www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 9 September 2014 Page No. 8012-8015

A novel approach to make NLP predictive and non-ambigious in Punjabi Language

Parneet Kaur[1], Er.Vandana Pushe[2]

[1]Department of CSE, Desh Bhagat University,
Mandi Gobindgarh, Punjab, India
parneetbenipal@gmail.com

[2]Department of CSE, Desh Bhagat University,
Mandi Gobindgarh, Punjab, India
vandanapushe@gmail.com

***Abstract*—This paper present a probabilistic generative model for learning semantic parsers from ambiguous supervision. Our approach learns from natural language sentences paired with world states consisting of multiple potential logical meaning representations. It disambiguates the meaning of each sentence while simultaneously learning a semantic parser that maps sentences into logical form. Compared to a previous generative model for semantic alignment, it also supports full semantic parsing.**

***Keywords*—Language Genertaor,Ambiguity, probabilistic generative model ,Ontology, , Distributional Semantics.**

I. Introduction

An important application of natural language processing is the interpretation of human instructions. The ability to parse instructions and perform the intended actions is essential for smooth interactions with a computer or a robot. Some recent work has explored how to map natural-language instructions into actions that can be performed by a computer (Branavan et al. 2009; Lau, Drews, and Nichols 2009). In particular, we focus on the task of navigation (MacMahon, Stankiewicz, and Kuipers 2006; Shimizu and Haas 2009; Matuszek, Fox, and Koscher 2010; Kollar et al. 2010; Vogel and Jurafsky 2010). The goal of the navigation task is to take a set of naturallanguage directions, transform it into a navigation plan that can be understood by the computer, and then execute that plan to reach the desired destination. Route direction is a unique form of instructions that specifies how to get from one place to another and understanding them depends heavily on the spatial context. The earliest work on interpreting route directions was by linguists (Klein 1982; Wunderlich and Reinelt 1982). While this domain is restricted, there is considerable variation in how different people describe the same route. Below are some examples from our test corpus of instructions given for the route shown in Figure 1:Paper proposed 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). In particular, Chen and Mooney (2008) introduced the problem of learning to sportscast by simply observing natural language commentary on simulated Robocup robot soccer games. The training data consists of natural language (NL) sentences ambiguously paired with logical meaning representations (MRs) describing recent events in the game extracted from the simulator. Most sentences describe one of the extracted recent events; however, the specific event to which it refers is unknown. Therefore, the learner has to figure out the correct matching (alignment) between NL and MR before inducing a semantic parser or language generator. Based on an approach introduced by Kate and Mooney (2007), Chen and Mooney (2008) repeatedly retrain both a supervised semantic parser and language generator using an iterative algorithm analogous to Expectation Maximization (EM). However, this approach is somewhat ad hoc and does not exploit a well-defined probabilistic generative model or real EM training.

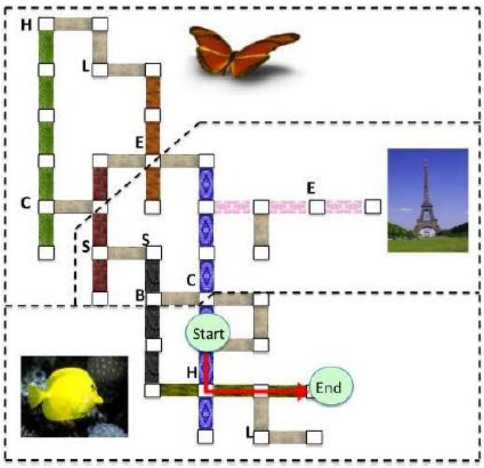


Figure 1: This is an example of a route in our virtual world. The world consists of interconnecting hallways with varying floor tiles and paintings on the wall (butterfly, fish, or Eiffel

Tower.) Letters indicate objects (e.g. 'C' is a chair) at a location

For example, in probabilistic logic, the synonymy relation between "man" and "guy" is represented by: 8x. man(x) , guy(x) | w1 and the hyponymy relation between "car" and "vehicle" is: 8x. car(x)) vehicle(x) | w2 where w1 and w1 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).

II. Related Work

The conventional approach to learning semantic parsers (Zelle and Mooney, 1996; Ge and Mooney, 2005; Kate and Mooney, 2006; Zettlemoyer and Collins, 2007; Zettlemoyer and Collins, 2005; Wong and Mooney, 2007b; Lu et al., 2008) requires detailed supervision unambiguously pairing each sentence with its logical form. However, developing training corpora for these methods requires expensive expert human labor. Chen and Mooney (2008) presented methods for grounded language learning from ambiguous supervision that address three related tasks: NL–MR alignment, semantic parsing, and natural language generation. They solved the problem of aligning sentences and meanings by iteratively retraining an existing supervised semantic parser, WASP (Wong and Mooney, 2007b) or KRISP (Kate and Mooney, 2006), or an existing supervised natural-language generator, WASP (Wong and Mooney, 2007a). During each iteration, the currently trained parser (generator) is used to produce an improved NL–MR alignment that is used to retrain the parser (generator) in the next iteration. However, this approach does not use the power of a probabilistic correspondence between an NL and MRs during training. On the other hand, Liang et al. (2009) proposed a probabilistic generative approach to produce a Viterbi alignment between NL and MRs. They use a hierarchical semi-Markov generative model that first determines which facts to discuss and then generates words from the predicates and arguments of the chosen facts. They report improved matching accuracy in the Robocup sportscasting domain. However, they only addressed the alignment problem and are unable to parse new sentences into meaning representations or generate natural language from logical forms. In addition, the model uses a weak bag-of-words assumption when estimating links between NL segments and MR facts. Although it does use a simple Markov model to order the generation of the different fields of an MR record, it does not utilize the full syntax of the NL or MR or their relationship. Chen et al. (2010) recently reported results on utilizing the improved alignment produced by Liang et al. (2009)'s model to initialize their own iterative retraining method. By combining the approaches, they produced more accurate NL– MR alignments and improved semantic parsers. Motivated by this prior research, our approach combines the generative alignment model of Liang et al. (2009) with the generative semantic parsing model of Lu et al. (2008) in order to fully exploit the NL syntax and its relationship to the MR semantics. Therefore, unlike Liang et al.'s simple Markov + bag-of-words model for generating language, it uses a tree-based model to generate grammatical NL from structured MR facts.

III. Background

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). 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).

Markov Logic Network: Markov Logic Network (MLN) (Richardson and Domingos, 2006) is a framework for probabilistic logic that employ weighted formulas in firstorder 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.

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. 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.

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.

IV. 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 semantic parser. It also points out the future work related to each part of the system.

Generative Model

Like Liang et al. (2009)'s generative alignment model, our model is designed to estimate P(w|s), where w is an NL sentence and s is a world state containing a set of possible MR logical forms that can be matched to w. However, our approach is intended to support both determining the most likely match between an NL and its MR in its world state, and semantic parsing, i.e. finding the most probable mapping from a given NL sentence to an MR logical form.

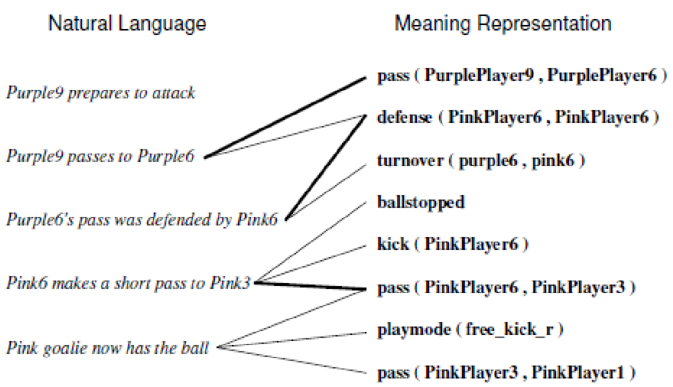
Our generative model consists of two stages:

• Event selection: P(e|s), chooses the event e in the world state s to be described.

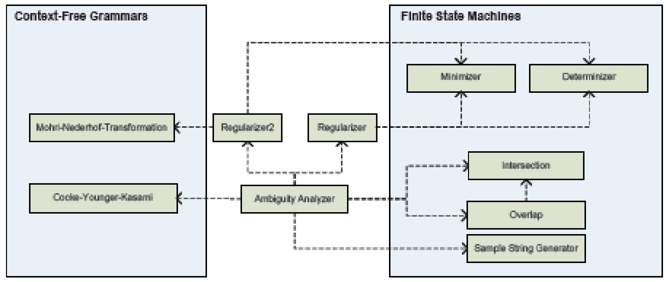
• Natural language generation: P(w|e), models the probability of generating natural-language sentence w from the MR specified by event e.

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.



Algorithms and their dependencies in *grambiguity*

input Navigation instructions and the corresponding navigation
plans (e_1, p_1), . . . , (e_n, p_n)
output Lexicon, a set of phrase-meaning pairs
main
for n-gram w that appears in e = (e_1, . . . , e_n) **do**
for instruction e_i that contains w **do**
Add navigation plan p_i to meanings(w)
end for
repeat
for every pair of meanings in meanings(w) **do**
Add intersections of the pair tomeanings(w)
end for
Keep k highest-scoring entries ofmeanings(w)
until meanings(w) converges
Add entries of meanings(w) with scores higher than threshold t to Lexicon
end for
end main

V. Conclusion

We have presented a novel system that learns a semantic parser for interpreting navigation instructions by simply observing the actions of human followers without using any prior linguistic knowledge or direct supervision.We demonstrated the need to model landmarks when executing longer, more complex instructions.We also introduced a plan refinement algorithm that fairly accurately infers the correct navigation plan specified in the instructions by using a learned semantic lexicon to remove extraneous information. Overall, our approach demonstrates an interesting and novel form of grounded language learning for a complex and useful task.

VI. References

[1]. 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).

[2]. 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).

[3]. Islam Beltagy, Cuong Chau, Gemma Boleda, Dan Garrette, Katrin Erk, and Raymond Mooney. 2013.

[4]. Montague meets Markov: Deep semantics with probabilistic logical form. In Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM-13).

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

[6]. 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).

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

[8]. 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).

[9]. 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).

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