Assessing Confidence of Knowledge Base Content with an Experimental Study in Entity Resolution

Michael Wick Sameer Singh Ari Kobren Andrew McCallum
School of Computer Science
University of Massachusetts Amherst MA
{mwick, sameer, akobren, mccallum}@cs.umass.edu

ABSTRACT

The purpose of this paper is to begin a conversation about the importance and role of confidence estimation in knowledge bases (KBs). KBs are never perfectly accurate, yet without confidence reporting their users are likely to treat them as if they were, possibly with serious real-world consequences. We define a notion of confidence based on the probability of a KB fact being true. For automatically constructed KBs we propose several algorithms for estimating this confidence from pre-existing probabilistic models of data integration and KB construction. In particular, this paper focuses on confidence estimation in entity resolution. A goal of our exposition here is to encourage creators and curators of KBs to include confidence estimates for entities and relations in their KBs.

1. INTRODUCTION

Automated information extraction and integration systems are now able to populate knowledge bases (KBs) from multiple data sources at unprecedented scales. Although recent advances in machine learning and natural language processing has improved the accuracy of such systems, they are still known to be much less accurate than humans. This is unfortunate because errors in the KB can negatively and profoundly impact decision-making, causing user-frustration in the most benign cases and serious real-world consequences in the worst cases. For this reason and others, we argue it is important to enrich the KB to include a measure of *confidence* about the entities and relations therein.

Confidence could be tremendously useful in mitigating the negative effects of KB errors. For example, confidence can greatly assist decision makers because it would enable different users to request different views of the knowledge base according to some desired level of confidence (which might depend on the domain and problem). This could be useful because some users are willing to sift through more (potentially errorful) data in order to find a particular answer, while others may require immediate high-precision answers (perhaps at the expense of recall/coverage). Further, confidence can be used to annotate the visualization of the KB entities and relations so that users can quickly assess the likelihood that a particular fact is true. Confidence may also be useful for joint data integration in which the output of one integration component depends heavily on another. For example, the output of named entity recognition could contain multiple hypotheses each annotated with their confidence; a downstream component such as entity resolution could consume this output and incorporate the various confidence levels in its predictions.

CIKM 2013 San Francisco, California USA

In order to support these use cases and others, we arrive at a list of desiderata for confidence values in KBs.

- **Truthful:** the confidence values associated with each KB item should reflect how likely that KB item is to be true. KB content with higher confidence should be more likely to be true on average than KB content with lower confidence.
- **Interpretable:** confidence values should be interpretable in both a *relative* (comparing the confidence values of two KB objects should be meaningful) and *absolute* (meaningful as a solitary value) sense.
- **Semantically meaningful:** Confidence values should obey a formal semantics, allowing confidence to take part in formal queries of the KB.
- **Consistent:** two users querying the same confidence values should receive the same answer.

Thus, a natural and general definition of confidence that satisfies these desiderata is the marginal probability that a particular fact in the KB is true (for example, that some entity or relation exists). Since the components of most automated knowledge base construction systems are probabilistic, much of the machinery for computing these marginal probabilities already exists for many KBs. However, for many important data integration tasks such as coreference resolution, computing these marginals is intractable (and a computational speed vs accuracy trade-off is necessary in order to make confidence estimation feasible in practice).

In this work, we focus on the problem of estimating confidence values for the task of coreference resolution—the problem of clustering records in the KB (mentions) into the entities to which they refer—allowing us to provide confidence estimates for all the entities in the KB. In our setting, we define confidence as the marginal probability that a set of mentions all refer to a single entity (as opposed to two or more entities). However, because computing such a probability is intractable (requires summing over all possible clusterings of mentions into entities), we propose several approximate confidence estimation algorithms and compare them in terms of speed and accuracy on both small-scale (on which we can compare approximate and exact algorithms) and larger-scale data (on which we can compare the behavior of the approximation algorithms).

A summary of our goals and contributions:

- to initiate a dialog on the importance of confidence in KB construction
- formalize a definition of confidence that utilizes preexisting probabilistic data integration models
- evaluate and compare different algorithms for estimating confidence from these probabilistic data integration models

2. ENTITY RESOLUTION

In this paper we focus on estimating confidence values for coreference (or entity resolution), the problem of clustering *mentions* into the sets such that all the members of each set refer to the same *entity*. For example, to build a bibliographic knowledge base we would like to compile a publication list for each author in the KB. To do this, we need to cluster the author fields from citations (mentions) by the authors they refer to (entities). The problem is difficult to solve in general because (1) there are many people with the same first-initial last name combination, (2) authors are referred to in multiple ways (for example, by a nickname, by initials, by first full name, etc.), (3) there is often noise due to in spelling, typographical, and OCR errors.

In coreference resolution, the model can be defined as a scoring function f that takes a set of entities as input (each entity being a set of mentions), and outputs a real-valued number indicating the collective compatibility of the entities. More formally let $\mathcal{M}$ be the set of mentions and let $P = \{E_1, E_2, \cdots, E_n\}$ be a partitioning of the mentions $\mathcal{M}$ into disjoint entity sets. Let $S \subseteq P$ be a subpartitioning of the mentions. Then, the compatibility function f maps subpartitionings S to real-valued numbers. For example $f(P) = f(E_1, E_2, \cdots, E_n)$ is defined for an entire partitioning and $f(S) = f(E_4, E_9)$ is also defined for a subpartitioning consisting of only two entities (E_4 and E_9). This formulation of coreference encapsulates a number of existing coreference models, such as pairwise [10, 5, 13], entity-wise [4, 14], and hierarchical [15].

Coreference resolution can be solved by searching for a full partitioning $P = (E_1, E_2, \cdots, E_k)$ over $\mathcal{M}$ that maximizes the function f

$$P^\star = \arg\max_P f(P) \tag{1}$$

Although the coreference optimization does not require f to be defined on subpartitionings, it is convenient for explaining some of the proposed confidence estimation methods.

3. CONFIDENCE OF ENTITY RESOLUTION

We define the confidence associated with a set of *query mentions* $A = \{m_1, m_2, \cdots, m_n\}, A \subseteq \mathcal{M}$ as the marginal probability that the mentions all refer to the same entity. To compute this marginal probability, we define a probability distribution over coreference configurations as induced by the compatibility function f:

$$\pi(P) = \frac{1}{Z}\exp(f(P)), \quad Z = \sum_{P'} \exp(f(P')) \tag{2}$$

where Z sums over all possible partitionings.

The confidence associated with the set of mentions A is

$$g^\star(A) = \sum_P \pi(P)\mathbb{1}\{\exists E_i \in P \text{ s.t. } A \subseteq E_i\} \tag{3}$$

i.e. the sum of marginal probabilities of all the configurations in which the mentions in A are coreferent. Since this value is intractable to compute in practice, we propose a number of methods to approximate the confidence.

Markov Chain Monte Carlo (MCMC) Sampling
This method uses an MCMC sampler to sample many configurations of the query mentions and counts the number of samples in which the query mentions appear in the same entity. More precisely, if $P^{(1)}, P^{(2)}, \cdots, P^{(n)}$ are a set of samples drawn from π, then the sampling estimate is

$$\hat{g}(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\exists E_i \in P^{(i)} \text{ s.t. } A \subseteq E_i\} \tag{4}$$

An advantage of this method is that it asymptotically converges to the true confidence. Additionally, it can provide any-time results (with more samples yielding better estimates of the confidence). However, the method mixes slowly in some cases and may require the use of sophisticated sampling techniques (such as tempering [12], distributed [13], or query-aware [16] algorithms) to scale to complex models and large datasets.

Query Assignment Score
This method uses the compatibility function to estimate the confidence of a single set of mentions

$$\hat{g}(A) = f(A) \tag{5}$$

An advantage of this method is that it is extremely efficient to compute. However, the query assignment score method yields unnormalized confidence estimates that have a large range, making them uninterpretable. Therefore, this method is only useful when ranking confidences (i.e, comparing different confidence estimates generated using this method).

Query Assignment Perturbation
Instead of sampling partitioning of all the mentions $Mentions$, this approach efficiently samples partitionings only of the query mentions A, and estimates the marginals of such perturbations using f. In particular, let $A^{(1)}, A^{(2)}, \cdots, A^{(k)}$ be the partitionings of the mentions in A, each having been generated by running MCMC for a few steps step from the configuration in which all mentions in A are clustered together. The query perturbation estimate is given by

$$\hat{g}(A) = \frac{1}{n}\sum_{i=1}^{k} \mathbb{1}\{A = A^{(i)}\} \tag{6}$$

This method is slower to compute than the query assignment score, however it produces confidence estimates that are locally normalized using the samples drawn from the local neighborhood of A. Therefore, we can interpret this

Properties	True Confidence	MCMC	Query Score	Perturbation	Conditional Perturbation
Accuracy	Exact	Asymptotic	Good	Fair	Good
Efficiency	Intractable	Slow	Fastest	Very Slow	Very Slow
Consistent	Y	N	Y	N	N
Probabilistic Measure	Y	Y	N	Y	Y
Relative/Absolute	Absolute	Absolute	Relative	Absolute	Absolute
Simultaneous Evaluation	Y	Y	Y	N	N
Depends on non-Query Mentions	Y	Y	N	N	Y
Depends on other KB Entities	N	N	N	N	Y

Table 1: Properties of the various Confidence Evaluation Approaches

value as an approximate probability.

Conditional Query Assignment Perturbation
This method is similar to the previously described method except that it is conditioned on a partitioning of the mentions that are not in the query set (i.e. $Q = \mathcal{M} - A$). Specifically, let $A^{(1)}, A^{(2)}, \cdots, A^{(k)}$ be sampled partitionings over the mentions in A conditioned on a fixed partitioning over Q (we use the predicted maximizing assignment, for example). The conditional query assignment perturbation confidence estimate can be computed using Equation 6. This confidence estimate can also be interpreted as an approximate *conditional* probability.

Query-aware MCMC
This is a general purpose sampling method for efficiently answering statistical queries for graphical models with MCMC sampling [16]. In our context, it is similar to the MCMC method discussed above, except that it prioritizes sampling of the query mentions A and samples other mentions in proportion to their statistical dependence on A.

The properties of the various confidence estimation methods are listed in Table 1. These properties expand upon to the desiderata outlined in the introduction.

4. EXPERIMENTS

In this section we compare the various confidence estimation algorithms for two types of entity resolution problems: author coreference (clustering the author names in citations according to the author entities they refer to), and citation matching (clustering citations into paper entities). For author coreference we use the REXA dataset [3], and for citation matching we use the CORA dataset. For each of these approaches, we implement a pairwise coreference model [10] that employs a pairwise compatibility function to evaluate how likely two mentions are to refer to the same entity. Thus, the global compatibility function f defined in Section 2 factorizes over mention pairs.

4.1 Comparison to Oracle

In this experiment we compare the approximate confidence estimation algorithms against the exact oracle confidence values. For each dataset (Rexa and Cora), we pick random subsets of 10 mentions to form our set of mentions $\mathcal{M}$. We pick 100 query entities A by selecting a random subset of size 4 from $\mathcal{M}$. We compare the confidence as computed exactly (by iterating over all possible configurations of $\mathcal{M}$) and as estimated using our proposed methods. In Figure 1a (Rexa) and Figure 1b (Cora), we plot the rank of each of the configuration by confidence against the rank according to the oracle (the raw confidence plots are provided in the appendix). We also provide the correlation of these rankings in Table 2. We find that MCMC is the most accurate algorithm for this small dataset (it should be asymptotically correct); however, it is likely to be slow for larger datasets. In contrast, the query-assignment-score method is fast and correlates well with the oracles; however, the values are not probabilities and are only interpretable in a relative sense (for comparison to confidence estimates for other entities).

Method	Kendall's τ		Spearman's ρ	
	Rexa	Cora	Rexa	Cora
Query Assignment Score	*0.942*	*0.588*	*0.991*	*0.800*
MCMC Sampling	**0.950**	**0.869**	**0.995**	**0.967**
Query Perturbation	0.904	0.571	0.984	0.779
Conditioned Perturb.	0.635	0.646	0.812	0.780

Table 2: Confidence Rank Correlations to the Oracle

Method	Kendall's τ	Spearman's ρ
Query Assignment Score	**0.691**	**0.857**
MCMC Sampling	0.111	0.141
Query Perturbation	*0.343*	*0.452*
Conditioned Perturbation	0.044	0.083

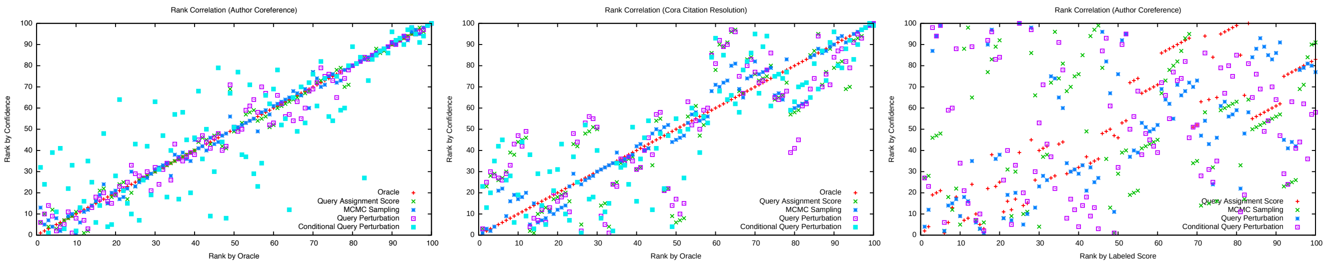
Table 3: Confidence Rank Correlations to the Labeled Score

4.2 Real-World Entity Resolution

In this experiment, we evaluate the approximate confidence estimation algorithms on the complete annotated Rexa dataset containing 1159 mentions ($\mathcal{M}$). Our query entities consist of 100 random sets of mentions from a predicted partitioning that are bigger than 5. Since computing the oracle confidence is intractable for this dataset, we treat the ground-truth loss functions (computed from the pairwise coreference labels) as a proxy for confidence, i.e. query mentions that appear as a single entity in the ground truth should have the highest confidence. We compare the rankings in Figure 1c and provide the correlations in Table 3. In this experiment, the most efficient algorithm is also the most accurate (query assignment score). The sampling based approaches do not work well because they were not able mix adequately in reasonable time. Future work will investigate more sophisticated sampling schemes such as tempering [12], distribution [13], and query-aware MCMC [16].

5. RELATED WORK

Although the field is still in its infancy, several recent



(a) vs Oracle on Rexa (small) (b) vs Oracle on Cora (small) (c) vs Labeled on Rexa

Figure 1: Scatter plots of Correlation Rank

approaches have taken significant steps towards estimating the confidence in predictions of information extraction systems. For models in which exact inference is possible, such as linear chain conditional random fields (CRF), exact inference may be used to estimate the joint marginal probability of the query assignment, this is similar to the *true confidence* estimation introduced in Section 3, and to the *Constrained Forward-Backward* algorithm for CRFs—a method that computes the probability of a query label subsequence given the data and model [2]. However, this approach is not practical for large and/or denser models that are common in information extraction, for example computing the partition function for clustering is exponential in complexity. For such models, sampling can be an efficient estimation scheme; we use MCMC as a sampling scheme for clustering, but it has also been used for long CRFs for which inference may be impractical [11].

Instead of estimating the confidence of predictions directly, some researchers estimate the confidence by training a separate model. The Open Language Learning for Information Extraction (OLLIE) framework, for example, uses the output of a logistic regression classifier that is trained on both positive and negative examples of extracted relations [9]. Outputs of multiple models trained on the same data can also be used to estimate this confidence, for example, training multiple perceptrons with varying number of hidden layers [6]. It is not quite clear how these ideas apply to much more complex models and the inference challenges that arise in automated knowledge base extraction. For example, it is possible to estimate the confidence in the coreference decision between a pair of mentions using a classifier, however extending this three or more set elements is non-trivial. For future work, it may be possible to combine such approaches with those that estimate confidence for clustering [7].

A separate approach from studying the properties of the model and/or inference is to use simple statistical rules to compute the confidence. For the Never Ending Language Learner (NELL), confidence of an extracted fact of a particular form is approximated by $1 - 0.5^c$, where c is the number of extracted facts of the same form [1]. Although such rules may work in many cases, they contain a number of disadvantages restricting their use in practice. For example, since the confidence increases exponentially as more facts are observed, noise in the data can have a significant impact on the confidence estimates. Further, because the facts are evaluated independently, this system could be confident in inconsistent facts.

In these previous approaches, confidence estimation has been identified as an extremely important component of information extraction pipeline, however the proposed approaches are restricted by their use of simple structured models and/or heuristic estimations of the confidence. Our work proposes a general framework for computing confidences in KB facts, and provides several alternate techniques to estimate this confidence for complex task of entity resolution.

6. DISCUSSION AND CONCLUSIONS

In this paper we proposed a list of desiderata for confidence values in KBs, proposed using marginal probabilities as a confidence measure, and finally proposed several approximate inference algorithms to estimate these marginal probabilities. We experimentally evaluated these methods on the problem of entity resolution in KBs: our results indicated that different methods are better suited for different desiderata (depending on whether speed, accuracy, or interpretability are important to the user or domain).

Although our initial results are promising, we caution that further experimentation is needed to verify if the proposed ideas are generically applicable. Different tasks in information extraction have a significantly different models and inference considerations, and a confidence estimation techniques that is efficient for one may not be practical for another. For example, the models used in relation extraction are significantly different from those used in entity resolution, and approaches such as belief propagation that cannot be used for many models of entity resolution may be a viable option for confidence estimation in relation extraction models.

We would also to describe a few shortcomings that became evident with using marginal probability as a confident measure. Since the marginal probability is defined in context of the configuration space of the variables, it does always provide intuitive values. For example, in coreference resolution, the marginal probability that a large set of mentions all refer to the same entity may be quite small because number of possible configurations in which the mentions are not coreferent far outnumbers the ones in which they do. In such cases the minute marginal probabilities may not be a suitable *absolute* measure of confidence. It is possible to mitigate this effect by give credit for partial entities, or by calibrating the model during the learning phase to yield values in the desired probability range. Another issue with our definition of marginal probabilities for confidence is that it primarily captures precision, not recall. As a result, in entity resolution, the confidence of a set of mentions monotonically decreases as new mentions are added to the set. This is not desirable behavior because more data often provides additional information that should make the model more confident about the entity.

In future work, we will address some of these shortcomings of the definition of confidence for predictions in a KB, and provide a more exhaustive empirical study of the various estimation techniques on a larger set of applications (such as relation extraction) and models.

7. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under NSF DGE-0907995.

8. REFERENCES

[1] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. H. Jr., and T. M. Mitchell. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)*, 2010.

[2] A. Culotta. Confidence estimation for information extraction. In *In Proceedings of Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT-NAACL*, 2004.

[3] A. Culotta, P. Kanani, R. Hall, M. Wick, and A. McCallum. Author disambiguation using error-driven machine learning with a ranking loss function. In *Sixth International Workshop on Information Integration on the Web (IIWeb-07)*, Vancouver, Canada, 2007.

[4] A. Culotta, M. Wick, and A. McCallum. First-order probabilistic models for coreference resolution. In *North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT)*, 2007.

[5] T. Finley and T. Joachims. Supervised clustering with support vector machines. In *International Conference on Machine Learning (ICML)*, pages 217–224, 2005.

[6] S. Gandrabur, G. Foster, and G. Lapalme. Confidence estimation for nlp applications. *ACM Trans. Speech Lang. Process.*, 3(3):1–29, Oct. 2006.

[7] F.-W. Gerstengarbe and P. Werner. A method to estimate the statistical confidence of cluster separation. *Theoretical and Applied Climatology*, 57(1-2):103–110, 1997.

[8] T. Kristjansson, A. Culotta, and P. Viola. Interactive information extraction with constrained conditional random fields. In *In AAAI*, pages 412–418, 2004.

[9] Mausam, M. Schmitz, S. Soderland, R. Bart, and O. Etzioni. Open language learning for information extraction. In *EMNLP-CoNLL*, pages 523–534, 2012.

[10] A. McCallum and B. Wellner. Toward conditional models of identity uncertainty with application to proper noun coreference. In *IJCAI Workshop on Information Integration on the Web*, 2003.

[11] A. Mejer and K. Crammer. Confidence in structured-prediction using confidence-weighted models. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 971–981. Association for Computational Linguistics, 2010.

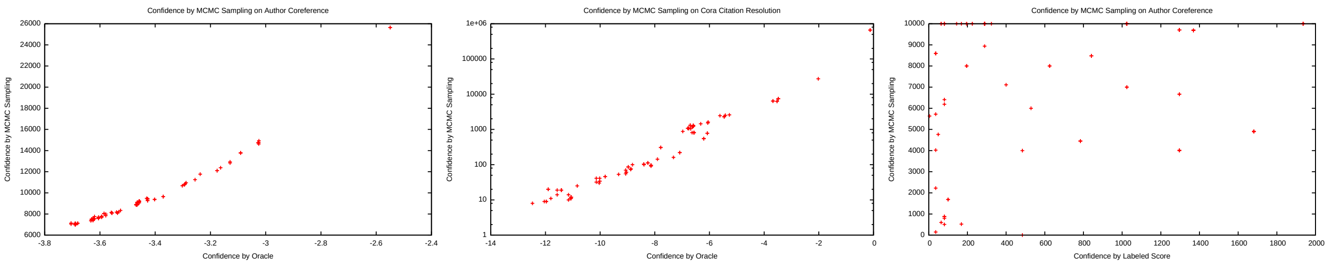
[12] R. M. Neal. Annealed importance sampling. *STATISTICS AND COMPUTING*, 11:125–139, 1998.

[13] S. Singh, A. Subramanya, F. Pereira, and A. McCallum. Large-scale cross-document coreference using distributed inference and hierarchical models. In *Association for Computational Linguistics: Human Language Technologies (ACL HLT)*, 2011.

[14] M. Wick, A. Culotta, K. Rohanimanesh, and A. McCallum. An entity-based model for coreference resolution. In *SIAM International Conference on Data Mining (SDM)*, 2009.

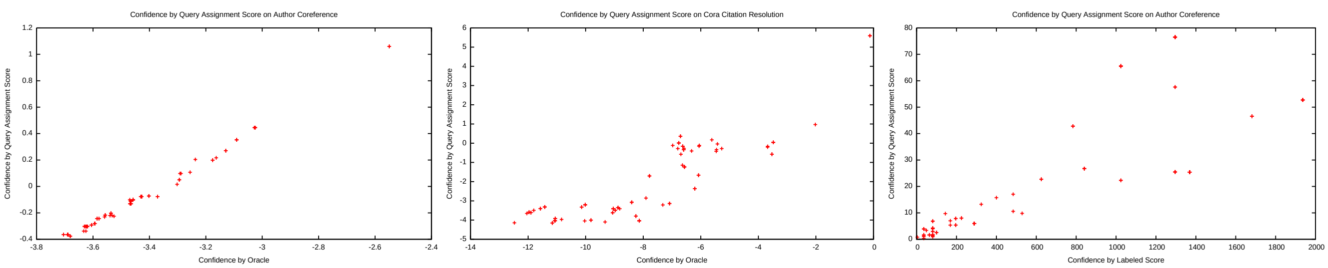
[15] M. Wick, S. Singh, and A. McCallum. A discriminative hierarchical model for fast coreference at large scale. In *Association for Computational Linguistics (ACL)*, 2012.

[16] M. L. Wick and A. McCallum. Query-aware McMC. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2564–2572. 2011.



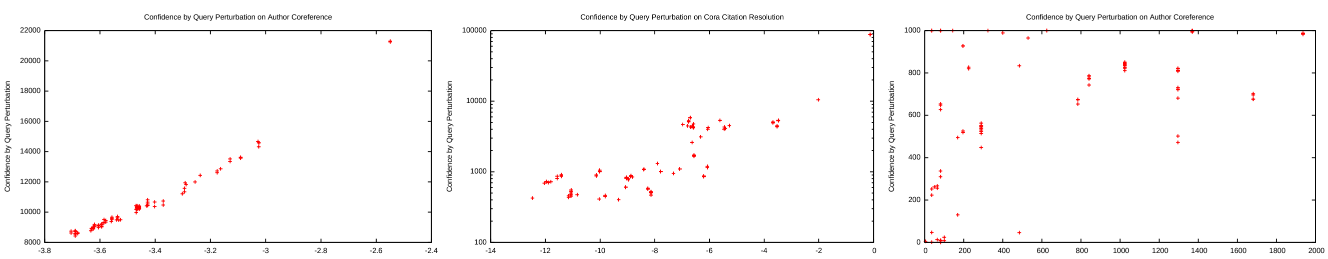
(a) vs Oracle on Rexa (small) (b) vs Oracle on Cora (small) (c) vs Labeled on Rexa

Figure 2: MCMC: Scatter plots of Confidence Scores



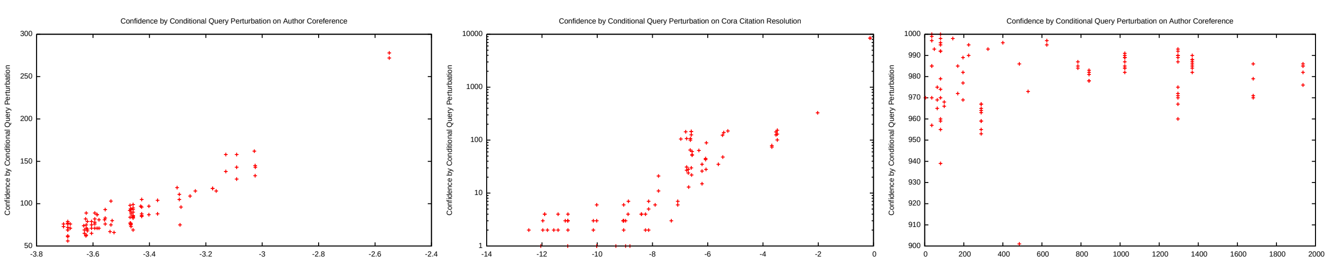
(a) vs Oracle on Rexa (small) (b) vs Oracle on Cora (small) (c) vs Labeled on Rexa

Figure 3: Query Assignment: Scatter plots of Confidence Scores



(a) vs Oracle on Rexa (small) (b) vs Oracle on Cora (small) (c) vs Labeled on Rexa

Figure 4: Query Perturbation: Scatter plots of Confidence Scores



(a) vs Oracle on Rexa (small) (b) vs Oracle on Cora (small) (c) vs Labeled on Rexa

Figure 5: Conditional Query Perturbation: Scatter plots of Confidence Scores