Data & Knowledge Engineering 90 (2014) 1–12

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Editorial

A semi supervised learning model for mapping sentences to logical forms with ambiguous supervision

Minh Le Nguyen *, Akira Shimazu

School of Information Science, JAIST, Nomi 1-1, Ishikawa 923-1292, Japan

ARTICLE INFO

Article history:
Received 11 June 2013
Received in revised form 8 August 2013
Accepted 20 December 2013
Available online 7 January 2014

Keywords:
Semantic parsing
Semi-supervised learning
Ambiguous supervision

ABSTRACT

Semantic parsing is the task of mapping a sentence in natural language to a meaning representation. The limitation of previous work on supervised semantic parsing is that it is very difficult to obtain annotated training data in which a sentence is paired with a semantic representation. To deal with this problem, we introduce a 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 with the maximum entropy model using our semantic learner. We evaluate the proposed models on standard corpora to demonstrate that our methods are suitable for semantic parsing. Experimental results show that the proposed methods work efficiently and well on ambiguous data and it is comparable to the state of the art methods.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Semantic parsing is the task of mapping a natural language sentence into a complete and formal meaning representation. This task is an interesting problem in Natural Language Processing (NLP) as it would very likely be part of any interesting NLP applications [2]. Semantic parsing also plays a very important role for natural language processing for database [14]. For example, the necessity of semantic parsing for most NLP applications and the ability to map natural language to a formal query or command language, are critical for developing more user-friendly interfaces.

Semantic parsing studies can be divided into two main approaches. The first one relies on a wide coverage semantic parser [4], and the second one considers building a domain-specific semantic parser. The issue with open-domain text is maybe that there is ore linguistic diversity, and especially more lexical variation. Therefore, semantic parsing for open domain text has still been a challenging task. In the scope of this paper, we focus on domain-specific semantic parsing, which has received much attention in recent years. There has been a significant amount of studies on learning to map sentences to semantic representations. Zelle and Mooney [38] and Tang [30] proposed the empirically-based methods using a corpus of natural language (NL) sentences and their formal representation for learning by inductive logic programming (ILP). The disadvantage of the ILP approach is that it is quite complex, and it is slow to acquire parsers for mapping long sentences to logical forms such as the Robocup corpus.

To overcome this problem, Kate and Mooney [15] proposed a method that used transformation rules estimated from the corpus of NL and logical forms, to transform NL sentences to logical forms.

In order to improve semantic parsing accuracy, Ge and Mooney (2005) presented a statistical method [12] merging syntactic and semantic information using a semantic augmented parsing tree (SAPT).[1]

Similar to Ge and Mooney (2005) [12], the approach by Nguyen, Shimazu, and Phan [27] also uses the corpus of SAPT trees to estimate their semantic parsing model. Their approaches exploit structured SVM models [33] to learn ensemble learning of semantic parsers.

* Corresponding author at: 1-1, Ishikawa 923-1292, Japan.
E-mail address: nguyenml@jaist.ac.jp (M. Le Nguyen).

[1] SAPT is a syntactic tree with semantic augmented at each non-terminal node.

0169-023X/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.datak.2013.12.001

Unlike those methods using SAPT, the works proposed by [39,40,21] map an NL sentence to its logical form using a Combinatory Categorical Grammar (CCG) with structured learning models. They have indicated that using online structured prediction along with CCG could lead to state of the art results in several domains (i.e. ATIS and Query language). Wong and Mooney (2006) [35] proposed a synchronous context-free grammar framework (SCFG) [1] to transform a language sentence (NL) to a meaning representation (MR). The system was extended to work with the formal language in λ-calculus (Wong and Mooney, 2007) to deal with logical variables.

All of the above methods require a human annotation process to create training data, which consists of a set of pairs of sentences and their logical representations. However, creating accurate training data is expensive and time-consuming. To cope with this problem, Kate and Mooney [18] initially proposed a semi-supervised learning model for semantic parsing using transductive SVMs. Along with the research line, there are several methods for semantic parsing using unlabeled data [37,9,31,29,28]. However, their approaches are designed for un-ambiguous supervision, while in real applications such as Robocup casting as well as weather-forecasting, we do not know exactly which semantic representation is paired with a sentence. Kate and Mooney (2007a) extended their semantic parser based on string kernel for ambiguous training data in which each sentence is only annotated with an ambiguous set of multiple, alternative potential interpretations. The merit of this work is that using a string kernel SVM one can deal with ambiguous and noisy data. Titov and Kozhevnikov [32] used unlabeled data for bootstrapping ambiguous semantic parsing, which does not rely on using large scale unlabeled data.

According to our knowledge, there is no research about using *external large unlabeled data* for semantic parsing with ambiguous supervision. In this paper, we would like to explore whether or not unlabeled data is useful for semantic parsing on ambiguous training data. To achieve this goal, we extend the method of [17] by proposing semi-supervised learning models using unlabeled-data with word-cluster models. The contributions of our proposed method are described as follows.

- Unlike previous work, we propose a semi-supervised learning method which allows incorporating unlabeled data to improve the performance of semantic parsing using ambiguous training data. The large amount of unlabeled data can be utilized to generate a cluster model to enrich the feature space of the learning model.
- Our goal is to answer a question whether or not unlabeled data is helpful for the task of mapping sentences to logical forms. In order to do that, we propose two semantic-parsing models, using unlabeled data to enrich the feature space of learning models. The first model is extended from the SVM-based semantic parsing model by incorporating unlabeled data to its string kernel. The second model is based on a maximum entropy model (MEM) in which unlabeled data is used to enrich the feature space of MEM. In addition, we also investigate the impact of using syntactic information (i.e. part of speech tagging, chunking labels) for mapping sentences to logical forms.

The rest of this paper is organized as follows: Section 2 describes the background on semantic parsing and machine learning models for semantic parsing. Section 3 gives an overview of semantic parsing for ambiguous supervision. Section 4 describes a semi-supervised method for semantic parsing in both unambiguous training data and ambiguous training data. Section 5 shows experimental results, and Section 6 discusses the advantage of our method and describes future work.

2. Background

To support readers in terms of understanding the paper easily, we briefly introduce some definitions, machine learning models, and base techniques used in the paper. We describe them in the following subsections.

2.1. Meaning representation

The semantic parsing process maps sentences to their computer-executable meaning representation (MRs). The MRs are expressed as informal languages, which are defined as meaning representation languages (MRLs). We assume that all MRLs are defined by a deterministic context-free grammar (we call MRL grammar) which ensures that every MR will have a unique parse tree. A learning system for semantic parsing is given a training corpus of NL sentences paired with their respective MRs from which it has to induce a semantic parser which can map novel NL sentences to their correct MRs. Fig. 1 illustrates how an NL query could be represented in a meaning representation. It shows the parse of the NL sentence and its productions for generating the tree. The non-terminals and terminals are shown in upper-case and lower-case, respectively. The terminals represent predicates, and we can obtain a parse tree of a sequence of terminals.

2.2. Machine learning models for semantic parsing

2.2.1. Support vector machine

Support Vector Machine (SVM) [10,34] is a classification model which uses a strategy that maximizes the **margin** between training samples and the hyperplane. SVMs have demonstrated their performance on natural language processing a number of problems in areas, including computer vision, handwriting recognition, pattern recognition, and statistical natural language processing. In the field of natural language processing, SVMs have been successfully applied to text categorization [13], text chunking [20], semantic parsing [16,27], and so on.

Download English Version:

https://daneshyari.com/en/article/378820

Download Persian Version:

https://daneshyari.com/article/378820

Daneshyari.com