Adapting DCS-Trees for Wide-Domain Semantic Parsing

Amelia Harrison
Department of Computer Science
University of Texas at Austin
ameliaj@cs.utexas.edu

Subhashini Venugopalan
Department of Computer Science
University of Texas at Austin
vsubhashini@cs.utexas.edu

Abstract

Recent work in semantic parsing for compositional question answering has shown that this task can be successfully approached using question-answer pairs rather than questions annotated with logical forms for training. However, existing datasets for this task are relatively small and it is unclear that even in absence of the burden of producing logical forms for training, that any existing approaches to compositional question answering can scale to larger datasets or datasets with wide coverage. In this paper, we explore the possibility of scaling one such approach to a larger dataset with wider coverage.

1 Introduction

In (**?**), Liang et al. present a novel approach to the semantic parsing task. Most prior approaches have focused on training parsing models using natural language utterances annotated with logical forms.[1] In this work, however, the logical forms are treated as latent variables, and a semantic parsing model is trained using natural language questions and their corresponding answers. This sort of "weak supervision" is advantageous because obtaining logical forms for natural language utterances requires expert annotators and is thus quite costly. This work is notable because it takes this weakly supervised approach, and also because it introduces an new formalism called dependency-based compositional semantics (DCS) trees for representing the meaning of a natural language question. In this note, we will briefly summarize the work in (**?**), and suggest some useful ways in which that work may be extended.

[1](**?**), (**?**), and (**?**) are notable exceptions.

Recent work has shown that the compositional question answering task can be successfully approached using question-answer pairs for training rather than questions annotated with logical forms (**?**) (). In (**?**),

2 Experimental Evaluation

This section describes the datasets used to evaluate our methods and the results obtained.

2.1 Data

We tested our approach on two datasets. The first is the standard GEO dataset (**?**). To test the scalability of our approach we also create a wide domain dataset by extracting data from DBpedia (**?**). The DBpedia data set consists of a multi-domain ontology derived from Wikipedia. Our project uses the DBpedia Infobox dataset (**?**; **?**) which contains specific facts about things (such as people, places, films, music, books, games) in Wikipedia articles. The Infobox data consists of

- Type information of the instances in wikipedia such as city, disease, computer game etc.
- Properties of the instances e.g. population of a city, date of birth.
- Specific properties specialized for some property value such as units of measurement for a person's height.

We build our DBpedia data set by extracting type and property information of all instances (as tuples - describe). This results in a large dataset of unary and binary predicates consisting of over five million facts.

Relevant wide-domain dataset We reduce the initial large dataset (DBP) to a more relevant smaller dataset consisting of 538,821 facts. To create this subset, we first considered all the binary predicates in the initial dataset (of the form predicate(key, value)). Next, we identify all items occuring as a key. We then count the frequency of occurence of each of these items as a value in the tuple and rank and pick the top 30000 most frequent items. We finally select all tuples containing any of these top items as a key. The idea is to select those items that are popular in the database and hence likely to also be more relevant.

2.2 Methodology

We evaluate our approach of generating lexical triggers with the DCS framework on the GEO dataset as outlined below. We split our evaluation into the following four categories:

- $ours$ - Lexical triggers generated purely from our approach.
- $+bridge$ - In addition to our trigger set, we add a small set of bridge predicate triggers described in (**?**). (expand?)
- $+general$ - This includes a set of handcoded a general trigger set of lexicons for mathematical operations such as sum and argmax (along with our triggers)
- all - The final set of triggers include all of the above, namely, our lexical triggers, the bridge predicate triggers and the general handcoded trigger set.

We split the lexical sets into the above categories for triggers generated using wordnet and wackypedia. In all the experiments, the training set consisted of 600 questions and the test set consisted of 280 questions.

Impact of number of lexical triggers per predicate. We further performed a series of tests to identify if the number of lexical triggers per predicate affected the performance in any way. This test was performed only on the lexical trigger set generated from wackypedia since the context lexical words have a total ordering (rank) and also because this set gave us the best performance. In our experiments we varied the number of lexical words from 1 to 12 and compared the performance of the DCS framework on the four categories described in 2.2.

3 Results

Lexicon Source Category	**WordNet**	**Wackypedia** (threshold=6)
ours	0.00	21.78
+bridge	0.00	42.5
+general	0.00	36.07
all	0.00	**69.64**

Table 1: Comparing performance of external lexical trigger sources on GEO dataset (600 train and 280 test queries)

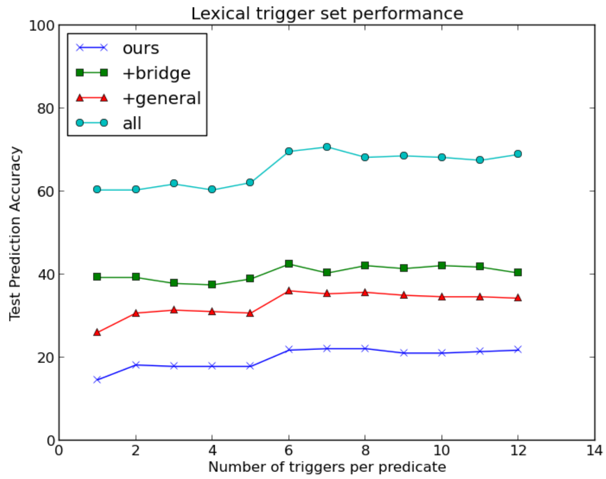

Figure 1: Graph showing the effect of varying the number of lexical triggers per predicate on GEO.

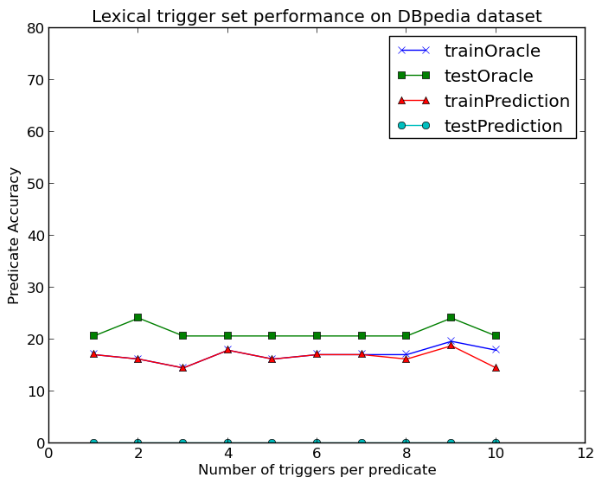

Figure 2: Performance of automatically generated lexical triggers on the wide-domain DBP dataset (80-20 split).

4 Related Work

Most prior approaches to the semantic parsing task focus on training parsing models using natural language utterances annotated with logical forms cite[Wong Mooney, Kate Moooney, Zettlemoyer Collins 2005/7/9]. Their algorithms learn the hidden associations between the logical forms and lexical and semantic elements. The task of annotation is expensive and recent works cite[CGRR11, CGCR10, PD09, LJK11] reduce the burden of annotation significantly by obviating the need for logical form annotations and focus on learning from question and answer pairs.

CGCR10 uses a feedback based approach to guide the learning algorithm to identify hidden alignments between questions and logical forms. In addition, their model learns to create the logical form query based on the composition rules of the meaning representation language and learning lexical mappings from external sources. While CGCR10 obviates the need for logical form annotations, they still rely on the framework of the meaning representations language which lacks in its expressiveness (in its ability to express) natural language queries. LJK11 proposes the DCS trees logical form that pushes towards a deeper representation of language. Each natural language question is associated with a number of permissible DCS trees which may correspond to the actual meaning and in turn evaluate to the right answer. The set of permissible trees is determined by a fixed set of lexical triggers which map words or sequences of words to with a predicate in the database. While the DCS framework appears to be a good candidate to tackle wide-domain semantic parsing, the model requires some non-trivial supervision to identify the set of lexical triggers. Our work addresses the task of automatically identifying the lexical triggers and reduce the burden of supervision further.

5 Other Issues

Those papers that had software and/or dataset submitted for the review process should also submit it with the camera-ready paper. Besides, the software and/or dataset should not be anonymous.

Please note that the publications of ACL 2013 will be publicly available at ACL Anthology (http://aclweb.org/anthology-new/) on July 28th, 2013, one week before the start of the conference. Since some of the authors may have plans to file patents related to their papers in the conference, we are sending this reminder that July 28th, 2013 may be considered to be the official publication date, instead of the opening day of the conference.

Acknowledgments

Thanks to Karl Pichotta and Ray Mooney for helpful discussions and healthy skepticism.

report.bib

References