An Annotated Corpus for Machine Reading of Instructions in Wet Lab Protocols

Abstract

We describe an effort to annotate a corpus of natural language instructions consisting of 622 wet lab protocols to facilitate automatic or semi-automatic conversion of protocols into a machine-readable format and benefit biological research. Experimental results demonstrate the utility of our corpus for developing machine learning approaches to shallow semantic parsing of instructional texts. We make our annotated Wet Lab Protocol Corpus available to the research community.

Introduction

As the complexity of biological experiments increases, there is a growing need to automate wet laboratory procedures to avoid mistakes due to human error and also to enhance the reproducibility of experimental biological research BIBREF0 . Several efforts are currently underway to define machine-readable formats for writing wet lab protocols BIBREF1 , BIBREF2 , BIBREF3 . The vast majority of today's protocols, however, are written in natural language with jargon and colloquial language constructs that emerge as a byproduct of ad-hoc protocol documentation. This motivates the need for machine reading systems that can interpret the meaning of these natural language instructions, to enhance reproducibility via semantic protocols (e.g. the Aquarium project) and enable robotic automation BIBREF4 by mapping natural language instructions to executable actions.

In this study we take a first step towards this goal by annotating a database of wet lab protocols with semantic actions and their arguments; and conducting initial experiments to demonstrate its utility for machine learning approaches to shallow semantic parsing of natural language instructions. To the best of our knowledge, this is the first annotated corpus of natural language instructions in the biomedical domain

that is large enough to enable machine learning approaches.

There have been many recent data collection and annotation efforts that have initiated natural language processing research in new directions, for example political framing BIBREF5 , question answering BIBREF6 and cooking recipes BIBREF7 . Although mapping natural language instructions to machine readable representations is an important direction with many practical applications, we believe current research in this area is hampered by the lack of available annotated corpora. Our annotated corpus of wet lab protocols could enable further research on interpreting natural language instructions, with practical applications in biology and life sciences.

Prior work has explored the problem of learning to map natural language instructions to actions, often learning through indirect supervision to address the lack of labeled data in instructional domains. This is done, for example, by interacting with the environment BIBREF8 , BIBREF9 or observing weakly aligned sequences of instructions and corresponding actions BIBREF10 , BIBREF11 . In contrast, we present the first steps towards a pragmatic approach based on linguistic annotation (Figure FIGREF4). We describe our effort to exhaustively annotate wet lab protocols with actions corresponding to lab procedures and their attributes including materials, instruments and devices used to perform specific actions. As we demonstrate in § SECREF6 , our corpus can be used to train machine learning models which are capable of automatically annotating lab-protocols with action predicates and their arguments BIBREF12 , BIBREF13 ; this could provide a useful linguistic representation for robotic automation BIBREF14 and other downstream applications.

Wet Lab Protocols

Wet laboratories are laboratories for conducting biology and chemistry experiments which involve chemicals, drugs, or other materials in liquid solutions or volatile phases. Figure FIGREF2 shows one

representative wet lab protocol. Research groups around the world curate their own repositories of protocols, each adapted from a canonical source and typically published in the Materials and Method section at the end of a scientific article in biology and chemistry fields. Only recently has there been an effort to gather collections of these protocols and make them easily available. Leveraging an openly accessible repository of protocols curated on the https://www.protocols.io platform, we annotated hundreds of academic and commercial protocols maintained by many of the leading bio-science laboratory groups, including Verve Net, Innovative Genomics Institute and New England Biolabs. The protocols cover a large spectrum of experimental biology, including neurology, epigenetics, metabolomics, cancer and stem cell biology, etc (Table TABREF5). Wet lab protocols consist of a sequence of steps, mostly composed of imperative statements meant to describe an action. They also can contain declarative sentences describing the results of a previous action, in addition to general guidelines or warnings about the materials being used.

Annotation Scheme

In developing our annotation guidelines we had three primary goals: (1) We aim to produce a semantic representation that is well motivated from a biomedical and linguistic perspective; (2) The guidelines should be easily understood by annotators with or without biology background, as evaluated in Table TABREF7 ; (3) The resulting corpus should be useful for training machine learning models to automatically extract experimental actions for downstream applications, as evaluated in § SECREF6 .

We utilized the EXACT2 framework BIBREF2 as a basis for our annotation scheme. We borrowed and renamed 9 object-based entities from EXACT2, in addition, we created 5 measure-based (Numerical, Generic-Measure, Size, pH, Measure-Type) and 3 other (Mention, Modifier, Seal) entity types. EXACT2 connects the entities directly to the action without describing the type of relations, whereas we defined and annotated 12 types of relations between actions and entities, or pairs of entities (see Appendix for a

full description).

For each protocol, the annotators were requested to identify and mark every span of text that corresponds to one of 17 types of entities or an action (see examples in Figure FIGREF3). Intersection or overlap of text spans, and the subdivision of words between two spans were not allowed. The annotation guideline was designed to keep the span short for entities, with the average length being 1.6 words. For example, Concentration tags are often very short: 60% 10x, 10M, 1 g/ml. The Method tag has the longest average span of 2.232 words with examples such as rolling back and forth between two hands. The methods in wet lab protocols tend to be descriptive, which pose distinct challenges from existing named entity extraction research in the medical BIBREF15 and other domains. After all entities were labelled, the annotators connected pairs of spans within each sentence by using one of 12 directed links to capture various relationships between spans tagged in the protocol text. While most protocols are written in scientific language, we also observe some non-standard usage, for example using RT to refer to room temperature, which is tagged as Temperature.

Annotation Process

Our final corpus consists of 622 protocols annotated by a team of 10 annotators. Corpus statistics are provided in Table TABREF5 and TABREF6 . In the first phase of annotation, we worked with a subset of 4 annotators including one linguist and one biologist to develop the annotation guideline for 6 iterations. For each iteration, we asked all 4 annotators to annotate the same 10 protocols and measured their inter-annotator agreement, which in turn helped in determining the validity of the refined guidelines. The average time to annotate a single protocol of 40 sentences was approximately 33 minutes, across all annotators.

Inter-Annotator Agreement

We used Krippendorff's INLINEFORM0 for nominal data BIBREF16 to measure the inter-rater agreement for entities, actions and relations. For entities, we measured agreement at the word-level by tagging each word in a span with the span's label. To evaluate inter-rater agreement for relations between annotated spans, we consider every pair of spans within a step and then test for matches between annotators (partial entity matches are allowed). We then compute Krippendorff's INLINEFORM1 over relations between matching pairs of spans. Inter-rater agreement for entities, actions and relations is presented in Figure TABREF7 .

Methods

To demonstrate the utility of our annotated corpus, we explore two machine learning approaches for extracting actions and entities: a maximum entropy model and a neural network tagging model. We also present experiments for relation classification. We use the standard precision, recall and F INLINEFORM0 metrics to evaluate and compare the performance.

Maximum Entropy (MaxEnt) Tagger

In the maximum entropy model for action and entity extraction BIBREF17 , we used three types of features based on the current word and context words within a window of size 2:

Parts of speech features which were generated by the GENIA POS Tagger BIBREF18 , which is specifically tuned for biomedical texts;

Lexical features which include unigrams, bigrams as well as their lemmas and synonyms from WordNet BIBREF19 are used;

Dependency parse features which include dependent and governor words as well as the dependency type to capture syntactic information related to actions, entities and their contexts. We used the Stanford dependency parser BIBREF20 .

Neural Sequence Tagger

We utilized the state-of-the-art Bidirectional LSTM with a Conditional Random Fields (CRF) layer BIBREF21 , BIBREF22 , BIBREF23 , initialized with 200-dimentional word vectors pretrained on 5.5 billion words from PubMed and PMC biomedical texts BIBREF24 . Words unseen in the pretrained vocabulary were randomly initialized using a uniform distribution in the range (-0.01, 0.01). We used Adadelta BIBREF25 optimization with a mini-batch of 16 sentences and trained each network with 5 different random seeds, in order to avoid any outlier results due to randomness in the model initialization.

Relation Classification

To demonstrate the utility of the relation annotations, we also experimented with a maximum entropy model for relation classification using features shown to be effective in prior work BIBREF26 , BIBREF27 , BIBREF28 . The features are divided into five groups:

Word features which include the words contained in both arguments, all words in between, and context words surrounding the arguments;

Entity type features which include action and entity types associated with both arguments;

Overlapping features which are the number of words, as well as actions or entities, in between the candidate entity pair;

Chunk features which are the chunk tags of both arguments predicted by the GENIA tagger;

Dependency features which are context words related to the arguments in the dependency tree according to the Stanford Dependency Parser.

Also included are features indicating whether the two spans are in the same noun phrase, prepositional phrase, or verb phrase.

Finally, precision and recall at relation extraction are presented in Table 5. We used gold action and entity segments for the purposes of this particular evaluation. We obtained the best performance when using all

feature sets.

Results

The full annotated dataset of 622 protocols are randomly split into training, dev and test sets using a 6:2:2 ratio. The training set contains 374 protocols of 8207 sentences, development set contains 123 protocols of 2736 sentences, and test set contains 125 protocols of 2736 sentences. We use the evaluation script from the CoNLL-03 shared task BIBREF29 , which requires exact matches of label spans and does not reward partial matches. During the data preprocessing, all digits were replaced by `0'.

Entity Identification and Classification

Table TABREF20 shows the performance of various methods for entity tagging. We found that the BiLSTM-CRF model consistently outperforms other methods, achieving an overall F1 score of 86.89 at identifying action triggers and 72.61 at identifying and classifying entities.

Table TABREF22 shows the system performance of the MaxEnt tagger using various features. Dependency based features have the highest impact on the detection of entities, as illustrated by the absolute drop of 7.84% in F-score when removed. Parts of speech features alone are the most effective in capturing action words. This is largely due to action words appearing as verbs or nouns in the majority of the sentences as shown in Table TABREF23 . We also notice that the GENIA POS tagger, which is is trained on Wall Street Journal and biomedical abstracts in the GENIA and PennBioIE corpora, under-identifies verbs in wet lab protocols. We suspect this is due to fewer imperative sentences in the training data. We leave further investigation for future work, and hope the release of our dataset can help draw more attention to NLP research on instructional languages.

Conclusions

In this paper, we described our effort to annotate wet lab protocols with actions and their semantic arguments. We presented an annotation scheme that is both biologically and linguistically motivated and demonstrated that non-experts can effectively annotate lab protocols. Additionally, we empirically demonstrated the utility of our corpus for developing machine learning approaches to shallow semantic parsing of instructions. Our annotated corpus of protocols is available for use by the research community.

Acknowledgement

We would like to thank the annotators: Bethany Toma, Esko Kautto, Sanaya Shroff, Alex Jacobs, Berkay Kaplan, Colins Sullivan, Junfa Zhu, Neena Baliga and Vardaan Gangal. We would like to thank Marie-Catherine de Marneffe and anonymous reviewers for their feedback.

Annotation Guidelines

The wet lab protocol dataset annotation guidelines were designed primarily to provide a simple description of the various actions and their arguments in protocols so that it could be more accessible and be effectively used by non-biologists who may want to use this dataset for various natural language processing tasks such as action trigger detection or relation extraction. In the following sub-sections we summarize the guidelines that were used in annotating the 622 protocols as we explore the actions, entities and relations that were chosen to be labelled in this dataset.

Actions

Under a broad categorization, Action is a process of doing something, typically to achieve an aim. In the

context of wet lab protocols, action mentions in a sentence or a step are deliberate but short descriptions of a task tying together various entities in a meaningful way. Some examples of action words, (categorized using GENIA POS tagger), are present in Table TABREF23 along with their frequencies.

Entities

We broadly classify entities commonly seen in protocols under 17 tags. Each of the entity tags were designed to encourage short span length, with the average number of words per entity tag being INLINEFORM0 . For example, Concentration tags are often very short: 60% 10x, 10M, 1 g/ml, while the Method tag has the longest average span of INLINEFORM1 words with examples such as rolling back and forth between two hands (as seen in Figure FIGREF28). The methods in wet lab protocols tend to be descriptive, which pose distinct challenges from existing named entity extraction research in the medical and other domains.

Reagent: A substance or mixture for use in any kind of reaction in preparing a product because of its chemical or biological activity.

Location: Containers for reagents or other physical entities. They lack any operation capabilities other than acting as a container. These could be laboratory glassware or plastic tubing meant to hold chemicals or biological substances.

Device: A machine capable of acting as a container as well as performing a specific task on the objects that it holds. A device and a location are similar in all aspects except that a device performs a specific set of operations on its contents, usually illustrated in the sentence itself, or sometimes implied.

Seal: Any kind of lid or enclosure for the location or device. It could be a cap, or a membrane that actively

participates in the protocol action, and hence is essential to capture this type of entity.

Amount: The amount of any reagent being used in a given step, in terms of weight or volume.

Concentration: Measure of the relative proportions of two or more quantities in a mixture. Usually in terms of their percentages by weight or volume.

Time: Duration of a specific action described in a single step or steps, typically in secs, min, days, or weeks.

Temperature: Any temperature mentioned in degree Celsius, Fahrenheit, or Kelvin.

Method: A word or phrase used to concisely define the procedure to be performed in association with the chosen action verb. It’s usually a noun, but could also be a passive verb.

Speed: Typically a measure that represents rotation per min for centrifuges.

Numerical: A generic tag for a number that doesn't fit time, temp, etc and which isn't accompanied by its unit of measure.

Generic-Measure: Any measures that don't fit the list of defined measures in this list.

Size A measure of the dimension of an object. For example: length, area or thickness.

Measure-Type: A generic tag to mark the type of measurement associated with a number.

pH: measure of acidity or alkalinity of a solution.

Modifier: A word or a phrase that acts as an additional description of the entity it is modifying. For example, quickly mix vs slowly mix are clearly two different actions, informed by their modifiers "quickly" or "slowly" respectively.

Mention: Words that can refer to an object mentioned earlier in the sentence.

Relations

Acts-On: Links the reagent, or location that the action acts on, typically linking the direct objects in the sentence to the action.

Creates: This relation marks the physical entity that the action creates.

Site: A link that associates a Location or Device to an action. It indicates that the Device or Location is the site where the action is performed. It is also used as a way to indicate which entity will finally hold/contain the result of the action.

Using: Any entity that the action verb makes ‘use’ of is linked with this relation.

Setting: Any measure type entity that is being used to set a device is linked to the action that is attempting to use that numerical.

Count: A Numerical entity that represents the number of times the action should take place.

Measure Type Link: Associates an action to a Measure Type entity that the Action is instructing to measure.

Coreference: A link that associates two phrases when those two phrases refer to the same entity.

Mod Link: A Modifier entity is linked to any entity that it is attempting to modify using this relation.

Settings: Links devices to their settings directly, only if there is no Action associated with those settings.

Measure: A link that associates the various numerical measures to the entity its trying to measure directly.

Meronym: Links reagents, locations or devices with materials contained in the reagent, location or device.

Or: Allows chaining multiple entities where either of them can be used for a given link.

Of-Type: used to specify the Measure-Type of a Generic-Measure or a Numerical, if the sentence contains this information.