Concept Learning from Natural Language Interactions

Shashank Srivastava

October 16, 2017

Abstract

Humans can efficiently learn about new concepts using natural language communications. For example, a human can learn the concept of a phishing email from natural language explanations such as ‘phishing emails often request your bank account number’. On the other hand, purely inductive learning systems typically require a large collection of labeled data for learning such a concept. If we wish to make computer learning as efficient as human learning, we need to develop methods that can learn from natural language interactions.

Learning from language presents two key challenges. The first is that of learning from interpretations, which refers to the mechanisms through which interpretations of language statements can be used to solve learning tasks in the environment. The second is the basic problem of learning to interpret language, which refers to an agent’s ability to map natural language explanations in pedagogical contexts to formal semantic representations that computers can process and reason over. We address aspects of both these problems, and provide an interface for guiding concept learning methods using language.

For learning from interpretation, we focus on concept learning (binary classification) tasks. We demonstrate that language can formulate concept learning tasks by defining rich and expressive features (e.g., ‘Does the email ask me to click a hyper-link?’), and show that methods for concept learning can benefit substantially from such explanations. We propose to address assimilation of language cues that implicitly constrain models for concept learning (e.g., ‘Most emails are not phishing emails’). In particular, we focus on quantifier expressions (such as usually, never, etc.) that reflect generality of specific observations, and can be incorporated into training of classification models. We also propose to explore the use of natural language to interactively engage with a teacher in a mixed-initiative setting to reduce the sample complexity of concept learning.

Apart from developing computational machinery that uses interpretations of language advice to guide concept learning, we develop complementary algorithms for learning to interpret language by incorporating different types of situational context, including conversational history and sensory observations. We show that environmental context can enrich models of semantic interpretation by not only providing discriminative features, but also reducing the need for expensive labeled data used for training them.
1 Introduction

Humans routinely use language to learn about new concepts in the environment. Verbal and written language form the core for much of human learning and pedagogy, as reflected in textbooks, lectures and student-teacher dialogues. In contrast, machine learning systems traditionally learn from passive collections of labeled examples. Not only is this paradigm unnatural compared to how a human might teach another human, but there are inherent statistical limitations on what can be learned from labeled examples alone. For example, the number of examples needed by an inductive learner coarsely scales as the log of the size of the hypothesis space. This can be intractable even for representations such as ontologies, which children learn with relatively few examples. A central reason for this discord is that for the most part, machine learning has ignored richer forms of input including explanations and clarifications, which can have significant consequences for learning research. If we wish to make computer learning as efficient as human learning, we need to develop methods that can learn from natural language interactions.

On the other hand, interpreting language is hard. It requires resolution of linguistic ambiguities as well as an understanding of non-linguistic situational context. From a practical perspective, obtaining data consisting of statements paired with their semantic interpretations is expensive. Hence, there is a need for NLP methods to expand an awareness of the environment, as well as to drive training of semantic parsers from indirect forms of supervision.

In this proposal, our goal is to develop algorithms that can learn through natural language interaction. Learning from language presents two key challenges: (a) learning from interpretation, which refers to using interpretations of language to learn and reason about the environment, and (b) learning to interpret, which refers to the primary ability to interpret language in situated contexts. We address aspects of both these problems, and provide a basic conceptual interface for guiding machine learning algorithms from language.

In general, learning can refer to a variety of things. For example, it may refer to developing the capacity to play a game, memorizing an alphabet, performing a procedure, etc. Here, we aim to learn function approximations using language. In particular, we focus on concept learning tasks. Similarly, by language interpretation we will refer to semantic parsing, i.e. decoding natural language sentences to logical representations grounded in a domain language.

Thesis statement: Machines can learn concepts in the environment from natural language explanations, in ways similar to a human. In particular, machines can use explanations to formulate concept learning tasks by defining rich and expressive features, constrain classification models by leveraging declarative knowledge, and interactively engage with humans to simplify their learning. At the same time, the ability to automatically interpret language can be guided by incorporating different types of non-linguistic contexts. Environmental context can not only enrich language interpretation by providing discriminative features,
but also reduce the need for expensive labeled training data.

1.1 Framework

Figure 1 shows a general framework for learning from language. An agent consists of a language interpreter and a task learner, and can interact with the external environment in two ways. The agent can receive input from the environment in the form of (1) observations through sensors, as well as (2) natural language instruction from a teacher. In turn, an agent can affect the environment by performing actions, or asking questions.

Formally, an agent can sense the state of the environment $e$ through a set of sensors $s$. At any time, the agent can access a set of sensor observations $s(e)$. A teacher, who is part of the environment, can provide natural language advice of two types. Firstly, the teacher can describe how to compute the quantities necessary for learning the target concept in terms of groundable sensor observations, $s(e)$. In terms of the learning problem, this corresponds to defining the features of a learning problem, $X = g(s(e))$. The learning problem then consists of finding a function approximation $Y = f(g(s(e)))$ to predict a concept in the environment, based on sensor readings $s(e)$. A second type of natural language advice can directly describe the target function to be learned, which we interpret as constraints on the model. Thus, the scope of knowledge conveyed from language explanations can refer to the sensor observations to use ($s(e)$), the reasoning or computation required to infer variables of interest ($g$), and
the mapping from those variables to the target concept \((f)\). We illustrate this further through an example in the next section.

The upper and lower halves of Figure 1, representing the direction of influence between language and the environment, correspond to the key challenges of ‘learning to interpret’ and ‘learning from interpretation’ that we described earlier. In general, an agent can also refer to a personal knowledge base of background knowledge to drive its learning, and learning can modify the knowledge base with new facts. While background knowledge is an important component of human learning and reasoning, we do not delve into this aspect in this work.

1.2 A Concrete Example

Let us consider an example to concretely understand the problems involved in learning from natural language interactions. Figure 2 shows a hypothetical scenario, where a physician explains the concept of patients at risk of heart attack to an agent. The interaction exemplifies some common facets of learning from language, including explanations, clarifications, and declarative transfer of knowledge.

**Learning task: Identifying risk of heart attack**

| Physician: I want you to identify patients in our ward who are at risk of heart attack |
| Learner: OK, Doc. What indications should I look for? |
| Physician: Check their LDL cholesterol levels |
| Learner: I’ll run a test. What else? |
| Physician: Look for a history of diabetes |
| Learner: I’ll check if there’s a record |
| Physician: Also, see if they exercise more than thrice a week |
| Learner: Will do that |
| Physician: Roughly about 1 in 5 of our patients is at risk of heart attack |
| Learner: OK |
| Physician: Male patients who are above 65 and have a BMI more than 30 are definitely at risk |
| Learner: I’m sorry, but I don’t know how to get a patient’s BMI |
| Physician: It’s the weight in kilos divided by the square of the height, in meters |
| Learner: Thanks! Can you provide some case-histories of patients who developed heart-disease in our records? |
| Physician: Yes, have a look |
| Learner: Is this new patient vulnerable? |
| Physician: Very good! You seem to have a hang of it now |

Figure 2: Hypothetical scenario of learning from natural language interaction.

Natural language instruction can encompass a very wide range of useful information, e.g., advice on when to learn, which teachers to trust, what questions
to ask during exploratory or confirmatory phases of learning, etc. Thus, it is very hard to have a definitive formulation for learning from language. However, from the perspective of machine learning, language can be especially valuable at three specific points in the learning process, beginning with formulating learning tasks. Most successful applications of machine learning rely on a process of careful feature engineering, even before any learning takes place. This involves design of features on part of the data analyst, which defines the variables on which the learning problem is based. Instead of hand-coding programs to extract such features, learning from natural language instruction can expand the scope of computer systems to learn from people with no machine learning expertise (such as the physician in this example).

Once a learning problem is formulated, language can guide the learning process itself through direct transfer of declarative knowledge about the task to be learned, which may be hard for a machine to learn by itself. Such knowledge can improve generalization by preferring models that concur with human provided advice. In our example, an intelligent learner can use such knowledge (in the form of the physician’s explanation) to avoid models that may predict a majority of patients to be at risk of heart attack. In this sense, language can itself be a source of supervision by constraining learning models through declarative knowledge. This can simplify learning since model constraints can guide learning towards more promising parts of the hypothesis space.

A third way in which language can facilitate learning is through interactive dialog on part of the agent. As seen in the current example, this may takes multiple forms: asking for (1) a clarification of a previous explanation, (2) seeking labeled examples, or (3) validating a learned model through feedback.

At the same time, in order to be able to learn from language, the learner needs the ability to interpret instructional language as feature definitions, different types of model constraints, and data measurements. For example, we assume that a statement such as ‘Check if their LDL cholestrol level is less than 150’ can be mapped to a logical form such as lessThan( getValueForField(LDL), 150), which can be executed as a query or a program against an existing database to retrieve the value of a feature. A corollary of this grounding is that the scope of meanings that can be conveyed through natural language is determined by the predicates in the domain logical language. Learning to map such instructional language to actionable logical representations is challenging, especially since creating data consisting of language statements paired with their semantic annotations is expensive.

In this section, we highlighted some of the principal sub-problems involved in concept learning from language. In this thesis, we will develop methods to address some of these. An important observation that we reiterate here is that learning tasks often depend on background knowledge, which we do not attempt to model here. For example, the physician might say that an important indicator for the risk of heart attack is the peak oxygen consumption rate \( \text{VO}_2 \text{max} \) for an individual. We may not be able to measure this value directly, but we may know that this is related to an individual’s resting heart rate (which we likely can observe), and use that as a proxy feature. Ideally, an agent would be able to reason and use background knowledge to infer such relevant features.
However, for sake of simplicity, we do not consider the use of such background knowledge in the current work.

1.3 Summary of Proposed contributions

The proposed contributions of this thesis work can be categorized in the following two directions (also depicted in Figure 1).

(a) Learning from Interpretation: When humans learn, they rely on rich forms of supervision, including explanations, examples and interactive dialogue for efficient learning. In contrast, pure inductive learning (the predominant paradigm in machine learning) is fundamentally limited by its dependence on passively acquired big data. We suggest that learning from language is a viable paradigm for automated systems, which presents the following advantages that can enable more efficient learning.

1. Language can be used to naturally formulate learning problems by describing rich and compositional features. By formulation, we refer to defining the relevant variables (features) in whose terms the output of the problem (label) needs to be learned. The value of these features may involve simple lookups in a data-base, or sequences of actions. For example, in predicting the risk of heart attack, a doctor can say: ‘Is the patient’s BMI more than 30?’. Such a question can parsed by a language interpreter to a structured query, and answered from a patient’s health record, thus defining an active new attribute to be considered for each new patient.

2. Natural language explanations can be used to constrain machine learning models, minimizing the need for labeled data. For example, everyday language contains quantification expressions (such as ‘all’, ‘some’, ‘rarely’, ‘usually’, etc.) that are explicit denoters of generality. Similarly, natural language often conveys explicit declarative knowledge about a domain that may be hard to learn inductively. Such explanations can be used to guide training of machine learning methods, by constraining models to emulate the teacher’s advice.

3. Language allows a natural medium for interactive dialog on the part of the agent. We will explore how this can be leveraged by an agent for improving concept learning performance by (1) seeking labels for specific examples, (2) validating predictions from a model, (3) asking clarifying questions for filling an information gap.

We propose to demonstrate that the above-mentioned aspects of natural language can be leveraged by machine learning algorithms to improve performance on concept learning tasks.

(b) Learning to Interpret: On the other hand, to learn from open-ended natural language explanations, we must develop better computational models for language interpretation in situated contexts. The grounding of language in the situated environment is a fundamentally valuable property, which makes it
useful above abstract symbol systems (Harnad, 1990). We show that models for language interpretation can benefit from leveraging this grounding in two ways.

4. Statistical models of language interpretation can employ *discriminative features* that depend on environmental context to improve parsing performance. Such features can capture associations between language and environmental context. For example, consider the following conversation sequence: ‘Physician: Check the BMI. Learner: OK. Physician: It should be less than 30.’ Here, the meaning of the last sentence cannot be inferred from its text alone, but this can be facilitated by a feature that keeps track of the subject in previous conversation.

5. We argue that in many scenarios, awareness of the situational context can itself *provide weak supervision* for training of semantic parsers. This can alleviate the need for human annotation, which is a major bottleneck in semantic parsing. For example, the statement ‘How high is LDL?’ can possibly refer to the LDL cholesterol level for a patient, or the stock price of a company (Lydall Inc). However, if we know that the statement was made by a physician in a hospital, pragmatics can guide us towards the correct interpretation, even if the semantic label is not provided.

In this thesis, we present novel algorithms for semantic parsing that incorporate two types of context: (i) conversational history, and (ii) sensory observations.

**Datasets:** We plan to collect and annotate data sets for exploring the above aspects of language interpretation and learning. Many of the problems we focus on are novel, and require creation of resources for their analysis and evaluation. Providing benchmark corpora for this new direction can facilitate further research towards natural language interfaces to machine learning.

We also plan to integrate ideas from different parts of the thesis work into a working demonstration that can interactively learn concepts using natural language explanations from human users.

The remainder of this proposal is organized as follows. In Section 2, we briefly review some background literature in three relevant areas: (i) inductive learning, (ii) statistical semantic parsing, and (iii) learning from language. In Section 3, we describe our proposed work and methods for formulating learning tasks and training supervised learning models from natural language explanations in detail (corresponding to itemized points 1 and 2 respectively), and present some preliminary results. We also discuss some possible directions for exploring learning through natural language interactions using mixed initiative dialogue (itemized point 3). In Section 4, we outline some directions for learning to interpret language using pragmatic cues from the environment (corresponding to itemized points 5 and 6) along with some empirical results. Finally, Section 6 provides a rough timeline of the thesis work.
2 Background and Related Work

In this section, we briefly overview three areas of previous work that are germane to this proposal document. We start with a brief outline of inductive learning methods that incorporate additional user-provided side information. We then present a brief summary of statistical semantic parsing methods, and explorations in learning from natural interactions.

**Inductive learning:** Inductive learning from labeled data has been widely successful in several domains, and has been a dominant focus of research in supervised learning (Caruana et al., 2008). However, this paradigm has some inherent drawbacks. Firstly, labeled training data may not be easily available for a wide range of possible learning problems. Secondly, with the absence of any background knowledge or biases to guide learning, the learner can only be as good as the data it receives. This makes inductive learning susceptible to spurious generalizations due to sample biases in the training data. Many notable approaches have explored incorporation of prior background knowledge in learning. These include the Generalized Expectation (Mann and McCallum, 2010) and Posterior Regularization (Ganchev et al., 2010) frameworks that integrate manually provided ‘side-information’ (feature and label constraints) to guide machine learning models. Earlier work on Explanation-based learning (Mitchell et al., 1986; DeJong and Mooney, 1986) leverages structured knowledge to analyze why an example belongs to a concept using a provided domain theory. Blum (1990) and Roth and Small (2009) explore learning of classification tasks through interactive supervision in presence of arbitrarily large feature spaces. Recent work such as by Lake et al. (2015) explores concept learning from few examples in limited settings, and presents encouraging results for one-shot learning by learning representations of instances over Bayesian programs. To the best of our knowledge, none of these approaches tackle the issue of concept learning from language, i.e., converting natural language supervision into a format suitable for training machine learning methods.

**Semantic parsing:** Semantic parsing refers to automatic mapping of natural language sentences to a domain-specific *logical form* that represents its meaning. Statistical methods for semantic parsing usually rely on approaches from structured prediction, and have been explored in diverse domains (Zelle and Mooney, 1996; Artzi and Zettlemoyer, 2013). While semantic parsers have traditionally relied on labeled datasets of statements paired with labeled logical forms (Zettlemoyer and Collins, 2005), recent approaches have focused on training semantic parsers from denotations of logical forms, rather than logical forms themselves (Krishnamurthy and Mitchell, 2012; Berant et al., 2013). Some of the methods in this proposal extends this paradigm by attempting to learn from still weaker signals, which have not been previously explored. The role of context in assigning meaning to language has been emphasized from abstract perspectives in computational semantics (Bates, 1976; Van Dijk, 1980), as well as in systems for task-specific applications (Larsson and Traum, 2000). Examples of the former include analyzing language from perspectives of speech acts (Searle, 1969)
Learning from Natural Interactions: The connection between language and human learning has been deeply explored from the perspective of linguistic and cognitive theories (Halliday, 1993; Lemke, 1990). Some statistical learning approaches have used different kinds of supervisory signals to guide language interpretation in context of performing external tasks (Liang et al., 2009; Branavan et al., 2009). Parsing natural language instructions from user manuals have been studied in game playing frameworks by Branavan et al. (2012); Eisenstein et al. (2009). Our work is also related to work by Goldwasser and Roth (2014); Clarke et al. (2010), who train semantic parsers in weakly supervised contexts, where language interpretation is integrated in real-world tasks, such as learning the rules of solitaire. Other approaches have explored the use of language input for tasks such as QA (Sukhbaatar et al., 2015), without explicitly modeling the process of semantic interpretation. More generally, the idea of learning through human interactions has been explored in several settings such as behavioral programming (Harel et al., 2012), natural language programming (Biermann, 1983), learning by instruction (Azaria et al., 2016), etc.

3 Learning from Interpretation

As discussed in Section 1, there can be several ways through which natural language can aid learning about the environment. In this section, we describe possible directions that we might explore in more detail:

3.1 Formulating learning tasks using NL

Status: Completed
Reference: Srivastava et al. (2017b)

The process of feature design transforms raw observations about the data into meaningful inputs that are provided to a learning algorithm. We suggest that language can be an natural medium for communicating such information, and present an approach that uses NL explanations to define expressive features that can characterize concepts. For example, learning the concept of ‘negative product reviews’ can benefit from a NL explanation such as ‘Negative reviews often mention phrases such as “poor quality”, “too expensive” and “disappointed”’. Such a statement can be interpreted as a feature function that will have a non-zero value if a product review contains the word ‘disappointed’.

Method: The main idea here is that semantic parsing can be used to map natural language descriptions to logical forms, which can denote feature functions, and can be evaluated in context of different instances to yield the value of a feature. Interpreting explanations as such attribute definitions can yield a set of relevant and informative features. A discriminative classifier such as a log-linear
Figure 3: NL explanations can express rich and expressive feature definitions

model can be used on top of this formulation, and can be used to learn concept definitions in terms of these user-specified features. This pipeline is shown in Figure 3, and proceeds as follows. A user describes a concept such as ‘phishing emails’ using language explanation. These are parsed by a semantic parser to a logical forms, which are in turn evaluated in the context of new instances to yield a feature vector. As mentioned, a discriminative classifier is then trained on this representation of the data.

**Data:** We created a dataset of 1,030 emails paired with 235 natural language statements made by human users in the process of teaching a set of seven email-related concepts. The dataset was collected using the Amazon Mechanical Turk crowd-sourcing platform. We deployed two tasks in sequence. First, a Generation task required workers to create emails belonging to different email categories. These included concepts such as reminder emails, meeting invitations, requests from boss, internet humor, going out with friends, policy announcements, etc. Next, a Teaching task showed workers emails from a concept, and asked them to describe these concepts in their own language. The final data contains between 30 and 35 statements describing each category. While the data-collection process was general and includes several types of signal for learning, here we specifically focus on whether a machine can learn from NL feature definitions provided by users, while ignoring other aspects.

**Results:** We show that using open-ended language explanations for formulating concept learning tasks can lead to significantly better concept learning performance than using traditional classification methods, especially in the low data regime. Figure 6 shows the average F1 score for concept learning using
These emails usually close with a name or title.
Some reminders will have a date and time in the subject.
The body of the email may say funny, picture, or internet.
Messages to friends sometimes have jpg attachments.
Emails from a public domain are not office requests.

Table 1: Examples of explanations collected from the Teaching task.

<table>
<thead>
<tr>
<th>Explanation</th>
<th>Teaching task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1</td>
<td>Example 2</td>
</tr>
<tr>
<td>Example 3</td>
<td>Example 4</td>
</tr>
</tbody>
</table>

Figure 4: Teaching task used to collect NL explanations describing concepts.

Figure 5: Concept learning performance for learning from natural language (LNL) vs a logistic classifier on a bag-of-words (BoW) model.

our approach for learning from natural language (LNL) compared to a logistic classifier based on a bag-of-words representation. We observe that our approach consistently outperforms the bag-of-words model (BoW), and requires fewer examples to reach near optimal performance before it plateaus.

Summary and future work: We presented an approach for converting natural language explanations to executable feature functions via semantic parsing, which can be evaluated on new instances of a concept. Through a user study, we empirically demonstrated that human users’ explanations of everyday concepts can be leveraged by this approach to yield better concept learning performance than traditional machine learning classifiers in the domain of personal emails. In ongoing work, we plan to evaluate this method on other domains, and in applications such as information extraction.
3.2 Constraining learning models using NL

*Status: Proposed*

Traditional supervised learning requires large quantities of labeled examples for generalization. This is problematic as obtaining labeled data for each learning task is unfeasible. On the other hand, language is often rich in declarative knowledge, which can reduce the sample complexity of learning. In particular, we will focus on *leveraging quantifier expressions and conditional rules* expressed in language to provide supervision for concept learning tasks.

Everyday language is rich in quantifier expressions (such as determiners like ‘all’, ‘each’, ‘few’, and frequency adverbs like ‘always’, ‘usually’, ‘never’), which are explicit denoters of generality. Since learning is largely synonymous with generalization, it is natural to use such signals to expedite learning. Even as traditional logic has studied the simplest of such quantifiers ($\forall$ and $\exists$), statistical models have not yet leveraged their predictive potential. Similarly, humans often teach and learn using instructional language in the form of conditional rules (e.g., ‘If the subject of an email threatens to close an account, it is definitely spam’). By identifying the types of constraints such explanations express, and using this information in computational frameworks, we can minimize the labeled examples needed. In particular, such knowledge may potentially enable *zero-shot learning* of concepts, without any labeled examples. We propose to develop methods that can utilize such linguistic expressions as supervision to drive training of machine learning algorithms with less data than traditional methods.

![Figure 6: NL constraints can reduce sample complexity of learning](image)

**Data:** Collection of NL explanations for this part will follow the same crowdsourcing methodology as in the previous section. However, we plan to experiment with data from different domains (including synthetic data to analyze linguistic usage patterns by human teachers in controlled settings).

**Methods:** The first step towards incorporating such declarative knowledge is to develop language interpretation techniques that can map language explanations to different types of model constraints in terms of observable features and data labels. For example, the statements ‘Spam emails often mention Viagra’ and ‘Emails that mention Viagara are always spam’ represent different types of constraints. In the next step, we need to develop training algorithms that can incorporate such constraints. The *Posterior Regularization paradigm*...
presents a flexible and overarching framework, which may be used to learn probabilistic models that follow a wide range of pre-defined constraints. Similarly, constrained conditional models are another class of methods that may be used to incorporate several types of model constraints.

**Evaluation:** We plan to experiment with extending such frameworks to learn concept models using natural language explanations with limited labeled data. The principal question we ask is whether constraining machine learning models using such constraints can enable learning with significantly fewer examples. We also propose to perform parallel user studies to gauge the extent to which such cues (such as quantifier expressions) in natural language explanations are helpful in learning among human subjects.

**Summary and future work:** We propose to explore integration of language cues, which provide quantitative constraints on machine learning models. In particular, we are interested in exploring whether quantifier expressions and conditional rules in everyday language can significantly reduce the sample complexity of concept learning by serving as a feasible mode of supervision.

### 3.3 Mixed-initiative learning using NL

*Status: Proposed*

Human language is inherently interactive, and this can be leveraged by an automated agent to engage users to learn more efficiently. Interactivity on the part of the learner can expedite multiple facets of learning, including both the previously discussed ones.

- In terms of formulating a learning task, an agent can use language to seek compositional explanations. For example, in our motivating example in Section 1, the physician refers to BMI during her/his explanations. Since the concept of ‘BMI’ was not recognizable to the learner, it asked the physician to ground it in terms of still simpler concepts, namely a patient’s height and weight, which the learner could understand. This involves two competencies on the part of the learner: (i) recognizing unfamiliar phrases and concepts that need further elaboration/teaching. (ii) invoking learning procedures to learn new auxiliary concepts, and use these to bootstrap learning of complex concepts.

- In terms of model learning, an agent can initiate dialog to solicit various kinds of measurements to simplify a learning problem. These include seeking of instance labels (e.g., ‘This email is spam’), feature labels (‘Emails that mention “Viagura” are spam’), label proportions (‘Spam emails are as common as non-spam email’), constraints on model expectations, etc. Thus, the agent needs to understand the different types of constraints and measurements that learning algorithms can accept, as well as which to seek during the learning process, and how to do this via language. A complete understanding of learning from language would require exploring the
role of such mixed initiative dialogues. We propose to: (1) analyze the value of such categories of measurements from human teachers through ablation studies, and (2) explore whether such interactive dialog on the part of the learner can significantly improve the downstream performance on a learning task.

**Summary and future work:** We propose to explore the role of mixed-initiative interaction between a learner and a human teacher in terms of the following two questions. First, what kinds of clarifications and explanations about the data are most useful in terms of concept learning. Second, whether proactive dialog on part of the learner can demonstrably improve performance on concept learning tasks.

### 4 Learning to Interpret

The previous section presented methods to solve learning tasks in the environment through information conveyed in NL statements (*learning from interpretation*). We now focus on the complementary prerequisite problem of *learning to interpret* language in situated contexts. In this section, we present new algorithms that assist semantic interpretation of language by incorporating contextual cues from the environment. In particular, we will focus on two types of context: conversational history, and sensory context.

#### 4.1 Language Interpretation from conversational context

**Status:** Completed  
**Reference:** Srivastava et al. (2017a)

Semantic parsers usually focus on understanding the semantics of a single sentence at a time. However, understanding conversations requires an understanding of conversational context and discourse structure across sentences. Figure 7 shows a real conversation between a human user and a digital email assistant. Observe that utterance U2 (‘Set to mom’s email’) cannot be parsed based on its content alone, but requires an understanding of the discourse. Based on the previous statement, setting the recipient field is the user’s likely intent. Similarly, utterances U3 and U4 show a repetition, where the agent first misinterprets (U3), and then correctly parses a statement (U4). Correctly interpreting the sentence requires modeling of the discourse structure of the conversation.

**Method:** We present an approach for semantic parsing by posing conversations as sequences of utterances to model ‘flow of discourse’. We consider the problem as a structured prediction, where we jointly learn preferences for collective assignments of logical forms for sentences in a sequence. Let $s = \{s_1 \ldots s_T\}$ denote a conversation sequence of $T$ utterances by a user. Let $l := \{l_1 \ldots l_T\}$ be the corresponding logical forms. We assume a supervised setting where we have labeled training sequences $\mathcal{T} := \{(s^{(1)}, l^{(1)}) \ldots (s^{(N)}, l^{(N)})\}$ consisting of utterances and their logical forms. Our model utilizes this sequential structure to incorporate information about discourse and pragmatics. We also associate
[U1] User: Compose an email
[A1] Agent: Composing new email. Its fields are: recipient, subject and body
[U2] User: Use mom’s email
[A2] Agent: I set the recipient field to ‘momthebest@email.com’
[U3] User: Set the subject to hello, please
[A3] Agent: I set the subject field to ‘hello, please’
[U4] User: Set the subject to hello, please
[A4] Agent: I set the subject field to ‘hello’
[U5] User: Set the body field to ‘I’ll be late’
[A5] Agent: I set the body field to ‘I’ll be late’
[U6] User: Send the email
[A6] Agent: I sent the email ...

Figure 7: Example of an interaction between a human user and an automated email assistant agent (Srivastava et al., 2017a).

a latent categorical variable denoted as $z_t$ with each user utterance $s_t$ to reinforce the modeling of the flow of discourse (see Figure 8). For a given utterance sequence $s = \{s_1 \ldots s_T\}$, our model predicts logical assignments, $\hat{l} = \{\hat{l}_1 \ldots \hat{l}_T\}$, and latent discourse states, $\hat{z} = \{\hat{z}_1 \ldots \hat{z}_T\}$ by finding the highest scoring assignment of logical forms, $\hat{l}$, and latent discourse states, $\hat{z}$, under a given model:

$$\langle \hat{l}, \hat{z} \rangle = \arg\max_{l \in \mathcal{L}(s), z} w^T \phi(s, l, z)$$

Figure 8: Traditional semantic parsing features depend on utterances $s_t$ and associated logical forms $l_t$ only. We allow structured features that can depend on previous logical forms $l_{t-1}$, latent variables $z_t$ representing the discourse state of the conversation at any step, and the previous utterances $s_1 \ldots s_t$.

Parameters $w$ can be trained via the latent Structured Perceptron algorithm.

**Data** We created a new dataset for semantic parsing of natural conversations, consisting of 113 real-life sequences of interactions of human users with an automated email assistant. The data contains 4759 natural language statements paired with labeled logical forms, obtained by annotating transcripts of interactions between users and an email assistant agent from Azaria et al. (2016).

**Results** We demonstrate that using conversational context yields significant improvements in parsing performance. Table 2 shows the performance of varia-
<table>
<thead>
<tr>
<th>Previous methods</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unstructured CCG</td>
<td>51.9</td>
</tr>
<tr>
<td>Seq2Seq</td>
<td>52.3</td>
</tr>
<tr>
<td>LEX</td>
<td>46.4</td>
</tr>
<tr>
<td>Our method (variants)</td>
<td></td>
</tr>
<tr>
<td>SPCon</td>
<td>54.2</td>
</tr>
<tr>
<td>SPCon + PMI</td>
<td>56.2</td>
</tr>
<tr>
<td>SPCon + Prob</td>
<td>56.9</td>
</tr>
<tr>
<td>SPCon + MF</td>
<td>62.3</td>
</tr>
<tr>
<td>SPCon + PP</td>
<td>59.8</td>
</tr>
</tbody>
</table>

Table 2: Test accuracies on Email Assistant dataset

We presented a structured prediction formulation for semantic parsing with conversational context (SPCon). Our baselines include: (i) Unstructured CCG: Traditional CCG parser (following Zettlemoyer and Collins (2005)) which uses the same lexicon and text-based features, but does not incorporate structural features (ii) Seq2Seq: Neural network based on sequence-to-sequence RNN model from Bahdanau et al. (2015), which directly maps utterances to logical forms (iii) LEX: Alignment-based model that chooses parses using a free grammar and lexical trigger scores.

Summary: We presented a structured prediction formulation for semantic parsing that incorporates conversational context by leveraging structural regularities in conversation sequences. This enables joint modeling of traditional text-based features, as well as structural features capturing the flow of discourse.

4.2 Language Interpretation from sensory context

Status: Completed, but may be extended
Reference: Srivastava et al. (2017b)

We argue that environmental context can not only provide richer features for supervised semantic parsing (as in Section 4.1), but can itself provide supervision for the training of semantic parsers. For this, we demonstrate an application where sensory context from the environment assists language learning, integrated in a real world concept learning task. In particular, we go back to the concept learning setting in Section 3.1). However, rather than presupposing a semantic parser trained from pairs of NL statements with their logical forms, we show that simple sensory context from the environment can be a source of supervisory signal for driving semantic parsing of situated language.

Method: We address the task of learning concepts from natural language statements with a small number of labeled examples. We map statements (such as ‘These emails contain a pdf file usually’) to logical interpretations (such as stringMatch(attachment stringVal ('pdf'))). These interpretations can be evaluated in context of a data instance \(x\) to return a binary value. Thus, each
natural language statement $s$ acts as a binary feature function $\{f_s(x) \in \{0,1\}\}$ that fires when the interpretation of a statement $s$ is true for a data instance $x$. The crux of our approach is that correct interpretations of natural language statements should be useful in discriminating instances of concepts. This assumption can be leveraged in the form of distant supervision, where the objective is to learn a parser that yields a discriminative classifier. e.g., a semantic parser may associate multiple incorrect interpretations with the statement above (such as `stringMatch(attachment stringVal ('usually'))`), which are unlikely to help in identifying instances of the concept. This signal can be used to guide the learning of a parser, while eventually facilitating concept classification.

**Evaluation:** We evaluate the parsing performance of our approach, which learns a semantic parser from only concept labels of examples. Table 5 evaluates parsing performance against the gold annotation logical forms for statements. In the table, full supervision refers to traditional training of a semantic parser using complete annotations of statements with their logical forms Zettlemoyer and Collins (2007). The results demonstrate that while not comparable to supervised parsing, this approach is relatively effective in learning semantic parsers with very weak supervision (concept labels of examples only).

**Summary:** We demonstrated an application where sensory context from the environment can be used to weakly supervise the training of a semantic parser, in absence of usual semantic annotations (sentences paired with labeled logical forms). For this, we leveraged pragmatics to prefer interpretations of natural language explanations that are more discriminative in context of concept learning. Our empirical results show that such context may be a viable and

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully Supervised (ZC07)</td>
<td>0.63</td>
</tr>
<tr>
<td>LNL-LR</td>
<td>0.30</td>
</tr>
<tr>
<td>LNL-NB</td>
<td>0.28</td>
</tr>
<tr>
<td>No training</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Table 3: Semantic parsing performance for our weakly supervised methods (learning from sensory context) vs full supervision (labeled logical forms)

Figure 9: Discriminative interpretations of natural language sentences are more likely to be correct. In context of concept learning, this can be judged by grounding a natural language statement in terms of a sensory context (email).
inexpensive source of supervision for training semantic parsers.

5 Integration in a working system

We plan on integrating aspects from the different parts of this thesis into a working demo, which exhibits abilities to interactively learn new concepts using natural language advice from users. A working demonstration such as this could be a proof-of-concept for some of the ideas proposed in this thesis, would help evaluate them in terms of practical utility, and identify further questions that future work could address. A possible use-case for such a system might be the Yahoo! In Mind mobile platform, where such a language based learner could help users in tasks like categorizing their email.

6 Timeline

Oct 2017: Thesis proposal

Dec 2017: Data collection and completion of ongoing work on constraining ML models using NL quantification.


May 2018: Integration of different components from this dissertation research into a working demo.


References


