

Title	神経意味解析における言語的特徴と事前訓練された言語モデルの注入
Author(s)	NGUYEN, MINH PHUONG
Citation	
Issue Date	2022-09
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/18128
Rights	
Description	Supervisor:Nguyen Minh Le, 先端科学技術研究科, 博士

氏名	NGUYEN, Phuong Minh		
学位の種類	博士 (情報科学)		
学位記番号	博情第 472 号		
学位授与年月日	令和 4 年 9 月 22 日		
論文題目	INJECTING LINGUISTIC FEATURES AND PRE-TRAINED LANGUAGE MODEL IN NEURAL SEMANTIC PARSING		
論文審査委員	Nguyen Le Minh	JAIST	Professor
	Satoshi Tojo	JAIST	Professor
	Kiyoki Shirai	JAIST	Assoc. Prof.
	Ken Satoh.	NII	Professor
	Naoya Inoue.	JAIST	Assoc. Prof.

論文の内容の要旨

Nowadays, in the quick development of science and technology, human life is getting more and more modern convenient with machines. Especially, computers and the internet are the major factors that help people connect together by storing, sharing, and searching for knowledge information in any domain. However, most of this information is written in unstructured data using human natural language, which is hard for logical searching as well as retrieval the meaningful information given complex questions. Therefore, this study aims to investigate the Semantic Parsing task in Natural Language Processing (NLP) that map a natural language sentence to machine-understandable information representation. In the developing history of this task, there are many kinds of semantic representations have been introduced and developed such as logical form, semantic frame, semantic graph, etc.

In this thesis, we introduce the effective methods using the neural network to solve the Semantic Parsing task. We focus on two kinds of widely used logic representations, logical form, and semantic frame as well as the issues of these semantic schemes. To this end, we propose the potential approaches to deal with the challenges of Semantic Parsing task, and present powerful methods for this tasks in the legal domain.

The first challenge we targeted is the local context integration in Semantic Parser. Inspired by grammar-based methods, the semantic representation of a sentence is the combination of sub-meaning representation generated by phrases in a sentence. Besides, the current state-of-the-art models using Transformer adapted from Neural Machine Translation task do not have components modeling phrase information. Therefore, we propose the Phrase Transformer - a new architecture incorporating representation of phrase via n-gram chunking into Self-Attention mechanism of the original Transformer. Our experimental results show that the proposed model works effectively and beat the original Transformer by utilizing local context features better.

The second issue we explored is the class imbalance in logical representation using semantic frames. The significant difference between the majority and minority classes causes the semantic parsing model confused in minority classes recognition. The observations on well-known datasets show that this problem is highly critical, special among Slot classes. To deal with this problem, we propose the Classify Anonymous Entities (CAE) mechanism by using multitask joint-learning to split the conventional Slot Filling task into two sub-task: detect anonymous entity by sequence tagging and classify recognized anonymous entities tasks.

Finally, we focus on constructing the semantic parser in the legal domain. The main challenges relate to the length and content of legal documents containing complex constraints about the conditions of articles. Besides, the limited annotated semantic parsing data also is a difficulty in this domain. Based on the DAPRECO Knowledge Base (KB), we firstly re-construct the GDPR (General Data Protection Regulation) Semantic Parsing dataset mapping a GDPR article points into its expression in DAPRECO KB. We also implement a Semantic Parser on this data and propose two mechanisms: Sub-expression intersection and Predicate REtrieval &

Sub-Expression Generation (PRESEG) to deal with the problems in the legal domain.

To summarize, our study is centered on dealing with fundamental problems to model Semantic Parser using a deep learning approach and adapting to the legal domain. The experimental results and detailed analysis proved the effectiveness of the proposed methods as well as the potential for domain adaptation. Despite the experiments being conducted on limited kinds of semantic schemes, the proposed models and solution ideas have the potential to be widely applied to other types of semantic representations or to various tasks in NLP in further research.

Keywords: Semantic Parsing, Phrase Transformer, Neural Machine Translation, Class Imbalance, Spoken Language Understanding, Legal Semantic Parsing.

論文審査の結果の要旨

Semantic parsing is one of the most challenging tasks in natural language processing due to language ambiguity issues. The thesis mainly exploits neural transformer models tackling semantic parsing problems. The research shows that incorporating local context integration in Semantic Parser can enhance the performance of semantic parsing in various domains. Therefore, the candidate proposes the Phrase Transformer - a new architecture incorporating representation of phrase via n-gram chunking into the Self-Attention mechanism of the original Transformer. The experimental results show that our proposed model work beat the original Transformer by utilizing local context features better. The second issue the research explored is the class imbalance in logical representation using semantic frames by proposing the Classify Anonymous Entities (CAE) mechanism. Finally, the thesis updates the phrase transformer model on constructing the semantic parser in the legal domain by dealing with the challenges related to the length and content of legal documents containing complex constraints about the conditions of articles. Overall, this is an excellent dissertation, and we approve of awarding a doctoral degree to Nguyen Minh Phuong-san.