科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　６　月　４　日現在

機関番号：13302
研究種目：若手研究（Ｂ）
研究期間：2010～2011
課題番号：22700139

研究課題名（和文）　文から論理表現への文脈依存変換を学習するための半教師モデル

研究課題名（英文）　Contextual information and semi-supervised models for transforming sentence to logical form representation

研究代表者
グエン　ミン　レ（Nguyen Minh Le）
北陸先端科学技術大学院大学・情報科学研究科・助教
研究者番号：30509401

研究成果の概要（和文）：
意味解析の改良をねらい，文脈素性を用いる方法を実装することを主に研究した．
また，意味解析モデルを実装するのにラベルなしデータがどのように有効かについても研究しました．これらの点に関して，注釈のない大規模コーパスを語クラスタモデルによりモデル化し，識別学習モデルのための素性を抽出しました．意味表現と自然言語文の同期モデルを学習するために，forest-to-string 法を適用した． この問題に対し，我々は，機械学習および線型計画法を用いて，法令条文の項(paragraph)の論理構造を２段階で学習する新しい枠組みし示した．

研究成果の概要（英文）

The main goal of our research is to implement the method of using contextual features for improving semantic parsing problems. We also study how unlabeled data could help to implement semantic parsing model further. As a result, we exploited word-cluster models to model a large un-annotated corpus, to extract features for discriminative learning models. In addition, we also introduce a novel semi supervised learning model for semantic parsing with ambiguous supervision. We applied the forest-to-string method for learning the synchronous model between semantic representation and natural language sentence. We also present a novel two-phase framework to learn logical structures of paragraphs in legal articles using machine learning and integer linear programming.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,100,000	330,000	1,430,000
2011 年度	700,000	210,000	910,000
年度			
年度			
年度			
総　計	1,800,000	540,000	2,340,000

研究分野：統計的自然言語処理テキスト要約，機械翻訳，言語理解
科研費の分科・細目：情報学・知能情報学
キーワード：　意味解析，言語理解

１．研究開始当初の背景

Recent works on semantic parsing are based on supervised machine learning approaches [ZET07][WON 07][KAT07][RE09][NGU08]. Like other works, our model [NGU08] is based on supervised learning model which needs a corpus of natural language sentences and its logical form. Those supervised learning models have focused on parsing sentence in isolation. Unlike those methods, [ZET09] considered the problem of learning to interpret sentence whose underlying meanings can depend on the context in which they appear. They have modified their previous model [ZET07] by introducing a hidden-variable variant of the perceptron algorithm.

Poon and Domigos [POO09] have developed an unsupervised semantic parsing model using Markov Logic. Liang et al (2009) have shown a less supervision learning model that learns to correspond to semantic representation and natural language sentence. The significant of this approach is that we do not need human effort in annotating training data. In addition, the unsupervised learning models could attain the state of the art results on several domains such as Robocup, sport-casting, and weather forecast [LIA09]. However, the limitation of unsupervised models is that it depends on too much on the language domain. In addition, unsupervised models work well on simple (short) sentences and simple meaning representation, but it is difficult for dealing with complex sentences and meaning representation [LIA09].

２．研究の目的

The purpose of this research aims at studying how well contextual information and unlabeled data are contributed to the performance of semantic parsing and natural language generation. We also studied how contextual information could be adaptively applied for legal domain.

３．研究の方法

In this project, the main goal of our research is to implement the method of using contextual features for improving semantic parsing problems. We also study how unlabeled data could be used to implement semantic parsing model further. As a result, we exploited word-cluster models to model a large un-annotated corpus, to extract features for discriminative learning models. For semantic parsing, we designed a semi-supervised model to the problem of semantic parsing in ambiguous supervision. We incorporated word-cluster model to enrich feature space and kernel matrix for semantic parsing problems. In addition to investigate appropriate semantic parsing models for general domain, we study how we can exploit semantic parsing model to the legal domain. The main problem is that a sentence in a legal domain is often long and complex. Multiple sentences within a paragraph are strongly related. We would like to utilize these relations within legal paragraph in order to exploiting semantic parsing.

One of the important issues for transforming NL sentence to logical form representation is how to evaluate them. In the project we focus on applying textual entailment for evaluating natural language generation.

４．研究成果

(1) We performed the proposed semi-supervised learning models for various natural language applications including part of speech tagging, legal processing, and text summarization. All the results are reflected that our semi-supervised model is effective.

(2) We then apply it to the problem of semantic parsing that maps a nature sentence to a meaning representation. The limitation of semantic parsing is that it is very difficult for obtaining annotated training data in which a sentence corresponding with a semantic representation. We proposed a semantic parsing approach for ambiguous training data using maximum entropy model. The semi-supervised learning model as mentioned above is then used for ambiguous training data. Experimental results on the standard data showed that the proposed method efficiently works well on ambiguous data.

(3) We have developed a semantic generation system, which used the contextual information to obtain an appropriate sentence forgiven a semantic representation. We applied the forest-to-string method for learning the synchronous model between semantic representation and natural language sentence. The multiple best outputs of the generation are obtained, and the appropriate one is selected using the contextual information with a re-ranking algorithm.

(4) In addition, we also introduce a novel semi supervised learning model for semantic parsing with ambiguous

supervision The main idea of our method is to utilize a large amount of data, to enrich feature space for machine learning models using in the proposed semantic learner. We empirically showed that word-cluster features obtained from unlabeled data, are effective for both the string kernel SVM method and the maximum entropy model. A careful evaluation of the proposed models on standard corpora showed that our methods are suitable for semantic parsing.

(5) One of the main goals I would like to achieve is to investigate current semantic parsing techniques to the legal domain. However, this adaptation task is not straightforward because legal sentences are often long and complex. In some cases, we need to parse all legal sentences within a paragraph, to determine the meaning of each sentence. The goals of this task are recognizing logical parts of law sentences in a paragraph, and then grouping related logical parts into some logical structures of formulas, which describe logical relations between logical parts. We present a novel two-phase framework to learn logical structures of paragraphs in legal articles using machine learning and integer linear programming.

(6) One of the important issues is the problem of recognition of textual entailment (TE). Our goal is to investigate this research direction for legal domain. In this year, we have recently published the paper on recognizing textual entailments for both Japanese and Vietnamese. The main idea of our method is to apply machine-learning models in which rich linguistic features are investigated. In addition, we proposed to use multilingual features for our textual recognition models. The multilingual features are obtained by using a machine translation system to translate textual entailment data from other languages (such as English). We participated the shared task competition for Japanese (RET NTCIR-9RITE 2011) in this year and our team achieved the first rank.

５．主な発表論文等

〔雑誌論文〕（計４　件）

[1] B.X. Ngo, M.L. Nguyen, T.T. Oanh, A. Shimazu. A Two-Phase Framework for Learning Logical Structures of Paragraphs in Legal Articles. *ACM Transactions on Asian Language Information Processing (ACM TALIP)*, 2012 (accepted)(reviewed)

[2] M.Q.N. Pham, M.L. Nguyen, B.X. Ngo, A. Shimazu. A learning-to-rank method for information updating task. *Applied Intelligence*, 2012 (accepted)(reviewed)

[3] B.X. Ngo, M.L. Nguyen, A. Shimazu. RRE Task: The Task of Recognition of Requisite Part and Effectuation Part in Law Sentences. *International Journal of Computer Processing Of Languages (IJCPOL)*, Volume 23, Number 2, pp. 109-130, 2011 (reviewed)

[4] M.L. Nguyen and A. Shimazu. Improving Subtree-based Question Classification Classifiers with Word-Cluster Models. Lecture Notes in Computer Science 6716 Springer 2011, ISBN 978-3-642-22326-6 (reviewed)

〔学会発表〕（計８　件）

[1] M.L. Nguyen, A. Shimazu, "A Semi Supervised Learning Model for Mapping sentences to logical form with ambiguous supervision" , *In Proceedings NLDB-2012, June 26-28, 2012, Groninge, Netherlands*

[2] M.Q.N. Pham, M.L. Nguyen, A. Shimazu. (2012). An Empirical Study of Recognizing Textual Entailment in Japanese Text. *In Proceedings of 13th International Conference on Intelligent Text Processing and Computational Linguistics (LNCS)*, pp. 438-449, March 11-17, 2012, New Delhi, India

[3] M.Q.N. Pham, M.L. Nguyen, A. Shimazu. (2012). Using Machine Translation for Recognizing Textual Entailment in Vietnamese, *2012 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF)*, Ho Chi Minh City, Vietnam, February 27 - March 1, 2012

[4] M. Pham, M.L. Nguyen and A. Shimazu., "Update Legal Documents Using Hierarchical Ranking Models and Word Clustering" , *the 23rd International Conference on Legal Knowledge and*

Information Systems. University of Liverpool (U.K.), 16th-17th December 2011

[5] M.L. Nguyen, N.X. Bach, A. Shimazu, Supervised and Semi-Supervised Sequence Learning for Recognition of Requisite Part and Effectuation Part in Law Sentences , *Proceedings of the 9th International Workshop on Finite-State Methods and Natural Language Processing (FSMNLP)*, Blois, France, July 2011

[6] V.C. Nguyen, M. L. Nguyen, A. Shimazu: Improving Text Segmentation with Non-systematic Semantic Relation. *12th International Conference on Intelligent Text Processing and Computational Linguistics (LNCS) (1) 2011: 304-315*, Feb 20-26, 2011, Tokyo, Japan

[7] B.X. Ngo, M.L. Nguyen, A. Shimazu. Exploring Contributions of Words to Recognition of Requisite Part and Effectuation Part in Law Sentences. In *Proceedings of the 4th International Workshop on Juris-Informatics* (JURISIN), pp. 121-132, Tokyo, Japan, November 2010

[8] M.L. Nguyen, B.X. Ngo, C.V. Nguyen, M.Q.N. Pham and A. Shimazu. , "A Semi-Supervised Learning Method for Vietnamese Part of Speech Tagging", *In Proceedings of the 2nd International Conference on Knowledge and Systems Engineering (KSE)*, 1, 141-146, 2010, October 7-9, 2010, Hanoi, Vietnam

６．研究組織

(1)研究代表者
グエン　ミン　レ (Nguyen Minh Le)
北陸先端科学技術大学院大学・情報科学
研究科・助教
研究者番号：30509401