Semantic Parsing using Distributional Semantics and Probabilistic Logic

Islam Beltagy[§] **Katrin Erk**[†] **Raymond Mooney**[§]
[§]Department of Computer Science
[†]Department of Linguistics
The University of Texas at Austin
Austin, Texas 78712
[§]{beltagy,mooney}@cs.utexas.edu
[†]katrin.erk@mail.utexas.edu

Abstract

We propose a new approach to semantic parsing that is not constrained by a fixed formal ontology and purely logical inference. Instead, we use distributional semantics to generate only the relevant part of an on-the-fly ontology. Sentences and the on-the-fly ontology are represented in probabilistic logic. For inference, we use probabilistic logic frameworks like Markov Logic Networks (MLN) and Probabilistic Soft Logic (PSL). This semantic parsing approach is evaluated on two tasks, Textual Entitlement (RTE) and Textual Similarity (STS), both accomplished using inference in probabilistic logic. Experiments show the potential of the approach.

1 Introduction

Semantic Parsing is probably best defined as the task of representing the meaning of a natural language sentence in some formal knowledge representation language that supports automated inference. A semantic parser is best defined as having three parts, a formal language, an ontology, and an inference mechanism. Both the formal language (e.g. first-order logic) and the ontology define the formal knowledge representation. The formal language uses predicate symbols from the ontology, and the ontology provides them with meanings by defining the relations between them.[1]. A formal expression by itself without an ontology is insufficient for semantic interpretation; we call it uninterpreted logical form. An uninterpreted logical form is not enough as a knowledge representation because the predicate symbols do not have meaning in themselves, they get this meaning from the ontology. Inference is what takes a problem represented in the formal knowledge representation and the ontology and performs the target task (e.g. textual entailment, question answering, etc.).

Prior work in standard semantic parsing uses a pre-defined set of predicates in a fixed ontology. However, it is difficult to construct formal ontologies of properties and relations that have broad coverage, and very difficult to do semantic parsing based on such an ontology. Consequently, current semantic parsers are mostly restricted to fairly limited domains, such as querying a specific database (Kwiatkowski et al., 2013; Berant et al., 2013).

We propose a semantic parser that is not restricted to a predefined ontology. Instead, we use distributional semantics to generate the needed part of an *on-the-fly* ontology. Distributional semantics is a statistical technique that represents the meaning of words and phrases as distributions over context words (Turney and Pantel, 2010; Landauer and Dumais, 1997). Distributional information can be used to predict semantic relations like synonymy and hyponymy between words and phrases of interest (Lenci and Benotto, 2012; Kotlerman et al., 2010). The collection of predicted semantic relations is the "on-the-fly ontology" our semantic parser uses. A distributional semantics is relatively easy to build from a large corpus of raw text, and provides the wide coverage that formal ontologies lack.

The formal language we would like to use in the semantic parser is first-order logic. However, distributional information is graded in nature, so the on-the-fly ontology and its predicted semantic relations are also graded. This means, that standard first-order logic is insufficient because it is binary by nature. Probabilistic logic solves this problem because it accepts *weighted* first order logic formulas. For example, in probabilistic logic, the

[1] For conciseness, here we use the term "ontology" to refer to a set of predicates as well as a knowledge base (KB) of axioms that defines a complex set of relationships between them

Proceedings of the ACL 2014 Workshop on Semantic Parsing, pages 7–11,
Baltimore, Maryland USA, June 26 2014. ©2014 Association for Computational Linguistics

synonymy relation between "man" and "guy" is represented by: $\forall x.\ man(x) \Leftrightarrow guy(x) \mid w_1$ and the hyponymy relation between "car" and "vehicle" is: $\forall x.\ car(x) \Rightarrow vehicle(x) \mid w_2$ where w_1 and w_1 are some certainty measure estimated from the distributional semantics.

For inference, we use probabilistic logic frameworks like Markov Logic Networks (MLN) (Richardson and Domingos, 2006) and Probabilistic Soft Logic (PSL) (Kimmig et al., 2012). They are Statistical Relational Learning (SRL) techniques (Getoor and Taskar, 2007) that combine logical and statistical knowledge in one uniform framework, and provide a mechanism for coherent probabilistic inference. We implemented this semantic parser (Beltagy et al., 2013; Beltagy et al., 2014) and used it to perform two tasks that require deep semantic analysis, Recognizing Textual Entailment (RTE), and Semantic Textual Similarity (STS).

The rest of the paper is organized as follows: section 2 presents background material, section 3 explains the three components of the semantic parser, section 4 shows how this semantic parser can be used for RTE and STS tasks, section 5 presents the evaluation and 6 concludes.

2 Background

2.1 Logical Semantics

Logic-based representations of meaning have a long tradition (Montague, 1970; Kamp and Reyle, 1993). They handle many complex semantic phenomena such as relational propositions, logical operators, and quantifiers; however, they can not handle "graded" aspects of meaning in language because they are binary by nature. Also, the logical predicates and relations do not have semantics by themselves without an accompanying ontology, which we want to replace in our semantic parser with distributional semantics.

To map a sentence to logical form, we use Boxer (Bos, 2008), a tool for wide-coverage semantic analysis that produces uninterpreted logical forms using Discourse Representation Structures (Kamp and Reyle, 1993). It builds on the C&C CCG parser (Clark and Curran, 2004).

2.2 Distributional Semantics

Distributional models use statistics on contextual data from large corpora to predict semantic similarity of words and phrases (Turney and Pantel, 2010; Mitchell and Lapata, 2010), based on the observation that semantically similar words occur in similar contexts (Landauer and Dumais, 1997; Lund and Burgess, 1996). So words can be represented as vectors in high dimensional spaces generated from the contexts in which they occur. Distributional models capture the graded nature of meaning, but do not adequately capture logical structure (Grefenstette, 2013). It is possible to compute vector representations for larger phrases compositionally from their parts (Landauer and Dumais, 1997; Mitchell and Lapata, 2008; Mitchell and Lapata, 2010; Baroni and Zamparelli, 2010; Grefenstette and Sadrzadeh, 2011). Distributional similarity is usually a mixture of semantic relations, but particular *asymmetric* similarity measures can, to a certain extent, predict hypernymy and lexical entailment distributionally (Lenci and Benotto, 2012; Kotlerman et al., 2010).

2.3 Markov Logic Network

Markov Logic Network (MLN) (Richardson and Domingos, 2006) is a framework for probabilistic logic that employ weighted formulas in first-order logic to compactly encode complex undirected probabilistic graphical models (i.e., Markov networks). Weighting the rules is a way of softening them compared to hard logical constraints. MLNs define a probability distribution over possible worlds, where a world's probability increases exponentially with the total weight of the logical clauses that it satisfies. A variety of inference methods for MLNs have been developed, however, their computational complexity is a fundamental issue.

2.4 Probabilistic Soft Logic

Probabilistic Soft Logic (PSL) is another recently proposed framework for probabilistic logic (Kimmig et al., 2012). It uses logical representations to compactly define large graphical models with continuous variables, and includes methods for performing efficient probabilistic inference for the resulting models. A key distinguishing feature of PSL is that ground atoms have soft, continuous truth values in the interval [0, 1] rather than binary truth values as used in MLNs and most other probabilistic logics. Given a set of weighted inference rules, and with the help of Lukasiewicz's relaxation of the logical operators, PSL builds a graphical model defining a probability distribution over the continuous space of values of the random variables in the model. Then, PSL's MPE inference (Most Probable Explanation) finds the overall interpretation with the maximum probability given a set of evidence. It turns out that this optimization problem is second-order cone program (SOCP) (Kimmig et al., 2012) and can be solved efficiently in polynomial time.

2.5 Recognizing Textual Entailment

Recognizing Textual Entailment (RTE) is the task of determining whether one natural language text, the *premise*, Entails, Contradicts, or not related (Neutral) to another, the *hypothesis*.

2.6 Semantic Textual Similarity

Semantic Textual Similarity (STS) is the task of judging the similarity of a pair of sentences on a scale from 1 to 5 (Agirre et al., 2012). Gold standard scores are averaged over multiple human annotations and systems are evaluated using the Pearson correlation between a system's output and gold standard scores.

3 Approach

A semantic parser is three components, a formal language, an ontology, and an inference mechanism. This section explains the details of these components in our semantic parser. It also points out the future work related to each part of the system.

3.1 Formal Language: first-order logic

Natural sentences are mapped to logical form using Boxer (Bos, 2008), which maps the input sentences into a lexically-based logical form, in which the predicates are words in the sentence. For example, the sentence "A man is driving a car" in logical form is:

$\exists x, y, z.\ man(x) \wedge agent(y, x) \wedge drive(y) \wedge patient(y, z) \wedge car(z)$

We call Boxer's output alone an uninterpreted logical form because predicates do not have meaning by themselves. They still need to be connected with an ontology.

Future work: While Boxer has wide coverage, additional linguistic phenomena like generalized quantifiers need to be handled.

3.2 Ontology: on-the-fly ontology

Distributional information is used to generate the needed part of an on-the-fly ontology for the given input sentences. It is encoded in the form of weighted inference rules describing the semantic relations connecting words and phrases in the input sentences. For example, for sentences "A man is driving a car", and "A guy is driving a vehicle", we would like to generate rules like $\forall x.\, man(x) \Leftrightarrow guy(x) \mid w_1$ indicating that "man" and "guy" are synonyms with some certainty w_1, and $\forall x.\, car(x) \Rightarrow vehicle(x) \mid w_2$ indicating that "car" is a hyponym of "vehicle" with some certainty w_2. Other semantic relations can also be easily encoded as inference rules like antonyms $\forall x.\, tall(x) \Leftrightarrow \neg short(x) \mid w$, contextonymy relation $\forall x.\ hospital(x) \Rightarrow \exists y.\ doctor(y) \mid w$. For now, we generate inference rules only as synonyms (Beltagy et al., 2013), but we are experimenting with more types of semantic relations.

In (Beltagy et al., 2013), we generate inference rules between all pairs of words and phrases. Given two input sentences T and H, for all pairs (a, b), where a and b are words or phrases of T and H respectively, generate an inference rule: $a \rightarrow b \mid w$, where the rule's weight $w = sim(\overrightarrow{a}, \overrightarrow{b})$, and sim is the $cosine$ of the angle between vectors $\overrightarrow{a}$ and $\overrightarrow{b}$. Note that this similarity measure cannot yet distinguish relations like synonymy and hypernymy. Phrases are defined in terms of Boxer's output to be more than one unary atom sharing the same variable like "a little kid" which in logic is $little(k) \wedge kid(k)$, or two unary atoms connected by a relation like "a man is driving" which in logic is $man(m) \wedge agent(d, m) \wedge drive(d)$. We used vector addition (Mitchell and Lapata, 2010) to calculate vectors for phrases.

Future Work: This can be extended in many directions. We are currently experimenting with asymmetric similarity functions to distinguish semantic relations. We would also like to use longer phrases and other compositionality techniques as in (Baroni and Zamparelli, 2010; Grefenstette and Sadrzadeh, 2011). Also more inference rules can be added from paraphrases collections like PPDB (Ganitkevitch et al., 2013).

3.3 Inference: probabilistic logical inference

The last component is probabilistic logical inference. Given the logical form of the input sentences, and the weighted inference rules, we use them to build a probabilistic logic program whose solution is the answer to the target task. A probabilistic logic program consists of the evidence set

E, the set of weighted first order logical expressions (rule base RB), and a query Q. Inference is the process of calculating $Pr(Q|E, RB)$.

Probabilistic logic frameworks define a probability distribution over all possible worlds. The number of constants in a world depends on the number of the discourse entities in the Boxer output, plus additional constants introduced to handle quantification. Mostly, all constants are combined with all literals, except for rudimentary type checking.

4 Tasks

This section explains how we perform the RTE and STS tasks using our semantic parser.

4.1 Task 1: RTE using MLNs

MLNs are the probabilistic logic framework we use for the RTE task (we do not use PSL here as it shares the problems of fuzzy logic with probabilistic reasoning). The RTE's classification problem for the relation between T and H, and given the rule base RB generated as in 3.2, can be split into two inference tasks. The first is finding if T entails H, $Pr(H|T, RB)$. The second is finding if the negation of the text $\neg T$ entails H, $Pr(H|\neg T, RB)$. In case $Pr(H|T, RB)$ is high, while $Pr(H|\neg T, RB)$ is low, this indicates Entails. In case it is the other way around, this indicates Contradicts. If both values are close to each other, this means T does not affect probability of H and that is an indication of Neutral. We train a classifier to map the two values to the final classification decision.

Future Work: One general problem with MLNs is its computational overhead especially for the type of inference problems we have. The other problem is that MLNs, as with most other probabilistic logics, make the Domain Closure Assumption (Richardson and Domingos, 2006) which means that quantifiers sometimes behave in an undesired way.

4.2 Task 2: STS using PSL

PSL is the probabilistic logic we use for the STS task since it has been shown to be an effective approach to compute similarity between structured objects. PSL does not work "out of the box" for STS, because Lukasiewicz's equation for the conjunction is very restrictive. We addressed this problem (Beltagy et al., 2014) by replacing Lukasiewicz's equation for the conjunction with an averaging equation, then change the optimization problem and the grounding technique accordingly.

	SICK-RTE	SICK-STS
dist	0.60	0.65
logic	0.71	0.68
logic+dist	0.73	0.70

Table 1: RTE accuracy and STS Correlation

For each STS pair of sentences S_1, S_2, we run PSL twice, once where $E = S_1$, $Q = S_2$ and another where $E = S_2$, $Q = S_1$, and output the two scores. The final similarity score is produced from a regressor trained to map the two PSL scores to the overall similarity score.

Future Work: Use a weighted average where different weights are learned for different parts of the sentence.

5 Evaluation

The dataset used for evaluation is **SICK**: Sentences Involving Compositional Knowledge dataset, a task for SemEval 2014. The initial data release for the competition consists of 5,000 pairs of sentences which are annotated for both RTE and STS. For this evaluation, we performed 10-fold cross validation on this initial data.

Table 1 shows results comparing our full approach (**logic+dist**) to two baselines, a distributional-only baseline (**dist**) that uses vector addition, and a probabilistic logic-only baseline (**logic**) which is our semantic parser without distributional inference rules. The integrated approach (**logic+dist**) out-performs both baselines.

6 Conclusion

We presented an approach to semantic parsing that has a wide-coverage for words and relations, and does not require a fixed formal ontology. An on-the-fly ontology of semantic relations between predicates is derived from distributional information and encoded in the form of soft inference rules in probabilistic logic. We evaluated this approach on two task, RTE and STS, using two probabilistic logics, MLNs and PSL respectively. The semantic parser can be extended in different direction, especially in predicting more complex semantic relations, and enhancing the inference mechanisms.

Acknowledgments

This research was supported by the DARPA DEFT program under AFRL grant FA8750-13-2-0026. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the view of DARPA, DoD or the US government. Some experiments were run on the Mastodon Cluster supported by NSF Grant EIA-0303609.

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of Semantic Evaluation (SemEval-12)*.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-10)*.

Islam Beltagy, Cuong Chau, Gemma Boleda, Dan Garrette, Katrin Erk, and Raymond Mooney. 2013. Montague meets Markov: Deep semantics with probabilistic logical form. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM-13)*.

Islam Beltagy, Katrin Erk, and Raymond Mooney. 2014. Probabilistic soft logic for semantic textual similarity. In *Proceedings of Association for Computational Linguistics (ACL-14)*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-13)*.

Johan Bos. 2008. Wide-coverage semantic analysis with Boxer. In *Proceedings of Semantics in Text Processing (STEP-08)*.

Stephen Clark and James R. Curran. 2004. Parsing the WSJ using CCG and log-linear models. In *Proceedings of Association for Computational Linguistics (ACL-04)*.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-13)*.

L. Getoor and B. Taskar, editors. 2007. *Introduction to Statistical Relational Learning*. MIT Press, Cambridge, MA.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-11)*.

Edward Grefenstette. 2013. Towards a formal distributional semantics: Simulating logical calculi with tensors. In *Proceedings of Second Joint Conference on Lexical and Computational Semantics (*SEM 2013)*.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic*. Kluwer.

Angelika Kimmig, Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2012. A short introduction to Probabilistic Soft Logic. In *Proceedings of NIPS Workshop on Probabilistic Programming: Foundations and Applications (NIPS Workshop-12)*.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*.

Tom Kwiatkowski, Eunsol Choi, Yoav Artzi, and Luke Zettlemoyer. 2013. Scaling semantic parsers with on-the-fly ontology matching. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-13)*.

T. K. Landauer and S. T. Dumais. 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*.

Alessandro Lenci and Giulia Benotto. 2012. Identifying hypernyms in distributional semantic spaces. In *Proceedings of the first Joint Conference on Lexical and Computational Semantics (*SEM-12)*.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical cooccurrence. *Behavior Research Methods, Instruments, and Computers*.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of Association for Computational Linguistics (ACL-08)*.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Journal of Cognitive Science*.

Richard Montague. 1970. Universal grammar. *Theoria*, 36:373–398.

Matthew Richardson and Pedro Domingos. 2006. Markov logic networks. *Machine Learning*, 62:107–136.

Peter Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research (JAIR-10)*.