Automating the Generation of Hardware Component Knowledge Bases

Luke Hsiao, Sen Wu, Nicholas Chiang[1], Christopher Ré, and Philip Levis

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

Stanford University and [1]Gunn High School

Overview

Introduction

Motivation: Hardware Component Selection is Hard

The process today...

- Creating embedded systems often requires developing new hardware.
- Searching for components that best meet system requirements is a significant portion of design time.
- Visit many web search pages, tuning parameters on each to get a handful of results, then inspect datasheets manually.

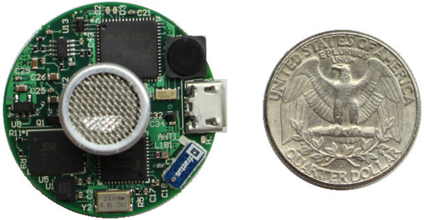


Figure 1: The Opo Sensor [1]

Motivation: Hardware Component Selection is Hard

The process today...

- Creating embedded systems often requires developing new hardware.
- Searching for components that best meet system requirements is a significant portion of design time.
- Visit many web search pages, tuning parameters on each to get a handful of results, then inspect datasheets manually.

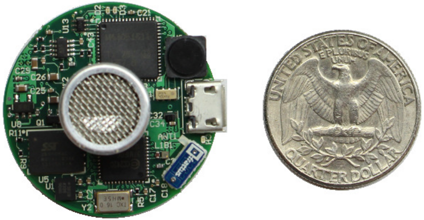


Figure 1: The Opo Sensor [1]

Downloading a datasheet is easy, but figuring out which datasheet to download is hard.

Motivation: The Opo Sensor Analysis

Operational Amplifier Requirements

- $1000\times$ gain to detect ultrasonic signal.
- Minimize number of gain stages.
- Low total current draw to preserve battery life.

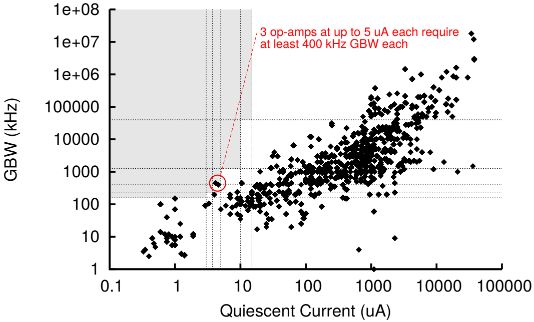


Figure 2: Original Opo Analysis using Digi-Key [1]

Motivation: The Opo Sensor Analysis

Operational Amplifier Requirements

- 1000× gain to detect ultrasonic signal.
- Minimize number of gain stages.
- Low total current draw to preserve battery life.

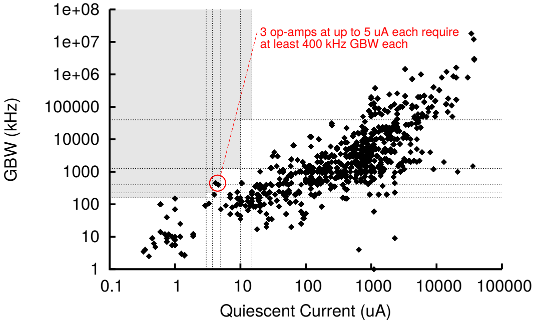


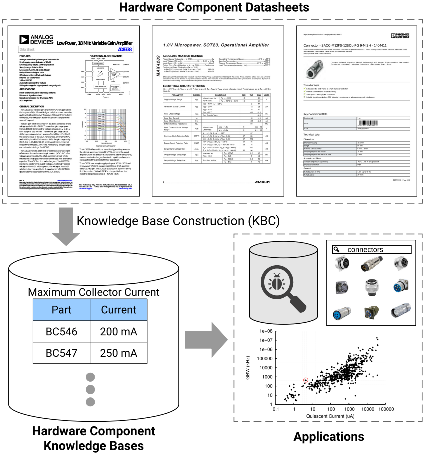
Figure 2: Original Opo Analysis using Digi-Key [1]

What if there was no Digi-Key?

Automating the Generation of Hardware Component Knowledge Bases

Contributions

1. A **general methodology** for building hardware component knowledge bases using state-of-the-art machine learning.
2. The evaluation of our methodology on **multiple hardware components**, extracting both **textual and non-textual** information.
3. **Application studies**, which highlight how these databases make hardware component selection easier.



The Challenges of PDF Datasheets

1. **Relational Data**: Traditional text-based search is insufficient.
2. **Technical Jargon**: Requires expertise to understand.
3. **Input Format**: Immense data variety in styles and formats makes heuristics insufficient.

Challenge 1: Relational Data

Relational Data

- A keyword search for "V_{OS}" and "1" may match 1000s of documents as both terms are commonly used.
- Instead, engineers want to query relational data (e.g., whether a specific part has a minimum "V_{OS}" value of "1 μV").
- Traditional unstructured text-based search is insufficient.

MAX44259/260/261/263

ABSOLUTE MAXIMUM RATINGS

IN+, IN-, OUT ....(V_{SS} - 0.3V) to (V_{DD} + 0.3V)
V_{DD} to V_{SS}...-0.3V to +6V
$\overline{\text{SHDN}}$, $\overline{\text{CAL}}$ ....-0.3V to +6V

ELECTRICAL CHARACTERISTICS

(V_{DD} = 3.3V, V_{SS} = 0V, V_{IN+} = V_{IN-} = V_{DD}/2, R_L = 10kΩ

PARAMETER		MIN	TYP	MAX	UNITS
DC CHARACTERISTICS					
V_{IN+} V_{IN-}		-0.1		VDD + 0.1	V
V_{OS}	MAX44259	10	50	800	μV
	MAX44260			100	
	MAX44261			500	
	MAX44263		100	800	

Challenge 2: Technical Jargon

Technical Jargon

- Datasheets use extensive technical jargon such as the symbols highlighted in red.
- Understanding a datasheet requires both technical expertise and deep experience.
- This precludes relying on untrained crowdsourcing services.

MAX44259/260/261/263

ABSOLUTE MAXIMUM RATINGS

IN+, IN-, OUT ....(V_{SS} - 0.3V) to (V_{DD} + 0.3V)
V_{DD} to V_{SS} ....-0.3V to +6V
SHDN, CAL ....-0.3V to +6V

ELECTRICAL CHARACTERISTICS

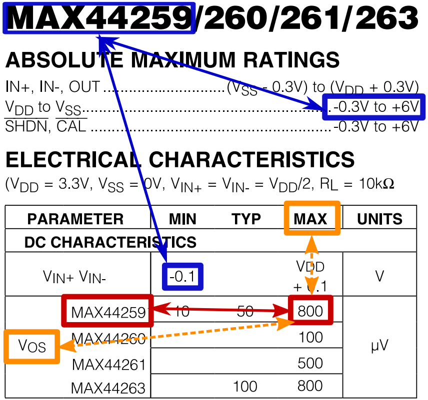
(V_{DD} = 3.3V, V_{SS} = 0V, V_{IN+} = V_{IN-} = $V_{DD}/2$, R_L = 10kΩ

PARAMETER		MIN	TYP	MAX	UNITS
DC CHARACTERISTICS					
V_{IN+} V_{IN-}		-0.1		VDD + 0.1	V
V_{OS}	MAX44259	10	50	800	μV
	MAX44260			100	
	MAX44261			500	
	MAX44263		100	800	

Challenge 3: Input Format

Input Format

- PDF documents lack structural information (e.g, explicit tables).
- Relationship must be inferred from the rendering of the text, vectors, and images.
- Cues like alignments, proximity, and emphasis are understandable to humans, but challenging for machines to interpret.
- The variety and non-uniformity of cues makes them difficult to address with heuristics.



MAX44259/260/261/263

ABSOLUTE MAXIMUM RATINGS

IN+, IN-, OUT ....(V_{SS} - 0.3V) to (V_{DD} + 0.3V)
V_{DD} to V_{SS} .... -0.3V to +6V
SHDN, CAL .... -0.3V to +6V

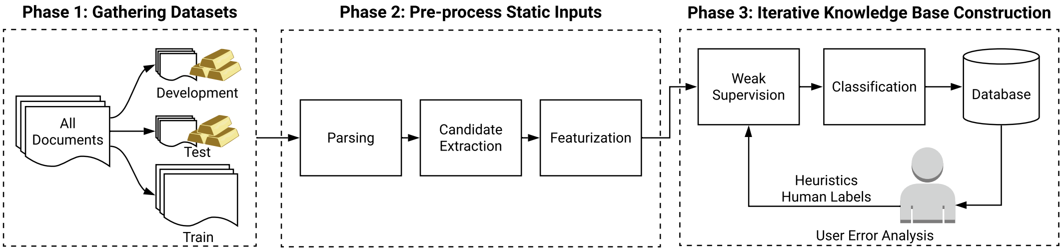
ELECTRICAL CHARACTERISTICS

(V_{DD} = 3.3V, V_{SS} = 0V, V_{IN+} = V_{IN-} = V_{DD}/2, R_L = 10kΩ

PARAMETER		MIN	TYP	MAX	UNITS
DC CHARACTERISTICS					
V_{IN+} V_{IN-}		-0.1		V_{DD} + 0.1	V
V_{OS}	MAX44259	10	50	800	μV
	MAX44260			100	
	MAX44261			500	
	MAX44263		100	800	

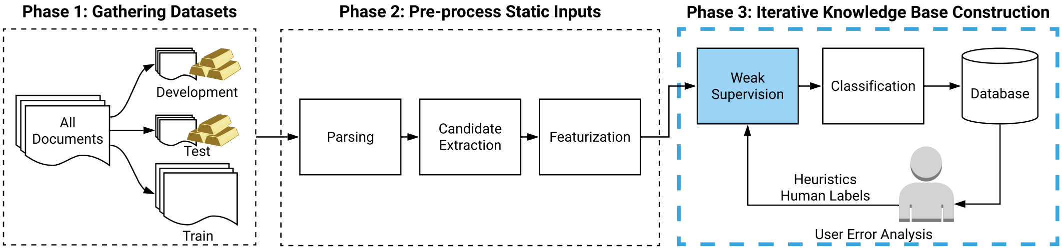
Methodology

A machine-learning approach



We formulate this as a weakly supervised machine-learning classification problem.

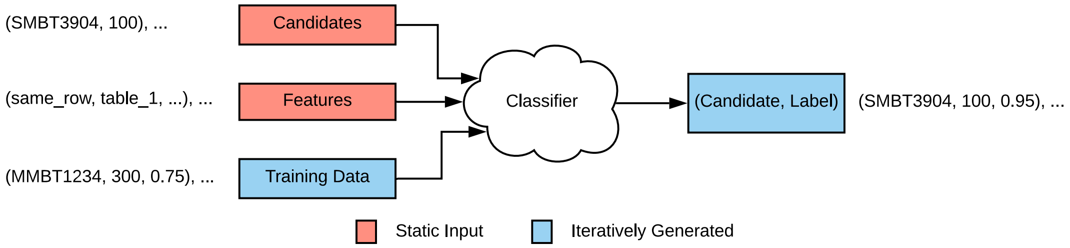
Generate Training Data with Weak Supervision



We use **weak supervision** [2] to generate training data.

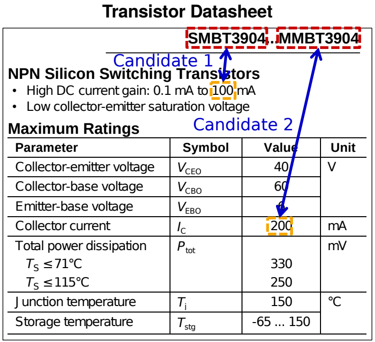
Inputs and Outputs for Supervised Classification

Populating the schema (Part Number, Maximum Collector Current) from transistor datasheets:



Rather than tuning features, we refine the training data itself.

Weak Supervision with Labeling Functions



Parameter	Symbol	Value	Unit
Collector-emitter voltage	V_{CEO}	40	V
Collector-base voltage	V_{CBO}	60	
Emitter-base voltage	V_{EBO}	6	
Collector current	I_C	200	mA
Total power dissipation	P_{tot}		mV
$T_S \leq 71$°C		330	
$T_S \leq 115$°C		250	
J unction temperature	T_j	150	°C
Storage temperature	T_{stg}	-65 ... 150	

Use **labeling functions** to programmatically apply weak supervision.

- Output true, false, or abstain from voting.
- Leverage heuristics, human annotations, etc.
- Relies on rich information captured by the Fonduer data model [4].

```
# Check if current is in same row as keyword "collector"
#   (SMBT3904, 100) -> ABSTAIN
#   (MMBT3904, 200) -> TRUE
def in_the_same_row_with(candidate):
    if "collector" in row_ngrams(candidate.current):
        return TRUE
    else:
        return ABSTAIN
```

Modeling Labeling Function Accuracy

Input

Candidate	LF 1	LF 2	LF 3
(SMBT3904, 100)	✖	⊘	✖
(SMBT3904, 200)	✔	✔	✖
(SMBT3904, 430)	✖	✖	⊘

Output

Candidate	Training Labels
(SMBT3904, 100)	0.23
(SMBT3904, 200)	0.85
(SMBT3904, 430)	0.15

Intuition: Data Programming

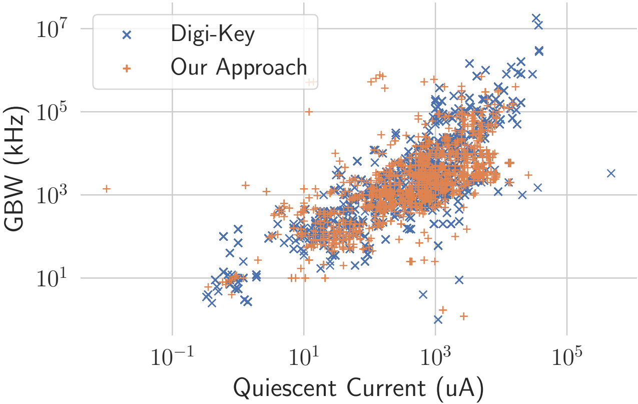
Use coverage, agreements, and disagreements to model the accuracy of each labeling function without ground truth [3].

Results

Electrical Characteristic Analysis

Our Approach vs. Digi-Key

- We identify the same Micrel MIC861/863 amplifier.
- Largely overlaps with Digi-Key.
- Our approach builds a database directly from PDF datasheets.
- Can be applied to new components or characteristics.



Comparing to Human-curated Knowledge Bases

Table 1: Quality of our approach vs. Digi-Key for compared to expert annotations.

Relation	Source	Precision	Recall	F1 score
Polarity	Digi-Key	**1.00**	0.67	0.80
	Our Approach	0.94	**0.94**	**0.94**
Max Collector-Emitter Volt.	Digi-Key	**0.97**	0.67	**0.79**
	Our Approach	0.75	**0.77**	0.76
Gain Bandwidth Product	Digi-Key	**0.91**	0.62	0.74
	Our Approach	0.88	**0.84**	**0.86**
Quiescent Current	Digi-Key	**0.93**	0.45	0.61
	Our Approach	0.89	**0.80**	**0.84**

- On average: improves on Digi-Key by 12 F1 points (recall +24 % and precision −9 %).
- Shifts class of errors from random human errors to systematic errors.

Conclusion

Summary

Hardware component knowledge bases empower academic research as well as industrial applications by making hardware data accessible.

Contributions

1. A **general methodology** for building hardware component knowledge bases using weak supervision on richly formatted data like PDF datasheets.
2. We achieve an average of 75 F1 points on **multiple hardware components**, extracting both **textual and non-textual** information, which is comparable with existing human-curated knowledge bases.
3. **Application studies**, which highlight how these databases make hardware component selection easier.

Questions? Come check out our poster!

References

[1] W. Huang, Y.-S. Kuo, P. Pannuto, and P. Dutta.
Opo: a wearable sensor for capturing high-fidelity face-to-face interactions.
In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems*, pages 61–75. ACM, 2014.

[2] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré.
Snorkel: Rapid training data creation with weak supervision.
Proceedings of the VLDB Endowment, 11(3):269–282, 2017.

[3] A. J. Ratner, C. M. De Sa, S. Wu, D. Selsam, and C. Ré.
Data programming: Creating large training sets, quickly.
In *Advances in Neural Information Processing Systems*, pages 3567–3575, 2016.

[4] S. Wu, L. Hsiao, X. Cheng, B. Hancock, T. Rekatsinas, P. Levis, and C. Ré.
Fonduer: Knowledge base construction from richly formatted data.
In *Proceedings of the 2018 International Conference on Management of Data*, pages 1301–1316. ACM, 2018.

Appendix

Dataset Summary

Table 2: Summary of the datasets used in our evaluation based on their size, number of files, average number of pages per document, and the number of relations extracted.

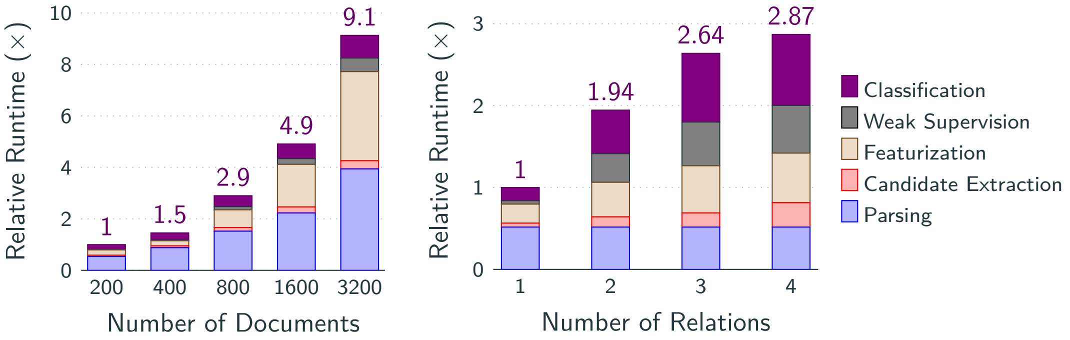
Dataset	**Size**	**#Docs**	**#Pgs/Doc**	**#Rels**
Bipolar Junction Transistors	3GB	7.0k	5.5	4
Circular Connectors	3GB	5.1k	3.2	1
Operational Amplifiers	5GB	3.3k	23.3	2

End-to-End Quality

Table 3: End-to-end quality in term of precision, recall, and F1 score for each dataset.

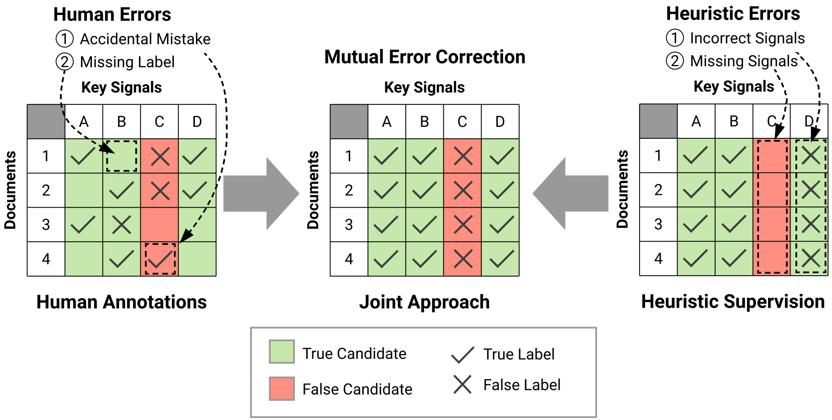
Dataset	Relation	Precision	Recall	F1 score
Trans.	Min. Storage Temp.	1.00	0.58	0.74
	Max. Storage Temp.	0.95	0.61	0.74
	Polarity	0.88	0.92	0.90
	Max. Collector-Emitter Volt.	0.85	0.77	0.81
Op. Amps.	Gain Bandwidth Product	0.72	0.76	0.74
	Quiescent Current	0.65	0.54	0.59
Circ. Conn.	Product Thumbnails	0.63	0.83	0.72

Performance at Scale



Runtime scales sub-linearly with documents and relations.

Benefits of a Joint Approach

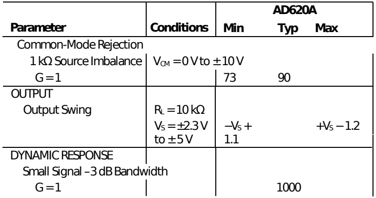


Our approach can benefit from both the recall of human annotations and the systematic consistency and prevision of heuristics-based weak supervision.

Future Work: Formatting Challenges

(a) Scanned documents

(b) Vector-drawn text



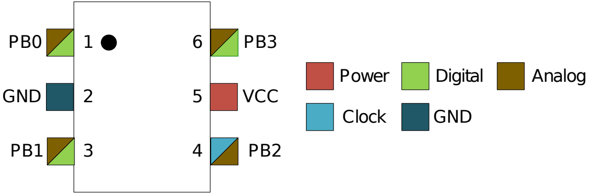
Parameter	Conditions	AD620A Min	AD620A Typ	AD620A Max
Common-Mode Rejection				
1 kΩ Source Imbalance	V_{CM} = 0 V to ±10 V			
G = 1		73	90	
OUTPUT				
Output Swing	R_L = 10 kΩ			
	V_S = ±2.3 V to ±5 V	$-V_S$ + 1.1		$+V_S$ − 1.2
DYNAMIC RESPONSE				
Small Signal −3 dB Bandwidth				
G = 1			1000	

(c) Breaking cell boundaries

Future Work: Implicit Relationships

CLASSIFICATION		A	B
h_{FE}	BC856	125~250	220~475
	BC857	125~250	220~475
	BC858	125~250	220~475

(a) Relationships to specific parts are implied by column headers alone. This example table is specifying that BC856A, BC857A, and BC858A have an h_{FE} of 125~250, while those with a B suffix have a value of 220~475.



(b) Relationships may also be specified using color matching.

Future Work: Open Information Extraction

Our approach extracts precise, pre-defined relations from documents.

- Requires explicitly defined schemas, each with corresponding labeling functions.
- Scales linearly with the number of target relations.

Can we utilize techniques in *open information extraction* to extract large sets of relations without requiring pre-defined specifications?