN
EV	⇒	EN	kansen'ika become liver fibrosis 肝繊維化 (Liver fibrosis)	⇒	kansen'isuu liver fibrosis 肝繊維腫	EV	⇒	EN

Table 4: Combination examples of semantic tags in axiom injection. ∗ indicates that the combination of tags assigned to the surface form and the standard disease name are the same. M_EN+ indicates that the M_EN tag appears one or more times. ⇒ indicates entailment relations.

Tohoku University[4] that was trained on Japanese Wikipedia data, and the tokenizers are MeCab and WordPiece. The top 5 combinations of semantic tags assigned to disease names are shown in Table 3. We perform axiom injection according to these combinations.

5.2 Axiom injection based on semantic tags

Table 4 shows the combination examples of semantic tag for the disease names in axiom injection. The asterisk ∗ indicates that the combinations of semantic tags are the same for both the surface form and the standard disease name. For example, the first row "∗ ⇒ ∗" indicates that the same semantic tags are assigned, such as "EV EN ⇒ EV EN" for "walking-impossibility ⇒ walking-difficulty". Similarly, the second row "M_EN ∗ ⇒ ∗" indicates that the semantic tags "M_EN EN" are assigned to the phrase "pseudo-intestinal obstruction" included in "chronic-pseudo-intestinal obstruction", and they match the tags assigned to "pseudo-ileus". "M_EN+" indicates that there are one or more "M_EN" tag present. As in the example where the semantic tags "EN ⇒ M_EN M_EN EN" are assigned to "PBC ⇒ primary-biliary-liver cirrhosis", even when there are multiple repetitions of the M_EN tag in the semantic tags assigned to the standard disease name, it is possible to generate an axiom. As an example of the generated axioms, we show the axiom to be injected when the premise sentence is (1) and the hypothesis sentence is (2) in Figure 2. The generated axioms are injected to the theorem prover as callable axioms during automated theorem proving and used for inference.

```
Parameter _PE : Entity -> Prop.
Parameter _塞栓症 : Entity -> Prop.
Parameter _肺 : Entity -> Prop.
Axiom e2pa:forall (x:Entity),
    _PE x -> _肺 x.
Axiom e2pa2:forall (x:Entity),
    _PE x -> _塞栓症 x.
Axiom e2pa3:forall (x:Entity),
    _PE x ->PartOf x x.
Hint Resolve e2pa e2pa2 e2pa3.
```

Figure 2: An example of the axiom to be injected.

[4]https://huggingface.co/cl-tohoku/bert-base-japanese

Model	Pretraining corpus	Tokenizer
Japanese RoBERTa base (RIKEN)	Wikipedia(Japanese)	MeCab + BPE
japanese-roberta-base (RINNA)	Wikipedia + CC-100 (Japanese)	Juman++ + sentencepiece
roberta-large-japanese (Waseda)	Wikipedia + CC-100 (Japanese)	Juman++ + sentencepiece

Table 5: RoBERTa models used in the experiments.

Model	RIKEN model	RINNA model	Waseda model	MedNER
Accuracy	97.2%	96.4%	97.0%	94.3%
Precision	83.4%	79.4%	76.3%	71.4%
Recall	82.3%	78.0%	77.2%	66.6%
F1-score	81.5%	77.3%	74.6%	66.1%

Table 6: Experimental results for NER of disease names.

6 Experiments

6.1 Experiments on the NER task of disease names

6.1.1 Experimental settings

To select a best performance model for the NER task for disease names to be combined with our inference system, we evaluate three RoBERTa models shown in Table 5. The Japanese RoBERTa base[5] (hereafter referred to as the RIKEN model), which is publicly available from RIKEN, was pretrained only on Japanese Wikipedia. The tokenizer uses MeCab (Kudo et al., 2004) for word segmentation and BPE (Sennrich et al., 2016) for subword segmentation. The Japanese RoBERTa base model publicly available from RINNAI (Zhao and Sawada, 2021) (hereafter referred to as the RINNA model), and the roberta-large-japanese model released by Waseda University[6] (hereafter referred to as Waseda model), were pretrained on both Japanese Wikipedia and the Japanese portion of the CC-100 dataset. Both models use Juman++ (Tolmachev et al., 2018) as a tokenizer for word segmentation and SentencePiece (Kudo and Richardson, 2018) for a subword segmentation.

For the NER task, we used 2,303 sentences from J-MedStd clinical texts as training data. We used 85% of the data as development data and 15% as validation data to finetune the pretrained language models. The training data contains 2,551 appearances of disease names. We randomly selected 326 sentences from J-MedStd clinical texts for evaluation. The 326 sentences that are used for evaluation do not overlap with the training data.

6.1.2 Evaluation

We trained and evaluated three BERT models, shown in Table 5, using our NER dataset in the experiment. We also performed a comparison with MedNER using our NER dataset and the experimental results are shown in Table 6. The RIKEN model had the highest score in terms of f1-score for predicting disease names. In the RINNA model, even unrelated text around disease names were extracted. The Waseda model had some disease names that were split in the middle of the name. MedNER tended to extract disease names that are closely related, such as "necrotic" and "granuloma" for "granuloma with necrosis", and "pain" and "pruritus" for "pain and pruritus". F1-score of MedNER decreased because the evaluation dataset created in this study was annotated with one disease name per annotation.

Based on the results of this experiment, we adopted the RIKEN model, which showed the highest performance, as the model for the NER task to be combined with our inference system.

6.2 Experiments on the RTE task

To evaluate the effectiveness of our inference system, we performed a comparison between our inference system and the previous inference system (Ishida et al., 2022).

[5]https://huggingface.co/liat-nakayama/japanese-roberta-base-20220905

[6]https://huggingface.co/nlp-waseda/roberta-large-japanese

Premise		Hypothesis
byouri-sindan ha **t u b 1** datta pathological diagnosis NOM tub1 be-PST 病理診断は**t u b 1**であった。 (The pathological diagnosis was **tub1**.)	⇒	byouri-sindan ha **gan** datta pathological diagnosis NOM canser be-PST 病理診断は**癌**であった。 (The pathological diagnosis was **cancer**.)
V A P ni yuukou de-aru **VAP** DAT valid be-PRS **VAP**に有効である。 (It is effective for **VAP**.)	⇒	**jinkou-kokyuuki-haienn** ni yuukou de-aru **mechanical-ventilator-pneumonia** DAT valid be-PRS 人工呼吸器肺炎に有効である。 (It is effective for **Ventilator-associated pneumonia**.)
kanja ha **homanzu-tyoukou**-yousei datta patient NUM **Homann's-sign**-positive be-PST 患者は**ホーマンズ徴候**陽性だった。 (The patient was positive for **Homann's sign**.)	⇒	kanja ha **sinbu-jomyaku-kessensyo**-yousei datta patient NUM **deep vein thrombosis**-positive be-PST 患者は**深部静脈血栓症**陽性だった。 (The patient was positive for **deep vein thrombosis**)

Table 7: Examples of our inference test set. The bolded part indicates disease name knowledge that is necessary for inference.⇒ indicates entailment relations.

6.2.1 Inference test set involving disease names

We constructed an inference test set that requires disease name knowledge injection and evaluated our proposed system. The inference test set is constructed to consist of sentence pairs whose relation is entailment and the experiment is conducted to test whether the system can correctly predict entailment relations. Table 7 shows examples of the inference tests set. For constructing the test set, we used sentences from J-MedStd-CR, a corpus of clinical case reports where the disease names mentioned in the sentences are different from the standard disease names in J-MeDic. We manually constructed a set of 149 pairs of simplified hypothesis sentences and corresponding premise sentences, where the hypothesis sentences were simplified versions of sentences containing disease names in the J-MedStd-CR corpus, and the disease names in the hypothesis sentences were replaced with their corresponding standard names.

6.2.2 Evaluation

We compared the accuracy between our inference system with the knowledge injection module and the previous system by Ishida et al. (2022) on our inference test set. Table 8 shows the results of the evaluation of inference. While the previous system failed to predict entailment relations for all examples, our system was able to make correct predictions for 106 out of 149 test sets.

Inference system	Accuracy
Ishida et al. (2022)	0/149 (0.0%)
Our system	106/149 (71.1%)

Table 8: Results on the RTE task.

6.2.3 Error analysis

We performed an error analysis on the cases where our inference system made incorrect predictions. Table 9 shows examples of error types and sentence pairs for the analysis of errors. There were many cases where the disease names written in English were not correctly extracted due to errors in NER. For errors related to syntactic analysis, the morphemes "itching" in (3) and "necrosis" in (4) were misclassified as verbs by the morphological analyzer, janome[7], when they should have been treated as nouns. These morphemes were treated as verbs such as "itch" and "become necrosis", and the wrong axioms were provided, which resulted in the failure of inference.

(3) sou-you
scraching-itching
そう_痒
(pruritus)

(4) kan-saibou-esi
liver-cell-necrosis
肝_細胞_壊死
(hepatic necrosis)

For the error caused by compound word analysis, the compound word semantic tagger by Ishida et al. (2022) classified the "cancer" in (5) as "PA" instead of being tagged with "EN", which resulted in a failure of inference. As a result, axiom injection could not be performed correctly.

(5) nyoukan-nyourozyouhi-gan
ureter-urothelium-cancer
尿管_尿路上皮_癌
(urothelial cancer)

Regarding errors due to syntactic analysis, an example is shown in Figure 3. Since "limbs pain" is a

[7]https://github.com/mocobeta/janome

Type	Number	Example
NER error	25	kanseiken ha AIH-you no syoken o teisi-teita liver-biopsy NOM AIH-like GEN finding ACC present-PST 肝生検ではＡＩＨ様の所見を呈していた。 kanseiken ha jikomen'ekisei-kan'en-you no syoken o teisi-teita Liver-biopsy NOM autoimmune-hepatitis–like GEN finding ACC present-PST ⇒ 肝生検では自己免疫性肝炎様の所見を呈していた。 (Liver biopsy showed findings suggestive of AIH ⇒ Liver biopsy showed findings suggestive of autoimmune hepatitis.)
Syntactic analysis error	14	sou-you-kan ga at-ta　　sou-you ga at-ta scratching-itching-sensation NOM be-PST　scratching-itching NOM be-PST 掻痒感があった。　⇒　そう痒があった。 (The patient had itch sensation. ⇒ The patient had pruritus.)
CW analysis error	3	kanja ha nyourozyouhi-gan de-atta　kanja ha nyoukan-nyourozyouhi-gan de-atta patient NOM urothelium-cancer be-PAST　patient NOM ureter-urothelium-cancer be-PAST 患者は尿路上皮癌であった。　⇒　患者は尿管尿路上皮癌であった。 (The patient had urothelial cancer.⇒ The patient had ureteral urothelial cancer.)
Semantic analysis error	1	sisi-toutuu ga zouakusi-ta　sisi-tuu ga zouakusi-ta limbs-pain NOM worsen-PST　limbs-pain NOM worsen-PST 四肢疼痛が増悪した。　⇒　四肢痛が増悪した。 (Limbs pain worsened. ⇒ Limbs pain worsened.)

Table 9: Types of errors and examples of sentence pairs.⇒ indicates entailment relations..

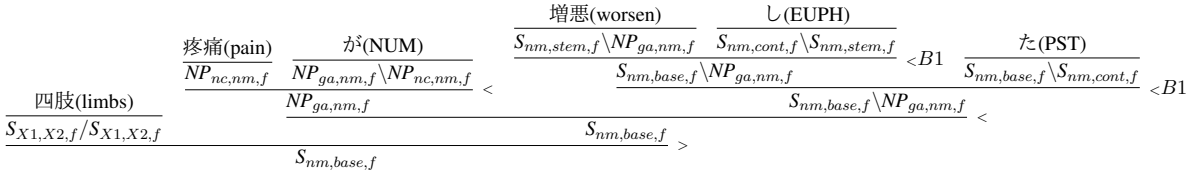

Figure 3: A semantic analysis result of “四肢疼痛が増悪した(Limbs pain worsened)”.

disease name, “limbs” and “pain” need to be combined first. However, according to the result of the syntax analysis, “pain worsened” was combined first, and then “limbs” was combined afterwards. This illustrates a case where the CCG syntactic structure for the disease name was not constructed correctly, leading to a failure to perform correct inference.

7 Conclusion

In this study, to flexibly perform inference involving knowledge for disease names, we extended the previous semantic analysis and logical inference system in the medical domain (Ishida et al., 2022). Specifically, we developed the knowledge injection module for the logical inference system, performing NER for disease names, searching for relevant knowledge using J-MeDic, and adding the resulting axioms to the theorem prover.

We also constructed a dataset for the NER task of disease names and an inference test set that requires knowledge injection of disease names. We evaluated our inference system using the constructed test set and as a result, we were able to perform correct inference for 106 out of 149 inference test cases. The future challenges are to expand the NER dataset and the inference test set, and to improve the knowledge injection module to further enhance the performance of the inference system. Furthermore, a comparison will be performed between the neural models trained on the expanded medical inference test set and the proposed method.

Acknowledgements

We thank the three anonymous reviewers for their helpful comments and suggestions, which improved this paper. This work was supported by JST, PRESTO grant number JPMJPR21C8, Japan.

References

Eiji Aramaki, Tomohide Iwao, Shoko Wakamiya, Kaoru Ito, Ken Yano, and Kazuhiko Ohe. 2018. A fundamental study on user utilization based on a trial operation of the medical case retrieval system. Japan Journal of Medical Informatics, 38(4):245–256 (in Japanese).

Daisuke Bekki. 2010. A Formal Theory of Japanese Grammar: The Conjugation System, Syntactic Structures, and Semantic Composition. Kuroshio. (In Japanese).

Yves Bertot and Pierre Castéran. 2013. Interactive Theorem Proving and Program Development: Coq'Art: The Calculus of Inductive Constructions. Springer Science & Business Media.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Shunsuke Doi, Takashi Kimura, Masaki Sekine, Takahiro Suzuki, Katuhiko Takabayashi, and Toshiyo Tamura. 2011. Management and evaluation of similar case searching system in homepage of medical society. Medical and Biological Engineering, 49(6):870–876 (in Japanese).

Takuya Goino and Tomoki Hamagami. 2021. Named entity recognition from medical documents by fine-tuning bert. In The 48th Intelligent Systems Symposium (in Japanese).

Yasunori Hokazono, Takahiro Hasegawa, Tomoki Watanabe, Kana Manome, Yukiko Yana, Hitomi Yanaka, Ribeka Tanaka, Koji Mineshima, Daisuke Bekki, et al. 2018. Semantic parsing in ccg2lambda and its application to financial document processing. In Proceedings of the 32th Annual Conference of JSAI, pages 3G105–3G105. The Japanese Society for Artificial Intelligence (in Japanese).

Mana Ishida, Hitomi Yanaka, and Daisuke Bekki. 2022. Compound words analysis and inferences of japanese clinical texts. In Proceedings of the 36th Annual Conference of JSAI, pages 1J4OS13a05–1J4OS13a05. The Japanese Society for Artificial Intelligence (in Japanese).

Kaoru Ito, Hiroyuki Nagai, Taro Okahisa, Shoko Wakamiya, Tomohide Iwao, and Eiji Aramaki. 2018. J-MeDic: A Japanese disease name dictionary based on real clinical usage. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to japanese morphological analysis. In Proceedings of the 2004 conference on empirical methods in natural language processing, pages 230–237.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Pascual Martínez-Gómez, Koji Mineshima, Yusuke Miyao, and Daisuke Bekki. 2016. ccg2lambda: A Compositional Semantics System. In Proceedings of ACL 2016 System Demonstrations, pages 85–90.

Pascual Martínez-Gómez, Koji Mineshima, Yusuke Miyao, and Daisuke Bekki. 2017. On-demand injection of lexical knowledge for recognising textual entailment. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 710–720, Valencia, Spain. Association for Computational Linguistics.

Tomohiro Nishiyama, Mihiro Nishidani, Aki Ando, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2022. NAISTSOC at the NTCIR-16 Real-MedNLP Task. In Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725. Association for Computational Linguistics.

Mark J. Steedman. 2000. The Syntactic Process. The MIT Press.

Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. 2018. Juman++: A morphological analysis toolkit for scriptio continua. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 54–59. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Hitomi Yanaka and Koji Mineshima. 2022. Compositional evaluation on Japanese textual entailment and similarity. Transactions of the Association for Computational Linguistics, 10:1266–1284.

Hitomi Yanaka, Koji Mineshima, Pascual Martínez-Gómez, and Daisuke Bekki. 2018. Acquisition of phrase correspondences using natural deduction proofs. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 756–766, New Orleans, Louisiana. Association for Computational Linguistics.

Takumi Yoshikoshi, Daisuke Kawahara, and Sadao Kurohashi. 2020. Multilingualization of a natural language inference dataset using machine translation. Technical Report 6, Kyoto University, Kyoto University / Presently with Waseda University, Kyoto University.

Tianyu Zhao and Kei Sawada. 2021. Release of pre-trained models for japanese natural language processing. JSAI Special Interest Group on Spoken Language Understanding and Dialogue Processing, 93:169–170 (in Japanese).

Appendix 1. Results of additional experiment

As an additional experiment, a comparison was made between Japanese BERT trained on a standard Japanese RTE dataset, not medical domain texts. The RTE dataset utilized for the experiment includes JSICK (Yanaka and Mineshima, 2022) (5,000 training examples) and JSNLI (Yoshikoshi et al., 2020) (approximately 530,000 training examples). JSICK is a manually translated Japanese dataset derived from the English RTE dataset SICK (Marelli et al., 2014), which consists of sentences encompassing various lexical, syntactic, and semantic phenomena. JSNLI is a large-scale Japanese RTE dataset created by machine translation from the English SNLI dataset. From here on, the BERT model trained on JSICK will be referred to as JSICK BERT, and the BERT model trained on JSNLI will be referred to as JSNLI BERT. We performed a three-class classifier (Entailment, Neutral, Contradiction) on the constructed 149 pair inference test set using JSICK BERT and JSNLI BERT.

BERT model	JSICK BERT	JSNLI BERT
Entailment	98	85
Neutral	49	47
Contradiction	2	17
Total	149	149

Table 10: RTE results using BERT.

There are no examples that both JSICK BERT and JSNLI BERT classify as contradiction, but there are 33 examples that neither of them classify as entailment.