Exploiting Search Logs to Aid in Training and Automating Infrastructure for Question Answering in Professional Domains

Filippo Pompili
Thomson Reuters
Centre for AI & Cognitive Computing
120 Bremner Blvd.
Toronto, Ontario M5J 0A8
Canada
filippo.pompili@tr.com

Jack G. Conrad[†]
Thomson Reuters
Research & Development
610 Opperman Drive
Saint Paul, Minnesota 55123
USA
jack.g.conrad@thomsonreuters.com

Carter Kolbeck
Thomson Reuters
Centre for AI & Cognitive Computing
120 Bremner Blvd.
Toronto, Ontario M5J 0A8
Canada
carter.kolbeck@tr.com

ABSTRACT

Developing an AI question answering system for the legal and regulatory domain requires significant ground truth annotations for what constitutes a good answer for a given question. Collecting these annotations from qualified legal and regulatory professionals is time consuming and expensive. By making use of user activity data from the query logs of existing legal and regulatory search engines, it is possible to speed up the annotation collection process as well as supplement annotations with imputed labels. We used signals from user activity logs indicating that a user affirmatively engaged with an answer after entering a query. We leveraged these signals to infer suitable answers to questions without needing to rely on annotators. In previous research efforts, such identification was known as Implicit Relevance Feedback (IRF). Our investigations have determined that 90% of our IRF candidates contain either a complete or partial answer. Given such an elevated baseline, the next phase of this project involved harvesting data derived from such IRF (we've termed it "silver data" in contrast to expert-annotated "gold data") and extending the process to significantly larger sets of data. We examine how the approach affects performance ranging from zero, and very low amounts of gold data to substantially higher amounts of gold data. Such efforts can result in producing appreciably more reliable amounts of training data for next generation QA systems as well as establishing the means to automate the infrastructure that supports such systems. We investigate the impact of including silver data alongside gold data on the performance of a QA system. Specifically: how does silver data impact the cold start challenge (when no gold data exists initially), how much gold data is needed to achieve comparable performance to a model trained on a given amount of silver data, and what performance gains can be realized by introducing silver data to graduated amounts of gold data? We show that leveraging silver data can establish a preliminary QA system in the absence of gold data, and boost the system's performance once the gold data workstream is in place. We further show the relative efficacy of silver data to gold data by conducting performance comparisons for models trained on varying ratios of each type of data.

CCS CONCEPTS

• **Information systems → Information Retrieval**; *Information retrieval query processing*; Query log analysis; **Information Retrieval**; *Evaluation of retrieval results*; Relevance assessment; **Data analytics**; *Environment-specific retrieval*; Clustering.

KEYWORDS

Data Mining, Data Analysis, Query Log Analysis, Legal Applications, Evaluation

ACM Reference Format:
Filippo Pompili, Jack G. Conrad, and Carter Kolbeck. 2019. Exploiting Search Logs to Aid in Training and Automating Infrastructure for Question Answering in Professional Domains. In *Seventeenth International Conference on Artificial Intelligence and Law (ICAIL '19), June 17–21, 2019, Montreal, QC, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3322640.3326738

1 INTRODUCTION

Over the past decade, increased use of machine learning and other artificial intelligence technologies has significantly expanded legal professionals' abilities to efficiently access, process, and analyze digital information. AI breakthroughs continue to improve everything from advanced search to information extraction and from data summarization to classification and review. When investigating new capabilities for state-of-the-art search or question answering systems, obtaining a sufficient amount of expert labeled training data is often a daunting and costly challenge.

For general web-based retrieval as well as for domain-specific applications, research has shown that practical alternatives to such human-generated "gold data" exist. For web-based applications, the supplemental training data can come in the form of selections users have clicked on in the course of their activity. This activity can involve documents, navigational paths, or commercial selections. By contrast, in areas of domain-specific or professional search, surrogate training data can derive from users' in-depth interaction with particular documents (or content at other meaningful levels of granularity such as sections, paragraphs or sentences) represented in

[†] Corresponding author.

ICAIL '19, June 17–21, 2019, Montreal, QC, Canada

ACM ISBN 978-1-4503-6754-7/19/06.
https://doi.org/10.1145/3322640.3326738

result sets. In other words, such interactions exceed merely viewing documents or content appearing in result sets.

In this work, we denote such sources of alternatively labeled training data "silver data" in contrast with the more laboriously obtained and costly "gold data." We study the value of such low cost implicit relevance feedback (IRF) derived from user activity logs in at least two distinct ways. First, we explore and quantify its potential contribution to a challenge often known as the "cold start" problem. That is, how do we begin to train a machine learning algorithm underlying a search engine when we have not yet assembled human-produced labels for training data? The argument is that in the absence of gold data, silver data may serve as a useful substitute. Secondly, as established legal information providers and start-ups alike "operationalize" their increasingly more powerful AI-fueled capabilities in expanding competitive settings, the need to improve and retrain models with the most up-to-date assessments will make "time to market" a growing imperative. For large-scale, increasingly cloud-based production environments that may serve millions of legal and regulatory professionals in the U.S. and Canada, the UK, and elsewhere, the demand to automate infrastructure has never been so critical. Advantages in the marketplace include secure and cost-effective on-demand scalability, flexibility and reliability which typically offer ease of use and high performance as well. As reliability statistics will demonstrate in a subsequent section, given the performance properties of our silver data, if one can scale its procurement, one could automate the consistent retraining of the engine based upon the latest user behavior (interaction with their results). This subject will be revisited later in the work.

The contribution of this work is to formulate and answer three fundamental research questions involving the role and extent that IRF in the form of silver data can play in establishing a question answering system in the legal and regulatory domain when little or no gold training data is available. As limited initial gold data becomes available, we discuss in terms of silver to gold data ratios the added value that silver data can play to help boost search engine performance.

2 USE CASES

The scope of use cases associated with user activity logs is considerable, especially when a production-calibre system is involved. Some of these use cases are illustrated below.

In the context of Information Retrieval (Search) systems applied to domains such as legal and regulatory as well as others such as finance, we have identified a set of use cases tied to exploiting search queries and the documents returned from them. One set of use cases that has proven to be valuable in leveraging user activity logs, commonly referred to as query logs, is silver data. We formally define silver data as question-answer (QA) pairs harvested from the logs that can be used to identify answer documents (that can be assigned positive labels) and non-answer documents (that can be assigned negative labels). In the case of our professional legal and regulatory search system,[1] we confirm a positive instance of silver data when a user has participated in a *heightened degree of interaction* with a document returned by his or her query, interaction that exceeds simply viewing a document, namely, printing, emailing, saving or exporting a document. We rely on this heightened interaction as a *strong indication of the relevance* of that document to the user's query. Given this definition of silver data, there are a number of related validation activities that can be conducted in conjunction with this data. For example, occasionally we have subject matter experts (SMEs) review the results of our silver data queries in order to verify that these documents so identified do indeed warrant a positive relevance judgment. Depending on how reliable the assignment process is, we can quantify our confidence in the silver data. (See Section 4.1.2.)

In addition to the procurement of silver data, the query logs can be utilized to differentiate jurisdictional and practice area-specific queries from one another, for example, federal queries from state queries and these from other practice areas such as auditing & accounting or pension & benefits queries based on the data collections users run their queries against. Furthermore, query results that are not assigned a positive label, either by human editors or by silver data harvesting, can be imputed to be negative examples and thus be assigned a negative label. Given a set of gold and silver labeled data, experiments can be conducted on mixtures of less expensive silver data and more expensive gold data in order to achieve search performance levels that meet business requirements and expectations. Beyond these fundamental use cases, enterprises can also exploit their user activity logs for exemplar user queries to populate query auto-suggest or auto-complete functions. One can also cluster similar queries in the logs to track query frequency and to avoid serving up highly similar queries in auto-suggest recommendation systems. Alternatively, such query clusters can be presented to subject matter experts for straightforward selection and validation of exemplar queries, thus eliminating the sometimes challenging task of generating novel queries while still making limited use of domain expertise in the constructive selection of queries for training data. Lastly, one can use rule-based or ML-classifier-based approaches to extract or filter certain classes of queries, sometimes called "frames" in the context of QA systems, for further analysis and application. If one can design a highly effective query frame classifier, one may be able to refine the performance of the system by focusing on the ability of the search engine to treat such classes of queries in special ways using a divide and conquer strategy.

3 PRIOR WORK

The primary approaches employed to improve search results fall into three categories: document-centric, query-centric, and ranking-centric. Past research has shown that each of these approaches can be improved by exploiting user data.

Document-centric approaches involve modifying documents to create "surrogate documents" that allow for more informative query-document pairs. One common approach is to add metadata to a document to create a surrogate document; in web search, this metadata is often anchor text from hyperlinks [6]. In [19], Xue et al. take advantage of clickthrough data to associate queries with documents and include the query text as metadata for the associated document. They develop a search algorithm that makes use of both the surrogate document as well as the original document.

[1] Here we are addressing legal and regulatory search systems like those provided by Thomson Reuters, LexisNexis, Wolters Kluwer, Bloomberg and others.

Query-centric approaches involve modifying queries to create more informative query-document pairs. One common approach is to expand a query using a thesaurus based on statistics of word occurrences in a corpus; the corpus considered can be a global one, or a local one consisting of initial retrieval results [5] [18]. In [7], Hang et al. take advantage of clickthrough data to determine which documents were relevant to a given query. Correlations between query terms and the relevant documents' terms were used to select expansion terms to be added into the query. In [8], Custis and Al-Kofahi use clickthrough data for query expansion in the legal domain.

Ranking-centric approaches, like the one we present in this paper, involve creating features and/or labels for query-document pairs in order to train a machine learning model to rank search results. A common approach is to create text-based features from the query and document and to have domain experts manually select and label query-document pairs to be used for training.

In [11], Joachims et al. analyze how to take advantage of clickthrough data as implicit feedback to identify query-document pairs to use for training. While they show that web search engine clickthrough data is an effective form of implicit feedback, they also show that it is subject to biases: users' click behavior is influenced by the order in which a document appears in a result list, and by the content of the documents that appear with it in the search results.

In [1], Agichtein et al. expand on previous implicit feedback research to make it applicable to the real web search setting by looking at clickthrough data and other browsing behavior such as dwell time (the time a user spends on a page) from millions of search engine interactions (Bilenko et al. in [4] also show the utility of using dwell time as implicit feedback). Their methods treat user sessions as unreliable, noisy data points that can be aggregated to provide beneficial implicit feedback. Similarly, in [2], Agrawal et al. describe a method to aggregate probabilistic clickthrough behaviors of many users in order to directly generate labels to be used in reranking.

Agichtein et al. also show how implicit feedback techniques can be used to augment other common web search features. Notably, their experimental results show that a reranking algorithm that uses the popular BM25F score along with implicit feedback as features returns a relevant document in the top position 69% of the time, while the baseline BM25F returns the relevant document in the top position 53% of the time [1].

In [13], Liao and Moulinier investigate the task of re-ranking search results based on query log information. Earlier work has considered this problem either as the task of learning document rankings by using features based on user behavior, or as the task of enhancing documents and queries using log data. Their contribution combines both. They distill log information into event-centric surrogate documents (ESDs), and extract features from these ESDs to be used in a learned ranking function. Their experiments on a legal corpus demonstrate that features engineered on surrogate documents lead to improved rankings, in particular when the original ranking is of poor quality.

While the methods to identify relevant documents used in this paper could be applied to document-centric or query-centric approaches, we limit ourselves to looking at the ranking-centric application. The research cited here shows the effectiveness of using clickthrough and other web browsing behavior as implicit feedback for training a machine learning algorithm for search ranking. We show that the specific ways legal and regulatory researchers interact with documents in our domain-specific search environment can also be used as effective implicit feedback towards this end.

4 DATA

4.1 User Activity Logs

We initially restricted our investigation to a subset of our comprehensive query logs, one that involved subtopics such as tax. We began with approximately 5 million queries from our query logs. These queries included Boolean, keyword, and natural language searches. When we extracted from this set the queries involving natural language questions, the subset of remaining queries approached 500,000. Upon subsequently clustering this set of a half-million queries, and conflating duplicate and fuzzy duplicate queries, the set of queries was reduced to 100,000 queries. This is the set we worked with in the experiments described in Section 6. Examples from both the broader set of NL queries as well as a specific subset are shown below.

Set A: General: Legal
- What is a certificate of appealability?
- Who has the burden of proof for contempt?
- What are the elements of criminal trespass?
- When can an insurer seek reimbursement of defense costs from the insured?
- Can a non-signatory be liable for breach of contract?

Set B: Specific Subtopic: Tax
- Can I claim a hobby loss deduction for expenses paid for a horse show activity?
- What is the test to determine if a profit motive exists?
- How is fair market value of restricted stock determined at vest?
- Is the sale of property by a real estate holding company capital gain or ordinary income?
- Is covenant not to compete deductible as a business expense?

Subdomains aside, the common property that these queries possess is that they are reasonably well-formed natural language interrogatory statements. They represent more than short or simple keyword queries. They generally begin with a question word and can fall into a variety of question categories such as definitions, classifications, requirements, conditionals, time-constrained, etc. But such types are beyond the scope of this current work and will be reported on in a subsequent report.

4.1.1 Gold Data Grading Schema. When we have had SMEs review result sets for the purposes of grading, we have asked them to apply the following grading scale.

- A - Fully and completely answers the question and flows fluidly as a response. An A answer is so good that it could appear on the system immediately as a response to the user's question.
- C - Basically answers the question, but it is not as good as an A answer because it may flow oddly as a response or encloses the answer in extraneous facts, exceptions or circumstances that are not present in the question.

- D - Does not answer the question, but it is related enough to the issue that a user may understand why it would appear in the results as an answer.
- F - Does not answer the user's question and is completely unrelated to the issue. This would be embarrassing to include on the system as a response to the question.

4.1.2 Silver Data Validation Study. As described above, we have defined "silver data" as question-answer pairs originating from the query logs that are distinct from gold data insofar as no human has participated in the grading of the answer document. Instead, we select query result documents that the user has printed / saved / emailed / exported (i.e., engaged in a higher-level activity thus indicating a heightened degree of relevance) about which we have greater confidence in the relevance of the result.

To test this hypothesis, we asked one of our senior SMEs to "grade" a random selection of 125 QA pairs from our silver data set. She used the same grades and grading guidelines used by the editors for our gold data queries. In general, these queries were directed at Federal materials. As presented in Table 1, for 64% of the silver data queries examined, the SME assigned an 'A' grade, and in 90% of the instances, an 'A' or 'C' grade. In less than 10% of the cases did the query receive a 'D' grade and in 1% an 'F' grade. Given the 90% finding for an 'A' grade or within one grade of an 'A' grade, we would conclude that the silver data shows itself to be a valuable resource for initial training and automated infrastructure.

Grade	Count	Percent
A	80	64.0%
C	32	25.6%
D	12	9.6%
F	1	0.8%
TOTAL	125	100.0%

Table 1: Breakdown of Grades for Silver Data Sample

5 METHODOLOGY

The methodology described in this section involves work to monitor the development of a trial question answering system and the gold and silver data used to support that effort.

5.1 Re-ranker

Given a query, we use an internal engine coupled with our data repository (a.k.a., Novus) as our recall-focused first stage of search. The engine is a next generation derivative of the INQUERY search engine reported on by Turtle and Croft as well as Voorhees and Harman [16, 17]. It returns a set of candidate documents from our database of legal and regulatory documents. For our precision-tuned second stage of search, these documents are then re-ranked based on the features extracted from the query-document pairs formed by pairing the query to each document in the results set. Comparisons are made based on diverse categories of features that offer a spectrum of ways to measure similarity. These features enable the calculation of overlap between common elements in the query and document, such as n-grams, the cosine similarity of other representations, such as embeddings, the correspondence between parse trees, or even matching terms of art used (e.g., "guardianship", "owner-employee", "foreign tax shelter"). The documents are re-ranked using these features in conjunction with a state-of-the-art learning-to-rank algorithm that outputs a score signifying how well the document answers the query. The data used to train the re-ranking model is described below.

Juris-diction	Total Qrys	Total QA Pairs	Total Gold QA Pairs	A-graded QA Pairs	% of Total Gold Pairs
Federal	3,000	221,812	27,024	4,904	77.7%
State	1,200	107,811	7,768	2,389	22.3%
Combined	4,200	329,623	34,792	7,293	100.0%

Table 2: Statistics for Gold Data used for Training

5.2 Gold Data

The labels for the query-document pairs used for training come largely from subject matter experts (SMEs) who have evaluated tens of thousands of query-document pairs using the grading scale described in Section 4.1.1. We represented these grades on a scale from F=0 to A=3. The SME-provided labels serve as the basis for our re-ranking model and, along with their corresponding query-document pairs, are referred to as our "gold data."

The gold data queries were collected from the system user activity (query) logs. These queries were extracted based on two main natural language criteria: (i) contain a question word, and (ii) satisfy query length criteria in terms of tokens, for example, $5 \leq qry\_lgth \leq 30$. These queries were then clustered using embeddings so that similar questions were in the same cluster. SMEs were presented these clusters and chose queries (usually at most one per cluster) that were good representations of what might be asked by a legal and regulatory professional.

The editors were then presented with a set of candidate answer documents for each query. These candidate answer documents were the top documents returned when the query was entered into our existing search engine. Each of these query-document pairs was assigned a score.

5.3 Silver Data

In addition to the gold data, we used query-document pairs that had labels assigned based on user behavior as represented in the user activity logs, a.k.a. "silver data."

The log data consists of records of user actions performed on documents that occur following the execution of a user query. After running a query, in addition to viewing a document, a user may perform the actions on it described in Section 4.1.2. We consider these actions to be strong signals that a user considers a document relevant to the query, and we assign such query-document pairs an 'A' label. Simply viewing/reading a document without performing further actions on it was not considered a strong signal of document relevance and thus assigning such a label was not justified.

5.4 Imputed Negatives

Silver data queries provide us with additional positive labels for query-document pairs. For additional negative labels for query-document pairs, we ran our existing search engine using gold data queries and selected documents from low ranks (outside the top

N, where *N* was well into double digit ranks) to be paired with the query. These were given labels of 'F'. The rationale behind such "imputed negative" label assignments was that any answer document would most likely appear in the top ranks, so that identifying documents located in the *lower* ranks would be a means of gathering additional negative training data in a cost effective manner. One might correctly observe that this approach does not investigate the issue of selecting the most "valuable" negatives for training data (i.e., those more likely to be mistaken for positives by the ranking model), but instead relies on relatively large amounts of question-answer pairs gathered on the strength in numbers proposition that in the majority of cases these lower ranking pairs are true negatives.

5.5 Principal Research Questions

Using these different types of training data, we were interested in investigating three key research questions:

(1) The Cold Start challenge (when no gold data exists initially) – does silver data provide a useful starting point?
(2) How much gold data is needed to reach a performance level comparable to what was achieved by the silver data?
(3) Does adding silver data to graduated amounts of gold data still add value?

We will discuss specific details of these research questions as they are addressed in the experiments reported on below.

6 EXPERIMENTS

6.1 Research Question 1

We are interested in assessing the value of silver data in the absence of sufficient gold data in the early stages of development of a dedicated domain specific search engine or question answering system (Test 1). We conducted a series of experiments that explored the effectiveness of silver data originating from our user activity logs. In a follow-up to these trials, we included gold data that originated from our gold data repository. We conducted experiments in two separate jurisdictions of the legal and regulatory domain: Federal and State (Table 3). We based the data collections for these trials on data sets already participating in training exercises associated with the development of a question answering system for the legal and regulatory domain. For Federal, we relied on one annotated compendium consisting of approximately 100,000 documents. For State, by contrast, we used a diverse set of "reporter" collections consisting of caselaw, statutes, regulations, rulings as well as annotated secondary materials comparable to legal summaries. The total number of documents in these State collections was close to double that for Federal.

Our first set of experiments involved training our ranking engine with gradually increasing percentages of gold data [0% | 2% | 5% | 10% | 20% | 35% | 50% | 65% | 75% | 100%], excluding a hold-out set of 200 gold data queries for testing. As a baseline set, these experiments involved no silver data. They focused on gold data only and were conducted averaging over 10 independent graduation "paths" whose performance was computed against the same 200 hold-out queries.[2] No imputed negatives were used from the gold data set of queries for these baseline runs.[3] The plotted results of these experiments for Federal and State are shown in Figures 1 and 5.[4] These plots tend to show a lower performance, especially at the lower percentage levels of gold data used for training (i.e., below 20%). We note that these experiments are of a preliminary nature, and more thorough testing is to be conducted over multiple independent test sets.

6.2 Research Question 2

In addition to the experiments described above, there is a separate study to determine when the amount of gold data attains the same performance level as that when all of the available silver data is used for training (Test 2). The response to this research question can be seen in the two Federal plots where silver data is used (Figures 3 and 4) and in the two State plots where silver data is used (Figures 7 and 8). In each of these instances, one can see the performance achieved by training on the silver data only in the very first data points to the left, corresponding to the 0% gold data mark on the x-axis. In order to determine how much gold data is required to match the performance level achieved by the silver data, one has to examine the two preceding plots (2 and 3 for Federal and 6 and 7 for State). Focusing on answers@rank 1, for silver (no imputed negatives), the performance level is 0.49 for Federal (Figure 3) and 0.47 for State (Figure 7). In order to answer the question posed above, one needs to look at the corresponding tables where gold data is used in the absence of silver data. For Federal, no imputed gold is at the 25% mark (Figure 1), and with imputed negatives, it is at the 5% mark (Figure 2). For State, for imputed gold, it is close to the 8% mark (Figure 6), but without imputed gold, our tests indicate that performance will not be able to reach the level established by silver (Figure 5). So the answer to the second research question posed is: not much for Federal (5% to 25% and less than 10% for State as long as imputed negatives are used).

6.3 Research Question 3

In the next series of experiments, we began by training the ranking engine with all of our available silver data, and then proceeded to train versions of the engine with the gradual introduction of gold data using the same gradations shown for the baseline gold data experiments illustrated above. In order to answer this question (Test 3), one has to compare the figures using no silver data to the corresponding plots using silver data, so Figure 3 with 1 and Figure 4 with 2 for Federal, and Figure 7 with 5 and Figure 8 with 6 for State. For Federal without use of imputed negatives, the answer

[2] At the risk of having our estimates being overly sensitive to which test queries were selected, we chose the 200 hold-out approach for reasons of efficiency. Given our numbers of repetitions and assessment of the variability of the graduation paths, if we changed our queries as well, we would likely need to repeat each of our experiments several hundred times.

[3] Worth noting is that for all experimental *testing*, only queries with an SME-verified answer were used.

[4] In the discussions here and below, 'figure' and 'plot' are used interchangeably and all figures referred to appear on the last two pages of this report. For clarity among the most important ranges of performance, these plots are cut off at the 50% of gold data mark along the x-axis. After 50%, the plots tended to be flat or approximately flat.

Jurisdiction	Collection Sources	Silver Data Qrys	% of Total
Federal	FEDANA (Annotated Compendium)	3203	92.6%
State	State Reporters	254	7.4%
Combined	Federal & State	3457	100.0%

Table 3: Quantity of Available Silver Data

is clearly yes for the lower percentages of gold. In Figure 1, the answers@rank1 are at 0.24, 0.36 and 0.42 for 2%, 5% and 10%, whereas with silver data in Figure 3, those points are at 0.47, 0.48 and 0.48. Only when we get up to 35% are the two comparable in the 51%-52% range. For Federal, the answer with the use of imputed negatives is still yes. In Figure 2, the answers@rank1 are at 0.43, 0.49 and 0.50 for 2%, 5% and 10%, whereas with silver data in Figure 4, those three points all stand at 0.48. So in this case, the silver data contributes more marginally, only in the 2% to 5% range.

For State, these differences appear to be *much* more pronounced. Without use of imputed negatives, the answer is clearly yes for the lower percentages of gold. In Figure 5, the answers@rank1 are at 0.16, 0.17, 0.18, 0.25, 0.30 and 0.31 for 2%, 5%, 10%, 20%, 35% and 50%, whereas with silver data in Figure 7, those points are at 0.52, 0.53, 0.55, and the rest 0.57. So across the board here, silver data clearly helps. For State, the answer with the use of imputed negatives is still yes. In Figure 6 the answers@rank1 are at 0.39, 0.44, 0.49, 0.54, 0.55 and 0.60 for 2%, 5%, 10%, 20%, 35% and 50% whereas with silver data in Figure 8, those points are 0.51, 0.53, 0.54, 0.58, 0.57 and 0.58. So in this case, the silver data continues to contribute, especially in the 2% to 20% range.

It would be possible to repeat this analysis above for answers in the top 2 ranks (labeled answers@rank2) in the figures, and similarly for answers in the top 3, 4 and 5 ranks as well. But since these other curves in each figure reasonably follow those for answers@rank1 the findings would not be dramatically different from those for the most important (top) rank, rank 1. Rank 1 possesses the most consequential information that may impact decisions involving performance. Note that the Figures shown in the Appendix and discussed in the next section display the curves for answers@rank1 through answers@rank5. When we refer to answers@rank_n where $n > 1$, what we mean is that at least one answer can be found between Rank 1 and Rank n.

7 RESULTS

The results we report on here derive from the two series of experiments described in Section 6 above. Given our principal motivations for this work — initial ranker training in the absence of gold data — our primary focus is on the left-hand side of the figures. For both jurisdictions examined, Federal and State, we want to compare ranker performance obtained from initial gradations of gold data with ranker performance from silver data alone, and, sequentially, from silver data in conjunction with modest portions of gold data. One observation worth noting for both jurisdictions is that for correct answers at Rank 1, the performance of the silver data plot with 0% to 2% gold data is higher than that for the plot using no silver data. For State, the trials harnessing silver data produce an answer at rank 1 of over 50% while that without it produce an answer *a third* of 50%. And for Federal, we observe a similar pattern, although much less pronounced for the answer@rank 1 plots.

7.1 Federal Results

(1) "Pure" gold data (i.e., no imputed negatives, no silver data, Figure 1), in contrast to State, gives the best results and holds the first spot as soon as 35% is reached (even though, at the cost of some noticeable variance on the samples of the graduation points).[5]

(2) Adding silver data only (i.e., with silver, without imputed negatives, Figure 3), helps the cold start from 0% up until around 20% (but only when compared against the gold only baseline shown in Figure 1),[6] after which it quickly settles to a slightly lower performance than the maximum achievable with gold-only labels, and without benefiting from a further increase of gold annotations. When compared against the scenario with imputed negatives only instead (Figure 2), this configuration holds the top performance only in the range from 0% up to a point in between 2% and 5% of gold labels.

(3) Imputed negatives also (without silver, Figure 2) are able to help at lower percentages similar to silver-only, even though they still cannot solve the 0% problem and have slightly lower performance than pure silver at very low gold percentages (requiring at least 2%-5% of gold labels before starting to outperform silver-only). On the other hand, they do not hinder performance as gold data keeps increasing.

(4) When including both silver and imputed negatives (Figure 4), we observe a result similar to scenario (2), i.e., we obtain a diminished, low variance growth, which has decent performance at low percentages, but which never clearly outperforms either of the two other options on the higher end of the gold data gradations.

7.2 State Results

(1) In striking contrast with Federal, the pure gold data baseline (i.e., no imputed negatives, no silver data, Figure 5), exhibits the lowest performing result by far among all experimental configurations. Its results also demonstrate a very high sensitivity to the particular choice of labelled data, indicating relatively high variance and "instability" of training data up to very high gradations of labelled data.

(2) Remarkably, the silver data only scenario (Figure 7) for State collections offers the greatest contribution to a substantial increase in performance at low and medium amounts of gold data, outperforming not only the gold-only baseline (Figure 5) at all percentages in the range shown, but also the imputed negatives-only alternative (Figure 6) from 0% up to 35%. In the same plots, we can also appreciate (a) the beneficial variance-reduction effect the silver data has against graduation sampling of gold data, and (b) its 0% "data-readiness." After the 35% mark, the gold with imputed negatives (Figure 6) reestablish their superiority and tend to outperform silver data-only (Figure 7), reaching answer@rank1 above 0.600 vs. 0.570.

(3) Finally, the plot with both silver and gold imputed negatives (Figure 8), in a manner similar to what was observed for

[5] About 9,460 gold data QA pairs as per Table 2.

[6] About 2,700 gold data QA pairs, at a silver to gold data ratio of roughly 1.2) (and including 0%) [cf: Tables 2 and 3].

Federal, matches in practice the behavior already observed in scenario (2), and also confirms that we do not need to discard any silver data if deciding to use them along with imputed negatives across an extended range of gold annotations.

So why is the contribution of silver data so much more effective when applied to State content versus Federal, especially when noting that there is so much less of it for State than there is for Federal (Table 3)? The first thing one should acknowledge is that we also have considerably less State gold data than Federal gold data to start with (Table 2). State gold data is less than a quarter of the total whereas State silver data is less than a *twelfth* of the total. This is not reflective of that dramatic a disparity, since 70% of the queries in this particular regulatory system logs are issued against Federal collections while only around 20% of the queries are issued against State collections. Federal is a more universally queried content set. One could argue that in the presence of a relatively small amount of gold data to begin with, silver data can have a greater impact on performance. Another dimension that might explain the much more significant impact that State-based silver data and the associated imputed negatives in general have on performance is related to a complementary property of imputed labels' *documents*, which is not directly related to the labels themselves, but rather, to the acquisition of additional statistics on the (unsupervised) distribution of the features within a dataset. This is a fact that may help the model better estimate the decision boundary, especially in cases where the variability of structure is more pronounced, as in the case of State documents. This is another topic that is being explored more systematically in our future work. Meanwhile, we discuss the role and characteristics of imputed negatives in the both of the next two sections.

8 CONCLUSIONS

Through this research, we have verified the role that silver data can play and its value in getting a QA system up and running. We have posed three essential research questions involving the role and degree of IRF that silver data can fill in the development of a preliminary question answering system in the legal and regulatory domain in the absence of training data. As limited amounts of gold data became available, we have also demonstrated what appropriate ratios of gold data to silver data could provide to practically and economically boost the performance of an early version QA system.

Of course there are limitations to the research reported on in this work. We have observed a variability in performance across practice areas. We have not presented a principled study into where (what ranks) the silver data originates. An abbreviated study has shown that they tend to come from the top ranks, not exclusively, but extensively, as in a diminishing Zipfian-shaped curve. At the same time we did not present the actual baseline performance nor the performance of the trained system following the introduction of silver and gold data. This is due to conditions imposed by the partnering businesses. Table 3 indicates a relatively small supply of silver data for the State practice area. This can be attributed to the fact that we are taking a subset of a subset. For the given application space, Federal queries are significantly more common than are State, and the natural language questions in the query logs examined are in the minority. In addition, our tests indicate that a relatively prompt performance saturation occurs with increased labeled silver data, but this is in relation to the gold data available at the time.

The silver data described in this report has thus far been used largely in an initial training role. As user activity logs grow, however, thanks to increased customer engagement with system data, and magnitudes more question-answer pairs and document interactions are recorded, it will be increasingly possible to harness such Q-A-Action triplets in an automated infrastructure capacity to facilitate the continuous training of models and the continuous delivery of upgraded performance to customer-oriented production systems.

Regarding the subject of automated infrastructure, such silver data can also be instrumental in delivering this capability to a related production system. By automating the procurement of such silver data, the system can train using many thousands of exemplar QA pairs, representing both positive and negative labels, thus establishing a system that realizes continuous integration, continuous delivery, and potentially continuous deployment as well. Up to this point, one has commonly followed enterprises such as Amazon, Netflix, Facebook, and others that have integrated such continuous processing into the life blood of their commercial ventures [10, 12, 15]. Yet given the magnitude and growth of user activity and the logs that record that activity in the Legal and Regulatory space, systems that rapidly retrain, deliver and deploy will be realizable and beneficial for legal professionals in major countries and jurisdictions where online research systems and their broad coverage of the legal landscape are the norm.

One final observation worth mentioning is the markedly beneficial role that a simple imputation strategy for negative labels can produce, as can be seen from the experimental plots and their analysis in Sections 7.1 and 7.2. The silver data approach can be differentiated from the imputed negative approach in several ways, ways that highlight the *complementary* nature of the two and which respond to the notion of using one exclusively in the absence of the other. The first difference is that silver labels are especially well suited to harvest the much needed and usually scarcely available *positive* labels (in contrast to the purely negative imputation). This means that, while it's true that we need positive labels as much as negative ones in a cold start scenario, it also implies that we cannot afford to drop the silver approach and use exclusively imputed negatives if that scenario is of particular interest. Also, silver data are by their very nature *dynamic*; that is they grow in quantity as the users keep interacting with the system, while the imputed negatives sampled from the *static* baseline are unable to adapt to the continuous stream of signals from users. Another difference is that imputed negatives arguably provide the engine with an expanded and almost arbitrarily large exposure to the depths of any underlying data collection, in contrast to both silver or graded gold data instances, which are only able to obtain a rather rarified sample of the collection from among the top ranks of results presented to users.

This simple exposure to corpus statistics is achieved without any specific attempt at harvesting particularly "hard" negative samples, i.e., those most likely to be confused as positive ones, and is driven by the optimization of imputed labels' misclassification costs, i.e., the optimization of a supervised objective function *on the final classification task* (e.g., in the form of a relevant or non relevant answer), rather than the optimization of either an unsupervised corpus modeling objective function based, e.g., on metric space proximity, or a supervised proxy as done, e.g., in the currently popular methods of deep "unsupervised" pre-training.[9][7] In contrast, our simple negative imputation heuristic can be seen as a simplified semi-supervised approach instantiated as a "single-step" self-training iteration, where the negative samples are drawn from the bottom end of the *baseline* reranked candidates, i.e., precisely those candidates that we're most confident in being true negatives (according to the baseline model). While we did not fully explore the venue of *multiple steps* iterative self-training, or even the much broader array of general semi-supervised approaches such as similarity clustering, manifold regularization, or fitting of mixture distributions, we do believe these techniques represent an interesting possibility to explore in future research. A more in-depth discussion on semi-supervised methods can be found in, e.g., [3, 14]. It remains to be studied what data characteristics affect the observed variability in the effectiveness of the imputed negatives approach (note, in particular, its different impact between Federal and State collections), even though the markedly higher variability in State collections with respect to Federal may hint at the more important role that corpus modeling can achieve in cases where document representations are strongly non-homogeneous.

9 FUTURE WORK

In future work, we plan to scale the amount of silver data we rely upon for our testing regime by harvesting larger sets from a greater number of years of user activity logs. More specifically, we will be extending the scope of our initial experiments that tested the ability of silver data to fill a gap in training data in the early deployment of a question answering system. We now know that leveraging imputed negative labels can play a valuable role in the overall training practices and that extracting silver data can be an effective way to get an engine up and running in advance of more robust testing and added training.

We have already referred to the fact that our State data contains more variability in structure across its reporter collections than the Federal data we used (greater variety in document length, in document format, in document material, etc.). We are thus interested in determining whether a more formal and systematic investigation into these properties may help confirm why such a relatively small amount of silver training data was able to produce such a meaningful positive impact on the model's overall performance. We also need to test this hypothesis against a broader array of State gold data as well. In short, State is not only a more varied practice area; it is also more illustrative of a rich, multi-dimensional content set as a whole. To treat and make new discoveries within such a content set is to begin to understand it and benefit from it.

In addition, once our systems have the chance to mature by way of the introduction of still more training data, we want to report on how our models compare to the baselines we started with. Beyond such baseline comparisons, we plan to further automate our infrastructure by using this expanded set of silver data to automatically train versions of our QA system. The subject of continuous integration, delivery and deployment in our context – in short, the opportunity for continuous training – lends itself to independent studies into the subject of active learning, and this too, is an area we wish to probe in subsequent investigations.

Lastly, given the observed contributions that training data supplemented with imputed negatives can make, we also plan to investigate a set of analogous research questions, ones probing the role and impact that imputed negatives can provide to a system's overall training regime, in terms of complete sets or graduated sets comparable to what we have examined above with the silver data.

[7] Unsupervised here means that no human expert annotations are used, but the objective metric is still formulated as fully supervised; it is sometimes referred to as *self-supervised* learning.

REFERENCES

[1] Eugene Agichtein, Eric Brill, and Susan Dumais. 2006. Improving Web Search Ranking by Incorporating User Behavior Information. In *Proceedings of the 29th Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*. ACM, 19–26.

[2] R. Agrawal, A. Halverson, K. Kenthapadi, N. Mishra, and P. Tsaparas. 2009. Generating Labels from Clicks. In *Proceedings of the Second International Conference on Web Search and Data Mining (WSDM '09)*. ACM, 172–181.

[3] Mikhail Belkin, Partha Niyogi, and Vikas Sindwani. 2006. Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. *Journal of Machine Learning Research* 7 (December 2006), 2399–2434.

[4] Mikhail Bilenko and Ryen W. White. 2008. Mining the Search Trails of Surfing Crowds: Identifying Relevant Websites from User Activity. In *Proceedings of the 17th International Conference on World Wide Web (WWW '08)*. ACM, 51–60.

[5] Claudio Carpineto, Renato de Mori, Giovanni Romano, and Brigitte Bigi. 2001. An Info-theoretic Approach to Automatic Query Expansion. *Trans. Inf. Syst.* 19, 1 (Jan. 2001), 1–27.

[6] Nick Craswell, David Hawking, and Stephen Robertson. 2001. Effective Site Finding Using Link Anchor Information. In *Proceedings of the 24th Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01)*. ACM, 250–257.

[7] Hang Cui, Ji-Rong Wen, Jian-Yun Nie, and Wei-Ying Ma. 2003. Query Expansion by Mining User Logs. *Trans. on Knowledge and Data Eng.* 15 (2003), 829–839.

[8] Tonya Custis and Khalid Al-Kofahi. 2008. Investigating External Corpus and Clickthrough Statistics for Query Expansion in the Legal Domain. In *Proceedings of the 17th International Conference on Information and Knowledge Management (CIKM'08)*. ACM, 1363–1364.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). http://arxiv.org/abs/1810.04805

[10] Nicole Forsgren, Jez Humble, and Gene Kim. 2018. *Accelerate: State of DevOps Strategies for a New Economy.* Report. DevOps Research & Assessment (DORA).

[11] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately Interpreting Clickthrough Data As Implicit Feedback. In *Proceedings of the 28th Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*. ACM, 154–161.

[12] Gene Kim, Jez Humble, Patrick Debois, and John Willis. 2016. *The DevOps Handbook:: How to Create World-Class Agility, Reliability, and Security in Technology Organizations.* IT Revolution Press, Portland, OR.

[13] Wenhui Liao and Isabelle Moulinier. 2009. Feature Engineering on Event-centric Surrogate Documents to Improve Search Results. In *Proceedings of the 18th International Conference on Information and Knowledge Management (CIKM'09)*. ACM, 629–632.

[14] Paven Kumar Mallapragada, Ring Jin, Anil K. Jain, and Yi Liu. 2009. SemiBoost: Boosting for Semi-supervised Learning. *Transactions on Pattern Analysis and Machine Intelligence* 31, 11 (2009), 2000–2014.

[15] Andi Mann. 2018. *2018 State of DevOps Report.* Report. Puppet & Splunk.

[16] Howard Turtle and W. Bruce Croft. 1991. Evaluation of an Inference Network-based Retrieval Model. *Trans. Inf. Syst.* 9, 3 (July 1991), 187–222. https://doi.org/10.1145/125187.125188

[17] Ellen M. Voorhees and Donna K. Harman. 2005. *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing).* The MIT Press, Chapter 11: University of Massachusetts INQUERY retrieval system.

[18] Jinxi Xu and W. Bruce Croft. 1996. Query Expansion Using Local and Global Document Analysis. In *Proceedings of the 19th Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR '96)*. ACM, 4–11.

[19] Gui-Rong Xue, Hua-Jun Zeng, Zheng Chen, Yong Yu, Wei-Ying Ma, Wensi Xi, and Weiguo Fan. 2004. In Optimizing web search using web click-through data. *Proceedings of the 13th International Conference on Information and Knowledge Management*, 118–126.

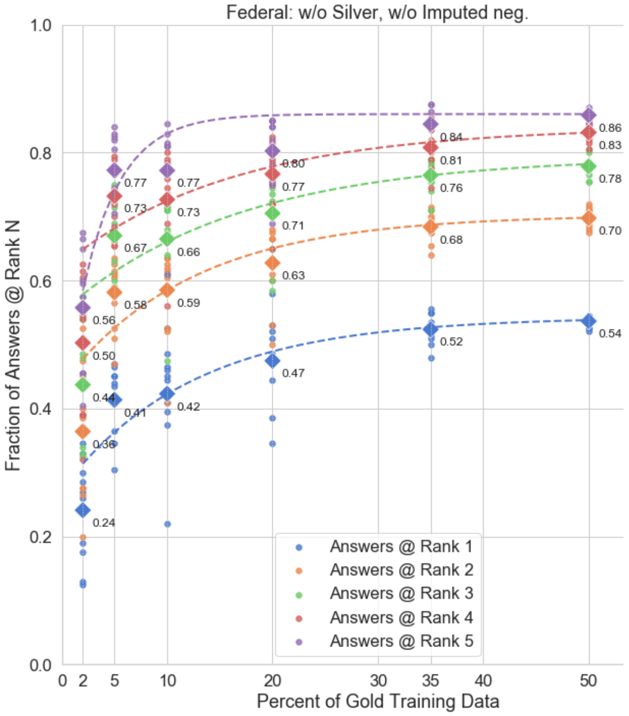

Figure 1: Fed. Perf. Curves Gold Data w/o Silver, w/o Imp Neg

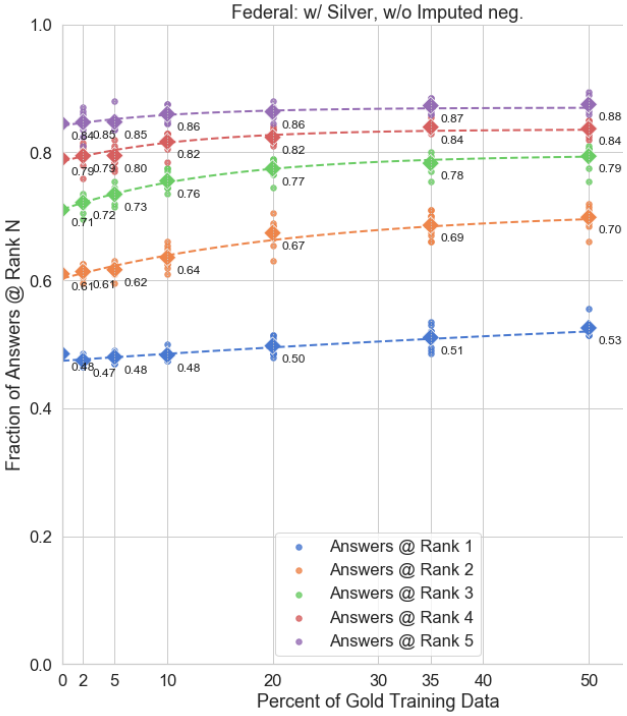

Figure 3: Fed. Perf. Curves: Gold Data w/ Silver, w/o Imp Neg

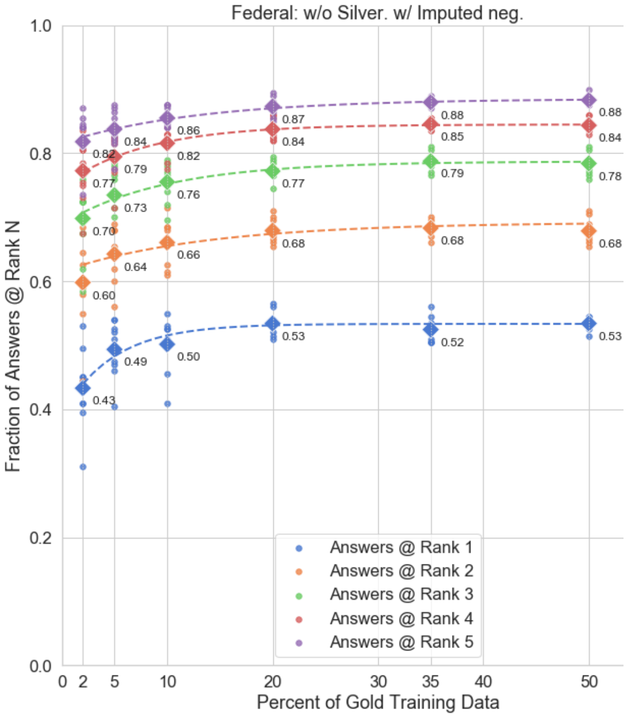

Figure 2: Fed. Perf. Curves Gold Data w/o Silver, w/ Imp Neg

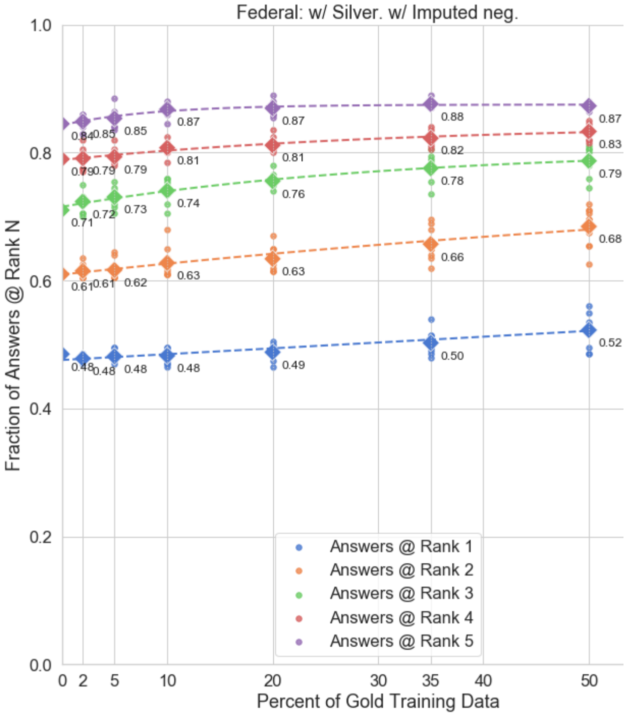

Figure 4: Fed. Perf. Curves: Gold Data w/ Silver, w/ Imp Neg

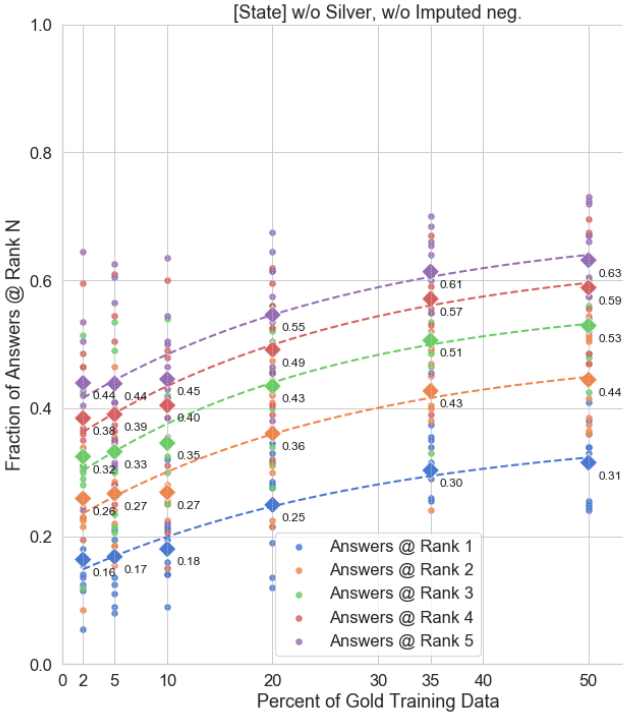

Figure 5: State Perf. Curves: Gold Data w/o Silver, w/o Imp Neg

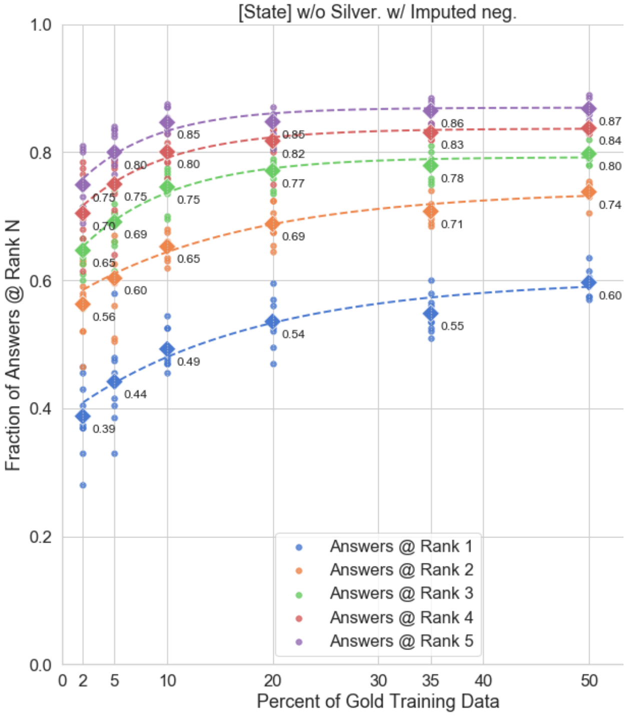

Figure 6: State Perf. Curves: Gold Data w/o Silver, w/ Imp Neg

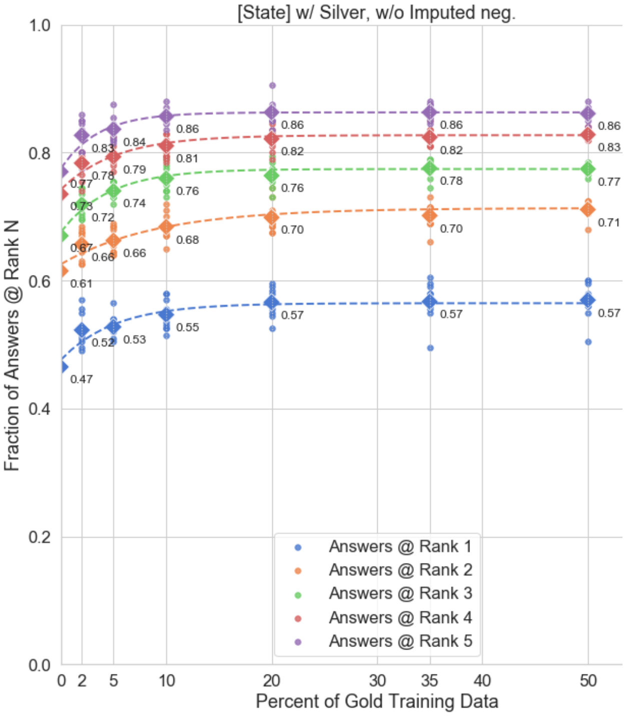

Figure 7: State Perf. Curves: Gold Data w/ Silver, w/o Imp Neg

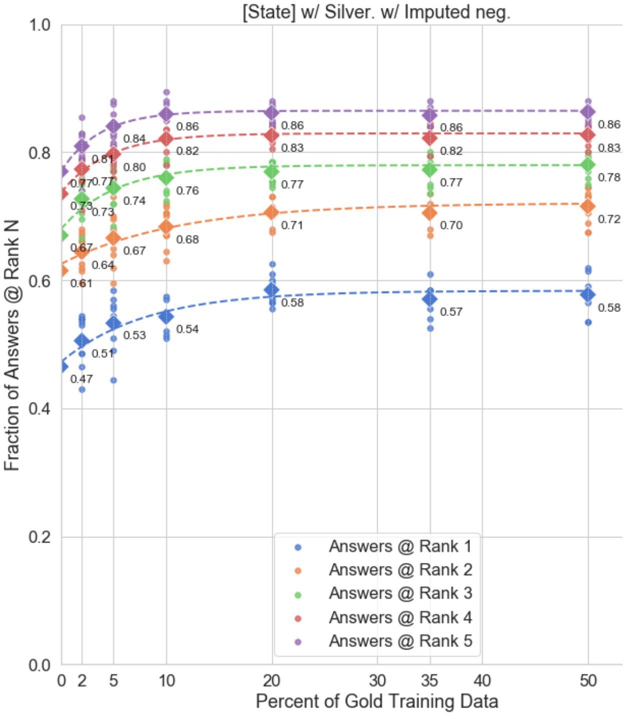

Figure 8: State Perf. Curves: Gold Data w/ Silver, w/ Imp Neg