Semantic Parsing for Information Extraction

Eunsol Choi Tom Kwiatkowski Luke Zettlemoyer
University of Washington
185 NE Stevens Way
Seattle, WA 98105
eunsol,tomk,lsz@cs.washington.edu

1 Introduction

We consider the challenge of learning semantic parsers that scale to large, open-domain problems, such as question answering or knowledge base completion with Freebase. In such settings, the sentences cover a wide variety of topics and include many phrases whose meaning is difficult to represent in a fixed target ontology. For example, even simple phrases such as 'daughter' and 'number of people living in' cannot be directly represented in Freebase, whose ontology instead encodes facts about gender, parenthood, and population. Here, we introduce a new semantic parsing approach that learns to resolve such ontological mismatches. The parser uses a probabilistic CCG (Steedman, 1996; Clark and Curran, 2007) to build linguistically motivated logical-form meaning representations, and includes an ontology matching model that adapts the output logical forms for each target ontology. In this abstract, we describe an ongoing work applying this approach for information extraction.

2 Information Extraction

We propose a new method to use semantic parsing for information extraction by mapping sentences to logical forms built from Freebase. We focus on noun phrases, which are very common in English and contain rich, compositional structure. When a sentence has a form NP_1 is NP_2, for example, *Barack Obama is 44th and current President of United States*, we know that NP_2 describes a set of attributes on NP_1. When we can link NP_1 to an entity in KB, by parsing NP_2 we can extract information about NP_1. Here we aim to train a semantic parser by aligning the semantic parser's logical representation with the Freebase query, and compare the entities labeled to noun phrases to entities returned by executing Freebase query.

Previous information extraction methods have either focused on a relatively small set of relations (Yao et al., 2011) or extracted large number of relations that are not canonicalized to any ontology (Banko et al., 2007). Our goal here is to generalize beyond relations and support extraction as much of the overall structure found in Freebase as possible, through rich ontology matching.

3 Dataset

YAGO, a semantic knowledge base with more than 10 million entities and 120 million facts, contains a dataset that fits our needs exactly. The YAGO category dataset is composed of 7 million pairs of noun phrases (i.e, 20th century American poets) and entities belonging to them, as well as mapping from YAGO entity to Freebase entity. For example, Figure 1 shows one such phrase "Symphonic Poems by Jean Sibelius" and its associated entities. Such noun phrases are automatically derived from human generated categories in Wikipedia. We propose a new way of utilizing this dataset as a large-scale resource to train an open-domain semantic parser.

4 Learning a two-stage semantic parser

We are using the two-stage semantic parsing paradigm introduced in (Kwiatkowski et al., 2013). The first parsing stage uses a probabilistic combinatory categorial grammar (CCG) to map sentences to new, *underspecified* logical-form meaning representations, which closely mirror the linguistic structure of the sentence but do not contain constants from the target ontology O. For example, in Figure 1, l_0 denotes the underspecified logical forms paired with each sentence x. The parser then maps this intermediate representation to a logical form that uses constants from O, such as y in the example. The ontology-matching step considers a large number of type-equivalent domain-specific meanings. It incorporates a number of cues, including the target on-

x :	Symphonic Poems by Jean Sibelius
l_0 :	$\lambda x.Symphonic(x) \wedge Poems(x) \wedge by(JeanSibelius, x)$
y :	λx.composition.form(x, Symponicpoems) $\wedge$ composition.composer(JeanSibelius, x) $\wedge$ music.composition(x)
e :	{The Bard, Finlandia,Pohjola's Daughter, En Saga, Spring Song, The Swan of Tuonela, Tapiola... }

Figure 1: Examples of sentences x, domain-independent underspecified logical forms l_0, fully specified logical forms y, and entities belong to noun phrases e from YAGO category data.

		%
Schema		75.5
Schema with Entity		49.5
Schema with Overlapping Entity		36
YAGO coverage on Freebase	P	40
	R	32
	F1	26

Figure 2: Annotation study on 200 noun phrases.

tology structure and lexical similarity between the names of the domain-independent and ontology-specific constants, to construct the final logical forms in the ontology.

Such parsers can be learned automatically from examples of natural language phrases paired with the sets of entities they describe in the target knowledge base. For example, (Kwiatkowski et al., 2013) used the perceptron algorithm to estimate a linear model from such data, where most of the intermediate decisions, including the CCG parse steps and ontology matching decisions, are latent.

5 Challenge: Incomplete data

While YAGO provides a source of training data, its incompleteness can complicate learning. Knowledge base incompleteness is a well-known problem; it is very difficult to specify all true facts for the entities and some entities will inevitably be missing entirely. For example, a study from (West et al., 2014) showed that over 70% of people in Freebase have no known birth place, and 99% have no known ethnicity.

5.1 Annotation study

To understand the coverage of YAGO category information in Freebase and vice-versa, we manually annotated 200 YAGO category with Freebase queries. Given these annotations, we measured two things: (1) how many YAGO category noun phrases can be represented with a query against the Freebase schema and (2) how accurate the Freebase entity results are for the queries that can be written, as compared to the entities in the YAGO lists. The study showed 75.5% of YAGO category noun phrases can be fully expressed in Freebase queries. However, the majority of these queries returned the empty set, demonstrating the incompleteness in Freebase. For those queries which returned any entities, 99 out of 200, the overlap to YAGO-labeled entities was also sparse, as shown in Figure 2.

6 Proposed Approach

The sparse coverage of YAGO category information in Freebase poses a serious challenge in using the set for supervision. Directly comparing query output to the YAGO entity sets would not provide a strong learning signal, due to the high levels of incompleteness seen in our small study. We aim to resolve this by instead using the data as a source of weak supervision for learning features, which can then be tuned from a small supervised training set to produce the final model.

6.1 PMI

Although the incompleteness is a challenge for direct supervision, in aggregate the co-occurance information between the categories and their entities still provides strong cues about which words should be associated with the different parts of the Freebase schema. For example, in Figure 1 some, but not all, of the entities in e would have the attribute composition.form(x, Symponic_Poem), which should be paired with the word "Symphonic poem," providing a weak hint about the association. Aggregating across all categories with "Symphonic poem", such as "Symphonic poem in the 19th century", can strengthen the correct association. To capture this signal, we computed pointwise mutual information between phrase and Freebase attributes represented in 7 million noun phrase-entity pairs. For each English phrase, we select the YAGO categories containing it, and also the entities in those categories sets. Then compute the PMI between each English phrase and Freebase attributes for all of its selected entities. This

PMI score will be added as a feature to help ontology matching step in our semantic parser, as described above.

6.2 Distributional Word Vectors

In the second ontology matching step of the two-stage semantic parser, we must map words in natural language to constants in the target ontology. Here, distributional word vector representations can be used to associate semantically related words in the text and ontology labels. For example, such associations might allow us to mapping the word "players" to the attribute "`sports.pro_athlete`" and "located in" to "`location.containedby`".

We will consider pre-trained embeddings from general corpus such as Wikipedia (Collobert et al., 2011), by adding it in as a feature in the parser, and also consider co-occurances computed with OpenIE tuples. Finally, we would like to also consider learning with a joint objective, that estimated the embeddings along side the semantic parsing model.

7 Plan of Work

We will build this proposed approach on the YAGO data through a small set of supervised data, including (x, l_0, y, e) tuples, much like we see in Figure 1. Since the proposed features are unlexicalized, the learning model will have the potential to generalize from the development set to all of the 350,000 other noun phrases in the YAGO dataset. This parser, learned from large-scale corpus, can further be generalized to parse other noun phrase in different texts such as any sentences of format "$NP1$ is $NP2$", or appositives structures ("$Entity$, $Noun\_Phrase$,.....).

We will first evaluate the performance by asserting the system output query to annotated queries, but later it can be evaluated by sets of sentences with $NP1$ is $NP2$ structure, by testing whether the query output generated from $NP2$ contains $NP1$. The framework can also be extended to suggest new schema to the existing knowledge base, as the system is not restricted to pre-defined set of relations.

References

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 2670–2676, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

S. Clark and J.R. Curran. 2007. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4):493–552.

Ronan Collobert, Jason Weston, Lon Bottou, Michael Karlen, Koray Kavukcuoglu, Pavel Kuksa, and Michael Collins. 2011. Natural language processing (almost) from scratch. arxiv:1103.0398v1.

Tom Kwiatkowski, Eunsol Choi, Yoav Artzi, and Luke Zettlemoyer. 2013. Scaling semantic parsers with on-the-fly ontology matching. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1545–1556, Seattle, Washington, USA, October. Association for Computational Linguistics.

M. Steedman. 1996. *Surface Structure and Interpretation*. The MIT Press.

Robert West, Evgeniy Gabrilovich, Kevin Murphy, Shaohua Sun, Rahul Gupta, and Dekang Lin. 2014. Knowledge base completion via search-based question answering. In *WWW*.

Limin Yao, Aria Haghighi, Sebastian Riedel, and Andrew McCallum. 2011. Structured relation discovery using generative models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1456–1466, Stroudsburg, PA, USA. Association for Computational Linguistics.