BLAH3: PROPOSAL

Data programming for PubAnnotation via Snorkel

Juan M Banda

Correspondence:
jmbanda@stanford.edu
Center for Biomedical Informatics
Research, Stanford, 1265 Welch
Rd, Stanford, CA, 94305
Full list of author information is
available at the end of the article

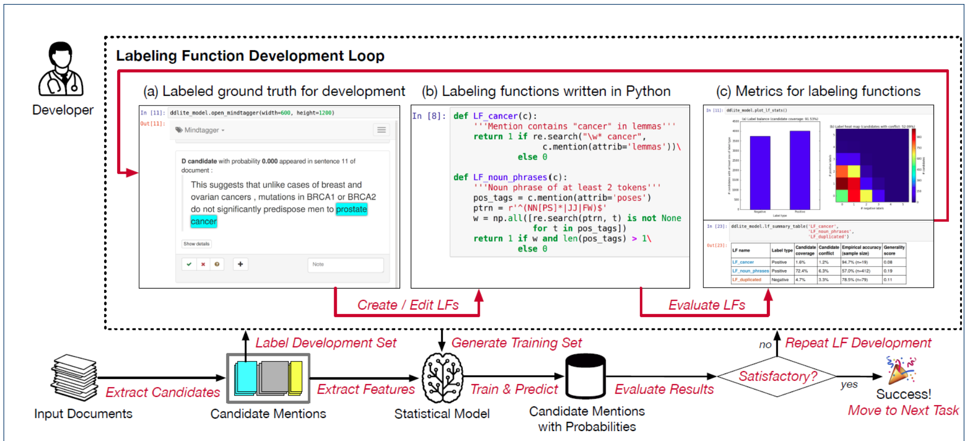
Abstract

As the volumes of medical data being generated continues to increase with wide adoption of electronic health records (EHR), cheaper sequencing technology, etc. and their subsequent use in medical publications increases, we are face with the challenge of producing high-quality labeled datasets for research. Many of the current NLP techniques rely on small hand-labeled datasets to annotate bigger corpora. While these NLP tools can alleviate some of these issues new methodologies, such as data programming, have been proposed to rely instead on 'lower quality' auto generated noisy data to train weakly supervised models and the use of generative models to denoise the noisy training data. In this proposal we propose to extend one of the leading tools, Snorkle, to support PubAnnotation format for both output of annotation and input of pre-existing labeled sets for validation of Snorkle labeling functions.

Keywords: NLP; Data Programming; Generative Models

Introduction

Data programming [1, 2] is defined as a paradigm to create training datasets in a programmatically manner. Extending the idea of distant supervision [3], where external knowledge (mappings) is used with an input dataset to create training examples. Data programming allows users to create heuristic labeling functions that provide a certain label to a subset of the data, while collectively generating a large but noisy training-set. The constructed labeling functions can be more general than in distant supervision mappings as they can additionally model an individual annotator's labels or leverage a combination of domain-specific rules and standard dictionaries. As a drawback, this can generate varying error rates, could potentially overlap, and even create conflicting rules. In order to address this, the labeling functions are modeled as a generative process, allowing for automatic denoising of



```
In [11]: ddlite_model.open_mindtagger(width=600, height=1200)
Out[11]:
```

D candidate with probability **0.000** appeared in sentence 11 of document :

This suggests that unlike cases of breast and ovarian cancers , mutations in BRCA1 or BRCA2 do not significantly predispose men to prostate cancer

```
In [8]: def LF_cancer(c):
            '''Mention contains "cancer" in lemmas'''
            return 1 if re.search("\w* cancer",
                              c.mention(attrib='lemmas'))\
                    else 0

        def LF_noun_phrases(c):
            '''Noun phrase of at least 2 tokens'''
            pos_tags = c.mention(attrib='poses')
            ptrn = r'^(NN[PS]*|JJ|FW)$'
            w = np.all([re.search(ptrn, t) is not None
                          for t in pos_tags])
            return 1 if w and len(pos_tags) > 1\
                    else 0
```

```
In [11]: ddlite_model.plot_lf_stats()
```

```
In [23]: ddlite_model.lf_summary_table('LF_cancer',
                                       'LF_noun_phrases',
                                       'LF_duplicated')
Out[23]:
```

LF name	Label type	Candidate coverage	Candidate conflict	Empirical accuracy (sample size)	Generality score
LF_cancer	Positive	1.6%	1.2%	94.7% (n=19)	0.08
LF_noun_phrases	Positive	72.4%	6.3%	57.0% (n=412)	0.19
LF_duplicated	Negative	4.7%	3.3%	78.5% (n=79)	0.11

Figure 1 Snorkle labeling procedure. Used from [1].

the resulting labeled sets by learning the correlation structure and accuracies of the labeling functions. Then, the resulting model of the labeled set is used to optimize a stochastic version of the loss function of the discriminative model that is being trained. Figure 1 gives an overview of this process. More details can be found on [1, 2], and a simplified version can be found on [4].

Snorkle (previously DDlite) is a Python framework for developing structured information extraction applications using this data programming paradigm. Snorkle learns a generative model of the noisy training set generated by the labeling functions. In other words, it learns which labeling functions are more accurate than others and trains a discriminative classifier with this knowledge. The framework is fully open source and freely available [5]. Using Juptyer/IPython notebooks, it allows researchers to focus on writing labeling functions instead of any feature extraction or complex NLP tasks. The framework contains a good amount of tutorials and features and active community.

Proposal goals

Major goals

1 Build extensions for Snorkle to produce all annotations in PubAnnotation format. This will allow us to take advantages of all of PubAnnotation's capabilities and cross functionality with other tools like Brat and Tagtog.

2 Build extensions for Snorkle to extract annotations from PubAnnotation for validation/testing of the labeling functions.

Minor goals

3 Find a collaborator that is providing an annotated corpus to BLAH3 and wants to build labeling functions to validate the power and usability of Snorkle.

Reach goals

4 Build a bare-bones web-interface to interact with a running instance of Snorkle. Essentially build a REST-like tool that allows rules to be constructed and tested via a browser with minimal interaction with Python.

Competing interests

The authors declare that they have no competing interests.

References

1. Ehrenberg, H.R., Shin, J., Ratner, A.J., Fries, J.A., Ré, C.: Data programming with ddlite: Putting humans in a different part of the loop. In: Proceedings of the Workshop on Human-In-the-Loop Data Analytics. HILDA '16, pp. 13–1136. ACM, New York, NY, USA (2016). doi:10.1145/2939502.2939515. http://doi.acm.org/10.1145/2939502.2939515
2. Ratner, A.J., De Sa, C., Wu, S., Selsam, D., Ré, C.: Data programming: Creating large training sets, quickly. In: NIPS: Proceedings of the 29th Neural Information Processing Systems Conference. NIPS '16 (2016)
3. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2. ACL '09, pp. 1003–1011. Association for Computational Linguistics, Stroudsburg, PA, USA (2009). http://dl.acm.org/citation.cfm?id=1690219.1690287
4. Data Programming: Machine Learning with Weak Supervision. Accessed: 2016-11-15. http://hazyresearch.github.io/snorkel/blog/weak_supervision.html
5. Snorkel. Accessed: 2016-11-15. https://github.com/HazyResearch/snorkel