Automating the Generation of Hardware Component Knowledge Bases

Luke Hsiao
Stanford University, USA
lwhsiao@stanford.edu

Sen Wu
Stanford University, USA
senwu@cs.stanford.edu

Nicholas Chiang
Gunn High School, USA

Christopher Ré
Stanford University, USA
chrismre@cs.stanford.edu

Philip Levis
Stanford University, USA
pal@cs.stanford.edu

Abstract

Hardware component databases are critical resources in designing embedded systems. Since generating these databases requires hundreds of thousands of hours of manual data entry, they are proprietary, limited in the data they provide, and have many random data entry errors.

We present a machine-learning based approach for automating the generation of component databases directly from datasheets. Extracting data directly from datasheets is challenging because: (1) the data is relational in nature and relies on non-local context, (2) the documents are filled with technical jargon, and (3) the datasheets are PDFs, a format that decouples visual locality from locality in the document. The proposed approach uses a rich data model and weak supervision to address these challenges.

We evaluate the approach on datasheets of three classes of hardware components and achieve an average quality of 75 F1 points which is comparable to existing human-curated knowledge bases. We perform two applications studies that demonstrate the extraction of multiple data modalities such as numerical properties and images. We show how different sources of supervision such as heuristics and human labels have distinct advantages which can be utilized together within a single methodology to automatically generate hardware component knowledge bases.

CCS Concepts • **Information systems** → *Data mining*; • **Hardware** → *Electronic design automation.*

Keywords Knowledge Base Construction, Design Tools

LCTES '19, June 23, 2019, Phoenix, AZ, USA

ACM ISBN 978-1-4503-6724-0/19/06…$15.00
https://doi.org/10.1145/3316482.3326344

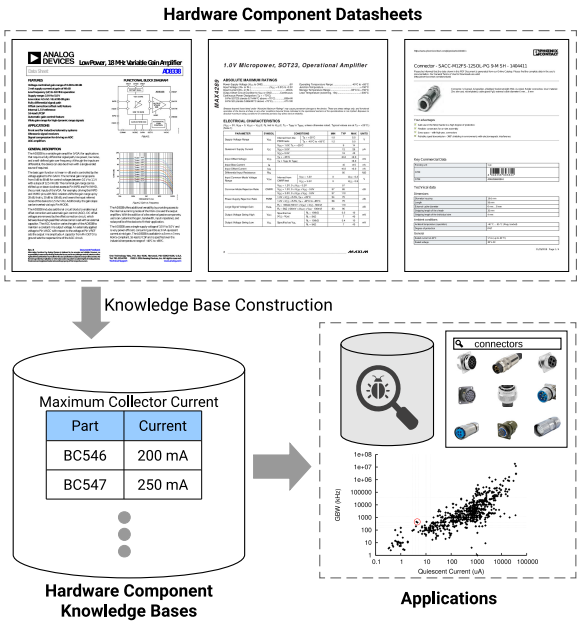

Figure 1. Hardware component knowledge bases are populated from datasheets and serve valuable applications such as cross validation, selecting components based on optimal electrical characteristics, or building rich search interfaces.

ACM Reference Format:
Luke Hsiao, Sen Wu, Nicholas Chiang, Christopher Ré, and Philip Levis. 2019. Automating the Generation of Hardware Component Knowledge Bases. In *Proceedings of the 20th ACM SIGPLAN/SIGBED Conference on Languages, Compilers, and Tools for Embedded Systems (LCTES '19), June 23, 2019, Phoenix, AZ, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3316482.3326344

1 Introduction

Creating embedded systems often requires developing new hardware. Searching for components that best meet system requirements constitutes a significant portion of design time. Downloading a datasheet is easy, but figuring out *which* datasheet to download is hard. The needed information is hidden in the datasheet itself, a complex document that is impenetrable to standard search engines. Requirements are