Automatic Knowledge Base Construction using Probabilistic Extraction, Deductive Reasoning, and Human Feedback

**Yang Chen**, **Sean Goldberg**, **Christian Grant**, **Kun Li**, **Daisy Zhe Wang**

Database Research Center, CISE, University of Florida

{yang, sean, cgrant, kli, daisyw} @cise.ufl.edu

**ABSTRACT**

We envision an automatic knowledge base construction system consisting of three interrelated components. **MADDen** is a knowledge extraction system applying statistical text analysis methods over database systems (DBMS) and massive parallel processing (MPP) frameworks; **PROBKB** performs probabilistic reasoning over the extracted knowledge to derive additional facts not existing in the original text corpus; **CAMEL** leverages human intelligence to reduce the uncertainty resulting from both the information extraction and probabilistic reasoning processes.

**MADDEN**

MADDen is our system for in-database text analytics over structured and unstructured text.

**MADDEN Features**

- SQL interface for non-trivial ad hoc queries.
- A library of functions for deep analytic processing.
- Built over PostgreSQL and a similar massively parallel database management system.
- Function may mix external API as a part of DBMS query execution.

**PROBKB**

PROBKB is a probabilistic knowledge base that derives implicit knowledge from entities, relations, and rules extracted from a text corpus by knowledge extraction systems like MADDen. Specifically, PROBKB:

- Uses facts and rules extracted from web. These facts and rules are often uncertain due to the heterogeneous data sources and the inherently probabilistic nature of the extraction algorithms. PROBKB is designed to deal with these uncertainties.
- PROBKB models uncertain rules as Markov logic networks (MLNs) and tries to scale up inference in the resulting Markov network. Potential research directions are rule/data partitioning and parallelizing inference algorithms.

**CHALLENGES AND CURRENT WORK**

- Grounding and inference over large-scale MLNs.
- Probabilistic integration of turkers and ML algorithms.
- Influential node selection from CRF/BNet.
- Learning algorithm implementation in DBs.
- Large-scale coreference.

**CAMEL**

Crowd-Assisted Machine Learning, or CAMEL, provides a human feedback element to probabilistic knowledge bases like PROBKB. Facts difficult for the machine to extract are resolved by crowdsourcing queries through Amazon Mechanical Turk. The key steps to CAMEL are **Selection** and **Integration**:

- Selection: PROBKB contains extracted information labeled according to some distribution. CAMEL selects knowledge that is most uncertain, corresponding to the highest entropy of the distribution, and packages a Human Intelligence Task for users to respond to.
- Integration: For consistency and correctness, generally 3-5 users respond to each task and the results are aggregated. Instead of the standard majority voting, CAMEL uses Dempster-Shafer theory to form a new probability distribution over the crowd results which can be inputted back into the knowledge base.

**ACKNOWLEDGEMENTS**

This material is based upon work supported by the National Science Foundation Graduate Fellowship under Grant No. DGE-0802270 and a generous gift from Greenplum.