Elementary: Large-Scale Knowledge-Base Construction via Machine Learning and Statistical Inference

Feng Niu, Computer Sciences Department, University of Wisconsin-Madison, USA
Ce Zhang, Computer Sciences Department, University of Wisconsin-Madison, USA
Christopher Ré, Computer Sciences Department, University of Wisconsin-Madison, USA
Jude Shavlik, Computer Sciences Department, University of Wisconsin-Madison, USA

ABSTRACT

Researchers have approached knowledge-base construction (KBC) with a wide range of data resources and techniques. The authors present Elementary, a prototype KBC system that is able to combine diverse resources and different KBC techniques via machine learning and statistical inference to construct knowledge bases. Using Elementary, they have implemented a solution to the TAC-KBP challenge with quality comparable to the state of the art, as well as an end-to-end online demonstration that automatically and continuously enriches Wikipedia with structured data by reading millions of webpages on a daily basis. The authors describe several challenges and their solutions in designing, implementing, and deploying Elementary. In particular, the authors first describe the conceptual framework and architecture of Elementary to integrate different data resources and KBC techniques in a principled manner. They then discuss how they address scalability challenges to enable Web-scale deployment. The authors empirically show that this decomposition-based inference approach achieves higher performance than prior inference approaches. To validate the effectiveness of Elementary’s approach to KBC, they experimentally show that its ability to incorporate diverse signals has positive impacts on KBC quality.

Keywords: Information Extraction, Knowledge-Base Construction, Machine Learning, Machine Reading, Natural Language Understanding, Statistical Inference, Systems

INTRODUCTION

Knowledge-base construction (KBC) is the process of populating a knowledge base (KB) with facts (or assertions) extracted from text. It has recently received tremendous interest from academia (Weikum & Theobald, 2010), e.g., CMU’s NELL (Carlson, Betteridge, Kisiel, Settles, Hruschka, & Mitchell, 2010; Lao, Mitchell, & Cohen, 2011), MPI’s YAGO (Kasneci, Ramanath, Suchanek, & Weikum, 2008; Nakashole, Theobald, & Weikum, 2011), and from industry (Fang, Sarma, Yu, & Bohannon, 2011).
To construct high-quality knowledge bases from text, researchers have considered a wide range of data resources and techniques; e.g., pattern matching with dictionaries listing entity names (Riloff, 1993), bootstrapping from existing knowledge bases like Freebase and YAGO (Suchanek, Kasneci, & Weikum, 2007), disambiguation using web links and search results (Hoffart, Yosef, Bordino, Fürstenau, Pin, Spaniol, ... Weikum, 2011; Dredze, McNamee, Rao, Gerber, & Finin, 2010), rule-based extraction with regular expressions curated by domain experts (Derose, Shen, Fei, Lee, Burdick, Doan, & Ramakrishnan, 2007; Chiticariu, Krishnamurthy, Li, Raghavan, Reiss, & Vaithyanathan, 2010), training statistical models with annotated text (Lafferty, McCallum, & Pereira, 2001), etc. All these resources are valuable because they are complementary in terms of cost, quality, and coverage; ideally one would like to be able to use them all. To take advantage of different kinds of data resources, a major problem that KBC systems face is coping with imperfect or conflicting information from multiple sources (Weikum & Theobald, 2010). (We use the term “information” to refer to both data and algorithms that can be used for a KBC task.) To address this issue, several recent KBC projects (Carlson et al., 2010; Kasneci et al., 2008; Nakashole et al., 2011; Zhu et al., 2009; Lao et al., 2011) use statistical inference to combine different data resources.

Motivated by the above observation, we present Elementary, a prototype system that aims to enable quick development and scalable deployment of KBC systems that combine diverse data resources and best-of-breed algorithms via machine learning and statistical inference (Niu, 2012). This article provides an overview of the motivation and advantages of the Elementary architecture, while only briefly touching on individual technical challenges that are addressed in our other publications. We structure our presentation around two main challenges that we face in designing, implementing, and deploying Elementary: (1) how to integrate conflicting information from multiple sources for KBC in a principled way, and (2) how to scale Elementary for Web-scale KBC.

**Challenge 1: How do we handle conflicting information from multiple sources of data and algorithms for a KBC task in a principled way?** To perform knowledge-base construction with multiple sources of information, a critical challenge is the ability to handle imperfect and conflicting information (Challenge 1a). Following several recent KBC projects such as StatSnowball/EntityCube (Zhu et al., 2009) and SOFIE/Prospera (Suchanek, Sozio, & Weikum, 2009; Nakashole et al., 2011), Elementary uses a probabilistic logic language called Markov logic (Richardson & Domingos, 2006) that has been applied to a wide range of text-related applications (Poon & Domingos, 2007; Suchanek et al., 2009; Zhu et al., 2009; Andrzejewski, Livermore, Zhu, Craven, & Recht, 2011). In Markov logic, one can write first-order logic rules with weights (that intuitively model our confidence in a rule); this allows one to capture rules that are likely, but not certain, to be correct. A Markov logic program (aka Markov logic network, or simply MLN) specifies what data are available, what predictions to make, and what constraints and correlations there are.

Elementary adopts the classic Entity-Relationship (ER) model (Chen, 1976; Cali, Gottlob, & Pieris, 2010) for the target KB. While there is a direct map between the ER model and predictions in Markov logic, a crucial challenge is how to accommodate information from diverse sources when developing a KBC system (Challenge 1b). To address this issue, Elementary uses a two-phase architecture: as shown in Figure 1, Elementary first runs a feature-extraction step to convert input data resources into relational signals called evidence; it then feeds evidence into a Markov logic engine where Elementary performs statistical inference to construct a knowledge base. Besides evidence, one may also translate a source of information into MLN rules. For example, we can translate machine-learning models like logistic regression and conditional random fields (Lafferty et al., 2001) into MLN...