Human-Machine Cooperation with Epistemological DBs: Supporting User Corrections to Knowledge Bases

Michael Wick
University of Massachusetts
140 Governor's Drive
Amherst, MA 01002
mwick@cs.umass.edu

Karl Schultz
University of Massachusetts
140 Governor's Drive
Amherst, MA 01002
kschultz@cs.umass.edu

Andrew McCallum
University of Massachusetts
140 Governor's Drive
Amherst, MA 01002
mccallum@cs.umass.edu

Abstract

Knowledge bases (KB) provide support for real-world decision making by exposing data in a structured format. However, constructing knowledge bases requires gathering data from many heterogeneous sources. Manual efforts for this task are accurate, but lack scalability, and automated approaches provide good coverage, but are not reliable enough for real-world decision makers to trust. These two approaches to KB construction have complementary strengths: in this paper we propose a novel framework for supporting human-proposed edits to knowledge bases.

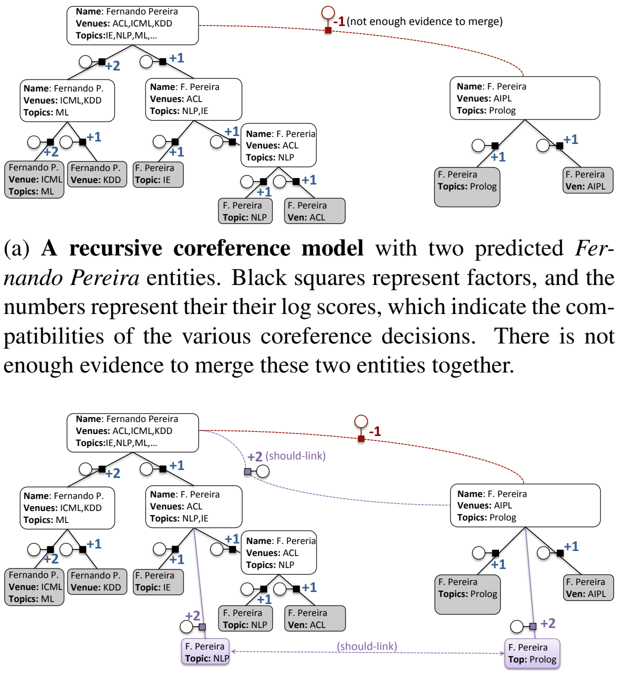
1 Introduction

Knowledge bases (KB) facilitate real-world decision making by providing access to structured relational information that enables pattern discovery and semantic queries. However, populating KBs requires the daunting task of gathering and assembling information from a variety of structured and unstructured sources at scale: a complex multi-task process riddled with uncertainty. Uncertainty about the reliability of different sources, uncertainty about the accuracy of extraction, uncertainty about integration ambiguity, and uncertainty about changes over time.

While this data can be gathered manually with high accuracy, it can be achieved at greater scale using automated approaches such as information extraction (IE). Indeed manual and automated approaches to knowledge base construction have complementary strengths: humans have high accuracy while machines have high coverage. However, integrating the two approaches is difficult because it is not clear how to best resolve conflicting assertions on knowledge base content. For example, it is risky to just allow users to directly modify the KB's notion of "the truth" because sometimes humans will be wrong, sometimes humans disagree, and sometimes the human edits become out-of-date in response to new events (and should be later over-written by IE).

We propose a new framework for supporting human edits to knowledge bases. Rather than treating each human edit as a deterministic truth, each edit is simply a new piece of evidence that can participate in inference with other pieces of raw evidence. In particular, a graphical model of "the truth" contains factors that weigh these various sources of evidence (documents culled by a web spider, outputs from IE systems, triples pulled from semantic web ontologies, rows streamed from external databases, etc.) against edits provided by enthusiastic groups of users. Inference runs in the background—forever—constantly improving the current best known truth. We call this an epistemological approach to KB construction because the truth is never observed (i.e., provided deterministically from humans or IE), rather, it is inferred from raw evidence with inference. Further, because the truth is simply a random variable in a graphical model, we can jointly reason about the value of the truth as well as the reliability of human edits (which we save for future work).

In the next section we describe the task of constructing a bibliographic KB, motivate the importance of coreference, and describe how to enable human edits in this context. Then we empirically


Name: Fernando Pereira Venues: ACL,ICML,KDD Topics:IE,NLP,ML,...
-1 (not enough evidence to merge)
+2
+1
Name: Fernando P. Venues: ICML,KDD Topics: ML
Name: F. Pereira Venues: ACL Topics: NLP,IE
Name: F. Pereira Venues: AIPL Topics: Prolog
+2
+1
+1
+1
Name: F. Pereria Venues: ACL Topics: NLP
+1
+1
Fernando P. Venue: ICML Topics: ML
Fernando P. Venue: KDD
F. Pereira Topic: IE
F. Pereira Topic: NLP
F. Pereira Ven: ACL
F. Pereira Topics: Prolog
F. Pereira Ven: AIPL
+1
+1
-1
+2 (should-link)
+2
+2
F. Pereira Topic: NLP
(should-link)
F. Pereira Top: Prolog

(a) **A recursive coreference model** with two predicted *Fernando Pereira* entities. Black squares represent factors, and the numbers represent their their log scores, which indicate the compatibilities of the various coreference decisions. There is not enough evidence to merge these two entities together.

(b) **How a human edit can correct the coreference error** in the previous figure. A human asserts that the "Prolog F. Pereira is also the NLP F. Pereira." This statement creates two mentions with a should-link constraint. During inference, the mentions are first moved into different entities. Then, when inference proposes to merge those two entities, the model gives a small bonus to this possible world because the two should-link mentions are placed in the same entity.

Figure 1: A recall coreference error **(top)**, is corrected when a user edit arrives **(bottom)**.

demonstrate that treating user edits as evidence allows corrections to propagate throughout the database resulting in an additional 43% improvement over an approach that deterministically treats edits as the truth. We also demonstrate robustness to incorrect human edits.

2 Supporting Human Edits in a Bibliographic KB

Reasoning about academic research, the people who create it, and the venues/institutions/grants that foster it is a current area of high interest because it has the potential to revolutionize the way scientific research is conducted. For example, if we could predict the next hot research area, or identify researchers in different fields who should collaborate, or facilitate the hiring process by pairing potential faculty candidates with academic departments, then we could rapidly accelerate and strengthen scientific research. A first step towards making this possible is gathering a large amount of bibliographic data, extract mentions of papers, authors, venues, and institutions, and perform massive-scale cross document entity resolution (coreference) and relation extraction to identify the real-world entities.

To this end, we implement a prototype "epistemological" knowledge base for bibliographic data. Currently, we have supplemented DBLP[1] with extra mentions from BibTeX files to create a database with over ten million mentions (6 million authors, 2.3 million papers, 2.2 million venues, and 500k institutions). We perform joint coreference between authors, venues, papers, and institutions at this scale. We describe our coreference model next.

2.1 Hierarchical Coreference inside the DB

Entity resolution is difficult at any scale, but is particularly challenging on large bibliographic data sets or other domains where there are large numbers of mentions. Traditional pairwise models (Soon et al., 2001; McCallum and Wellner, 2003) of coreference—that measure compatibility between pairs of mentions—lack both scalability and modeling power to process these datasets. Instead, inspired by a recently proposed three-tiered hierarchical coreference model (Singh et al., 2011), we employ an alternative model that recursively structures entities into trees. Rather than measuring compatibilities between *all* mention pairs, instead, internal tree nodes might summarize thousands of leaf-level mentions, and compatibilities are instead measured between child and parent nodes. For example, a single intermediate node might compactly summarize one-hundred "F. Pereira" mentions. Compatibility functions (factors) measure how likely a mention is to be summarized by this intermediate node. Further, this intermediate node may be recursively summarized by a higher level node in the tree. We show an example of this recursive coreference factor graph instantiated on two entities in Figure 1a.

For inference, we use a modified version of the Metropolis-Hastings algorithm that proposes multiple worlds for each sample (Liu et al., 2000). In particular, each proposal selects two tree nodes uni-

[1] http://dblp.uni-trier.de/xml/

formly at random. If the nodes happen to be in the same entity tree, then one of the nodes is made the root of a new entity. Otherwise, the two nodes are in different entity trees, then we propose to merge the two sub-tree's together by either merging the second subtree into the first subtree, or merging the second subtree into the root of the first subtree. If two leaf-level nodes (mentions) are chosen, then a new entity is created and the two mentions are merged into this newly created entity. We describe these proposals and the hierarchical coreference model in more detail in a forthcoming paper (Wick et al., 2012).

2.2 Human edits for entity resolution

Broadly speaking, there are two common types of errors for entity coreference resolution: recall errors, and precision errors. A recall error occurs when the coreference system predicts that two mentions do not refer to the same entity when they actually do. Conversely, a precision error occurs when the coreference error incorrectly predicts that two mentions refer to the same entity when in fact they do not. In order to correct these two common error types, we introduce two class of user edits: *should-link* and *should-not-link*. These edits are analogous to *must-link* and *must-not-link* constraints in constrained clustering problems; however, they are not deterministic, but extra suggestions via factors.

Each coreference edit in fact introduces two new mentions which are each annotated with the information pertinent to the edit. For example, consider the recall error depicted in Figure 1a. This is a real error that occurred in our system: there is simply not enough evidence for the model to know that these two *Fernando Pereira* entities are the same person because the co-authors do not overlap, the venues hardly overlap, and the topics they write about do not overlap. A user might notice this error and wish to correct it with an edit: "user X declared on this day that the Fernando Pereira who worked with Prolog is the same Fernando Pereira who works on natural language processing (NLP)". Presenting this edit to the bibliographic database involves creating two mentions, one with keywords about Prolog and the other with keywords about NLP, and both are annotated with a note indicating user X's belief: "user x: should-link". Then, special factors in the model are able to examine these edits in the context of other coreference decisions. As Markov chain Monte Carlo (MCMC) inference explores possible worlds by moving mentions between entities, the factor graph rewards possible worlds where the two mentions belong to the same entity. For example, see Figure 1b. In our experiments, a similar coreference error is corrected by an edit of this nature.

3 Experiments on Author Coreference

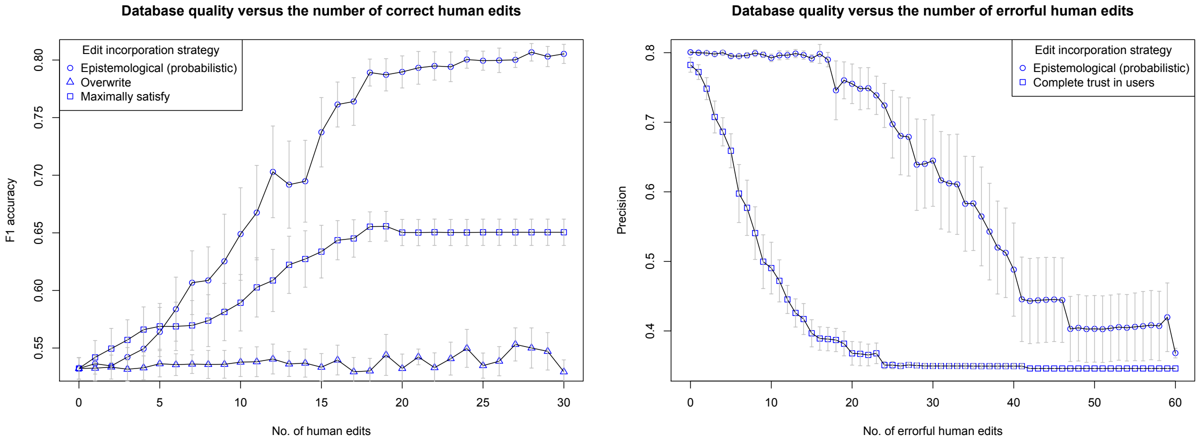
For the purpose of these experiments we focus on the problem of author coreference, which is a notoriously difficult problem due to common first and last names, spelling errors, extraction errors, and lack of "within document boundaries."

In order to evaluate our approach, we label a highly ambiguous "F. Pereira" dataset from BibTeX files.[2] We select this first-initial last name combination because it is fairly common in Portugal, Brazil and several other countries, and as a result there are multiple prominent researchers in the field of computer science. We construct this dataset with two strategies. First, from a publicly available collection of BibTeX files, we identify citation entries that have an author with last name "Pereira" and first name beginning with "F." Each of the *Pereira* mentions gathered in this manner are manually disambiguated by identifying the real-world author to which they refer. Second, we identified five prominent *Pereira* entities from the initial labeling and for three of them we were able to find their publication page and enter each publication into our dataset manually. The number of mentions in the five entities is as follows: (181 mentions, 92 mentions, 43 mentions, 7 mentions, 2 mentions).

3.1 Human edits

We argued earlier that users should not be allowed to directly edit the value of the truth because of the complications that may arise: domain-specific constraint/logical violations, disagreement about the truth, incorrect edits, etc. In this section, we test the hypothesis that the epistemological approach is better able to incorporate human edits than a more direct approach where users can directly edit the database content. To this end, we design two experiments to evaluate database quality as the number of

[2] http://www.iesl.cs.umass.edu/data/bibtex



(a) Good quality human edits

(b) Poor quality human edits

Figure 2: Sampling Performance Plots for 145k mentions

human edits increase. In the first experiment, we stream "good quality" human edits to the database, and in the second experiment we stream "poor quality" human edits (we will define what we mean by this in more detail later). For these experiments, we first create an initial database using the mentions in the "F. Pereira" dataset, and run MCMC until convergence reaching a precision of 80, and F1 of 54.

Next, given this initial database of predicted author entities, we measure how both "good quality" (correct) and "poor quality" (incorrect) human edits influence the initial quality of coreference. Although assessing the quality of a user edit is a subjective endeavor, we are still able to implement a relatively objective measure. In particular, we take the set of *Pereira* author entities initially discovered in the "original" DB and consider all possible pairs of these entities. If merging a pair into the same entity would increase the overall F1 score we consider this a correct human edit; if the merge would decrease the score we consider this an incorrect edit. Note that this reflects the types of edits that might be considered in a real-world bibliographical database where a user would browse two author pages and decide (correctly or incorrectly) that they should be the same entity. For example, one of the good quality pairs we discover in this way encodes the simulated "user's" belief that the "the Fernando Pereira who works on NLP is the same Fernando Pereira who works on machine learning". An example of a poor quality edit is "the Fernando Pereira that researches NLP is the same Fernando Pereira that works on MPEG compression".

Once we have determined which author pairs result in higher or lower F1 accuracy, we can then construct simulated edits of various quality. We consider three ways of incorporating these edits into the database. The first approach, *epistemological*, which we advocate in this paper, is to treat the edits as evidence and incorporate them statistically with MCMC. We convert each entity pair into edit-evidence as follows: two mentions are created (one for each entity), the attributes of the entities are copied into the features of these corresponding mentions, and a *should-link* constraint is placed between the mentions. The second two approaches simulate users who directly modify the database content. The first baseline, *overwrite*, resolves conflicts by simply undo-ing previous edits and overwriting them, and the second baseline, *maximally satisfy*, applies all edits by taking their transitive closure.

Good quality edits

In Figure 2a we compare our *epistemological* approach to the two baselines *overwrite* and *maximally satisfy* on the set of good user edits (averaged over 10 random runs). What is interesting about this result is that the *epistemological* approach, which is not obligated to merge the edited entities, is actually substantially better than the two baselines (which are deterministically required to merge the edited entities (provided by a ground truth signal)). After some error analysis, we determine that a major reason for this improvement is that the user edits propagate beyond the entity pair they were initially intended to merge. In particular, as the user edits become applied, the quality of the entities increase. As the quality of the entities increase, the model is able to make more accurate decisions about other mentions that were errorfully merged. For example, we observed that after MCMC inference merged the *natural language processing Fernando* with the *machine learning Fernando*, that an additional 18 mentions were correctly incorporated into the new cluster by inference. In a traditional approach, these corrections could not propagate thus placing the burden on the users to provide additional edits.

Poor quality user edits

In Figure 2b we evaluate the robustness of our *epistemological* database to poor quality (incorrect) human edits. In this figure, we evaluate quality in terms of precision instead of F1 so that we can more directly measure resistance to the over-zealous recall-oriented errorful must-link edits. The baseline approach that deterministically incorporates the errorful edits suffers rapid loss of precision as entities become merged that should not be. In contrast, the *epistemological* approach is able to veto many errorful edits when there is enough evidence to warrant such an action (the system is completely robust for twenty straight errorful edits). Surprisingly, the F1 (not shown) of the epistemological database actually increases with some errorful edits because some of the edits are partially correct, indicating that the this approach is well suited for incorporating partially correct information.

4 Related Work

An example of a structured database where there is active research in harnessing user feedback is the DBLife project (DeRose et al., 2007). Chai et al. (Chai et al., 2009) propose a solution that exposes the intermediate results of extraction for users to edit directly. However, their approach deterministically integrates the user edits into the database and may potentially suffer from many of the issues discussed earlier; for example, conflicting user edits are resolved arbitrarily, and incorrect edits can potentially overwrite correct extractions or correct user edits.

There has also been recent interest in using probabilistic models for correcting the content of a knowledge base. For example, Kasneci et al. (Kasneci et al., 2010) use Bayesian networks to incorporate user feedback into an RDF semantic web ontology. Here users are able to assert their belief about facts in the ontology being true or false. The use of probabilistic modeling enables them to simultaneously reason about user reliability and the correctness of the database. However, there is no observed knowledge base content taken into consideration when making these inferences. In contrast, we jointly reason over the entire database as well as user beliefs, allowing us to take all available evidence into consideration. Koch et al (Koch and Olteanu, 2008) develop a data-cleaning “conditioning” operator for probabilistic databases that reduces uncertainty by ruling-out possible worlds. However, the evidence is incorporated as constraints that eliminate possible worlds. In contrast, we incorporate the evidence probabilistically which allows us to reduce the probability of possible worlds without eliminating them entirely; this gives our system the freedom to revisit the same inference decisions not just once, but multiple times if new evidence arrives that is more reliable.

5 Conclusion

In this paper we described a new framework for combining human edits with automated information extraction for improved knowledge base construction. We demonstrated that our approach was better able to incorporate “correct” human edits, and was more robust to “incorrect” human edits.

6 Acknowledgments

This work was supported in part by the CIIR, in part by ARFL under prime contract #FA8650-10-C-7059, in part by DARPA under AFRL prime contract #FA8750-09-C-0181, and in part by IARPA via DoI/NBC contract #D11PC20152. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

Xiaoyong Chai, Ba-Quy Vuong, AnHai Doan, and Jeffrey F. Naughton. 2009. Efficiently incorporating user feedback into information extraction and integration programs. In *SIGMOD Conference*, pages 87–100.

Pedro DeRose, Warren Shen, Fei Chen, Yoonkyong Lee, Douglas Burdick, AnHai Doan, and Raghu Ramakrishnan. 2007. Dblife: A community information management platform for the database research community. In *CIDR*, pages 169–172.

Gjergji Kasneci, Jurgen Van Gael, Ralf Herbrich, and Thore Graepel. 2010. Bayesian knowledge corroboration with logical rules and user feedback. In *Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part II*, ECML PKDD'10, pages 1–18, Berlin, Heidelberg. Springer-Verlag.

Christoph Koch and Dan Olteanu. 2008. Conditioning probabilistic databases. *Proc. VLDB Endow.*, 1:313–325, August.

Jun S. Liu, Faming Liang, and Wing Hung Wong. 2000. The multiple-try method and local optimization in metropolis sampling. *Journal of the American Statistical Association*, 96(449):121–134.

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

Sameer Singh, Amarnag Subramanya, Fernando C. N. Pereira, and Andrew McCallum. 2011. Large-scale cross-document coreference using distributed inference and hierarchical models. In *ACL*, pages 793–803.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Comput. Linguist.*, 27(4):521–544.

Michael Wick, Sameer Singh, and Andrew McCallum. 2012. A discriminative hierarchical model for fast coreference at large scale. In *Association for Computational Linguistics (ACL)*.