DEBIASING NATURAL LANGUAGE EVALUATION WITH HUMANS IN THE LOOP AND STATISTICAL ESTIMATORS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Arun Tejasvi Chaganty
May 2018

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Percy S. Liang) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Christopher D. Manning)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Michael S. Bernstein)

Approved for the Stanford University Committee on Graduate Studies

Preface

Knowledge base population (KBP) systems take in a large document corpus and extract entities and their relations. Thus far, KBP evaluation has relied on judgements on the pooled predictions of existing systems. We show that this evaluation is problematic: when a new system predicts a previously unseen relation, it is penalized even if it is correct. This leads to significant bias against new systems, which counterproductively discourages innovation in the field. Our first contribution is a new importance-sampling based evaluation which corrects for this bias by annotating a new system's predictions on-demand via crowdsourcing. We show this eliminates bias and reduces variance using data from the 2015 TAC KBP task. Our second contribution is an implementation of our method made publicly available as an online KBP evaluation service. We pilot the service by testing diverse state-of-the-art systems on the TAC KBP 2016 corpus and obtain accurate scores in a cost effective manner.

Acknowledgments

I would like to thank…

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Harnessing the wealth of information present in unstructured text online has been a long standing goal for the natural language processing community. In particular, knowledge base population seeks to automatically construct a knowledge base consisting of relations between entities from a document corpus. Knowledge bases have found many applications including question answering (**???**), automated reasoning (**?**) and dialogue (**?**).

Evaluating these systems remains a challenge as it is not economically feasible to exhaustively annotate every possible candidate relation from a sufficiently large corpus. As a result, a pooling-based methodology is used in practice to construct datasets, similar to them methodology used in information retrieval (**??**). For instance, at the annual NIST TAC KBP evaluation, all relations predicted by participating systems are pooled together, annotated and released as a dataset for researchers to develop and evaluate their systems on. However, during development, if a new system predicts a previously unseen relation it is considered to be wrong even if it is correct. The discrepancy between a system's true score and the score on the pooled dataset is called *pooling bias* and is typically assumed to be insignificant in practice (**?**).

The key finding of this paper contradicts this assumption and shows that the pooling bias is actually significant, and it penalizes newly developed systems by 2% F_1 on average (Section **??**). Novel improvements, which typically increase scores by less than 1% F_1 on existing datasets, are therefore likely to be clouded by pooling bias during development. Worse, the bias is larger for a system which predicts qualitatively different relations systematically missing from the pool. Of course, systems participating in the TAC KBP evaluation do not suffer from pooling bias, but this requires researchers to wait a year to get credible feedback on new ideas.

This bias is particularly counterproductive for machine learning methods as they are trained assuming the pool is the complete set of positives. Predicting unseen relations and learning novel patterns is penalized. The net effect is that researchers are discouraged from developing innovative approaches, in particular from applying machine learning, thereby slowing progress on the task.

Our second contribution, described in Section **??**, addresses this bias through a new evaluation methodology, *on-demand evaluation*, which avoids pooling bias by querying crowdworkers, while minimizing cost by leveraging previous systems' predictions when possible. We then compute the new system's score based on the predictions of past systems using importance weighting. As more systems are evaluated, the marginal cost of evaluating a new system decreases. We show how the on-demand evaluation methodology can be applied to knowledge base population in Section **??**. Through a simulated experiment on evaluation data released through the TAC KBP 2015 Slot Validation track, we show that we are able to obtain unbiased estimates of a new systems score's while significantly reducing variance.

Finally, our third contribution is an implementation of our framework as a publicly available evaluation service at `https://kbpo.stanford.edu`, where researchers can have their own KBP systems evaluated. The data collected through the evaluation process could even be valuable for relation extraction, entity linking and coreference, and will also be made publicly available through the website. We evaluate three systems on the 2016 TAC KBP corpus for about $150 each (a fraction of the cost of official evaluation). We believe the public availability of this service will speed the pace of progress in developing KBP systems.

Bibliography

H. Adel, B. Roth, and H. Schütze. 2016. Comparing convolutional neural networks to traditional models for slot filling. In *Human Language Technology and North American Association for Computational Linguistics (HLT/NAACL)*.

G. Angeli, J. Tibshirani, J. Y. Wu, and C. D. Manning. 2014. Combining distant and partial supervision for relation extraction. In *Empirical Methods in Natural Language Processing (EMNLP)*.

J. A. Aslam, V. Pavlu, and E. Yilmaz. 2006. A statistical method for system evaluation using incomplete judgments. In *ACM Special Interest Group on Information Retreival (SIGIR)*, pages 541–548.

J. Berant, A. Chou, R. Frostig, and P. Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Empirical Methods in Natural Language Processing (EMNLP)*.

C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees. 2007. Bias and the limits of pooling for large collections. In *ACM Special Interest Group on Information Retreival (SIGIR)*.

C. Buckley and E. M. Voorhees. 2004. Retrieval evaluation with incomplete information. In *ACM Special Interest Group on Information Retreival (SIGIR)*, pages 25–32.

R. L. Burden and J. D. Faires. 1985. *Numerical Analysis (3rd ed.)*. PWS Publishers.

G. V. Cormack, C. R. Palmer, and C. L. A. Clarke. 1998. Efficient construction of large test collections. In *ACM Special Interest Group on Information Retreival (SIGIR)*.

H. T. Dang. 2016. Cold start knowledge base population at TAC KBP 2016. *Text Analytics Conference*.

J. Ellis, X. Li, K. Griffitt, and S. M. Strassel. 2012. Linguistic resources for 2012 knowledge base population evaluations. *Text Analytics Conference*.

A. Fader, L. Zettlemoyer, and O. Etzioni. 2014. Open question answering over curated and extracted knowledge bases. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1156–1165.

S. Han, J. Bang, S. Ryu, and G. G. Lee. 2015. Exploiting knowledge base to generate responses for natural language dialog listening agents. *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 129–133.

D. K. Harman. 1993. The first text retrieval conference (trec-1) rockville, md, u.s.a., 4-6 november, 1992. *Information Processing and Management*, 29:411–414.

L. He, M. Lewis, and L. Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Empirical Methods in Natural Language Processing (EMNLP)*.

H. Ji, R. Grishman, and H. Trang Dang. 2011. Overview of the TAC 2011 knowledge base population track. In *Text Analytics Conference*.

K. S. Jones and C. V. Rijsbergen. 1975. Report on the need for and provision of an "ideal test collection. *Information Retrieval Test Collection*.

A. Kalyanpur, B. K. Boguraev, S. Patwardhan, J. W. Murdock, A. Lally, C. A. Welty, J. M. Prager, B. Coppola, A. Fokoue-Nkoutche, L. Zhang, Y. Pan, and Z. M. Qui. 2012. Structured data and inference in deepqa. *IBM Journal of Research and Development*, 56:351–364.

A. Liu, S. Soderland, J. Bragg, C. H. Lin, X. Ling, and D. S. Weld. 2016. Effective crowd annotation for relation extraction. In *North American Association for Computational Linguistics (NAACL)*, pages 897–906.

C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. 2014. The stanford coreNLP natural language processing toolkit. In *ACL system demonstrations*.

A. B. Owen. 2013. *Monte Carlo theory, methods and examples*.

E. Pavlick, H. Ji, X. Pan, and C. Callison-Burch. 2016. The gun violence database: A new task and data set for NLP. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1018–1024.

L. Ratinov, D. Roth, D. Downey, and M. Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Association for Computational Linguistics (ACL)*.

S. Reddy, M. Lapata, and M. Steedman. 2014. Large-scale semantic parsing without question-answer pairs. *Transactions of the Association for Computational Linguistics (TACL)*, 2(10):377–392.

T. Sakai and N. Kando. 2008. On information retrieval metrics designed for evaluation with incomplete relevance assessments. In *ACM Special Interest Group on Information Retreival (SIGIR)*, pages 447–470.

D. Vannella, D. Jurgens, D. Scarfini, D. Toscani, and R. Navigli. 2014. Validating and extending semantic knowledge bases using video games with a purpose. In *Association for Computational Linguistics (ACL)*, pages 1294–1304.

W. E. Webber. 2010. *Measurement in Information Retrieval Evaluation*. Ph.D. thesis, University of Melbourne.

E. Yilmaz, E. Kanoulas, and J. A. Aslam. 2008. A simple and efficient sampling method for estimating AP and NDCG. In *ACM Special Interest Group on Information Retreival (SIGIR)*, pages 603–610.

J. Zobel. 1998. How reliable are the results of large-scale information retrieval experiments? In *ACM Special Interest Group on Information Retreival (SIGIR)*.