A Text Mining Framework to Map BMS data to BEM

Zhan Sicheng, Adrian Chong

Department of Building, National University of Singapore

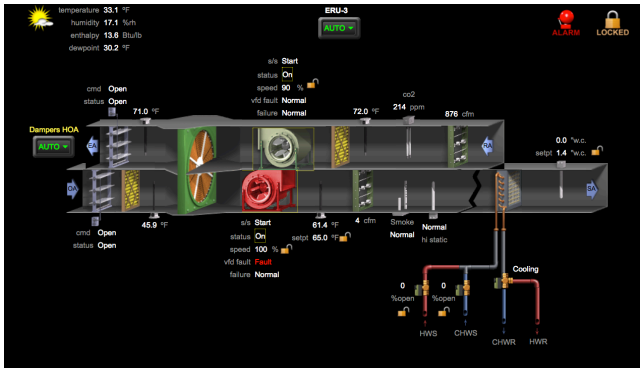
1 BACKGROUND

2 METHODOLOTY

3 RESULT & ANALYSIS

4 DISCUSSION

BACKGROUND



Building Management System (BMS)

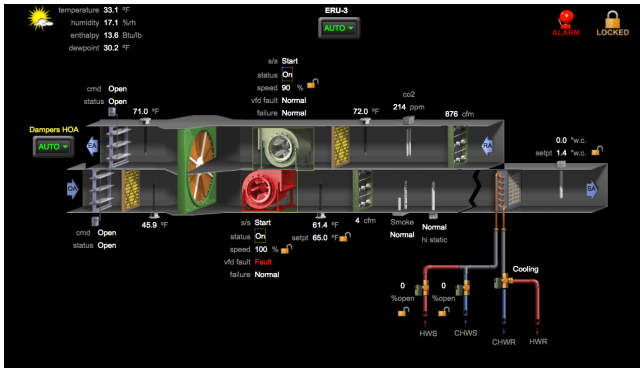
+

Building Energy Modeling (BEM)

Calibration

- Retrofit analysis
- Measurement and verification
- Fault detection and diagnosis
- Model predictive control
- …

BACKGROUND



Building Management System (BMS)

?

HOWEVER…

Building Energy Modeling (BEM)

Sample of raw BMS data tag:

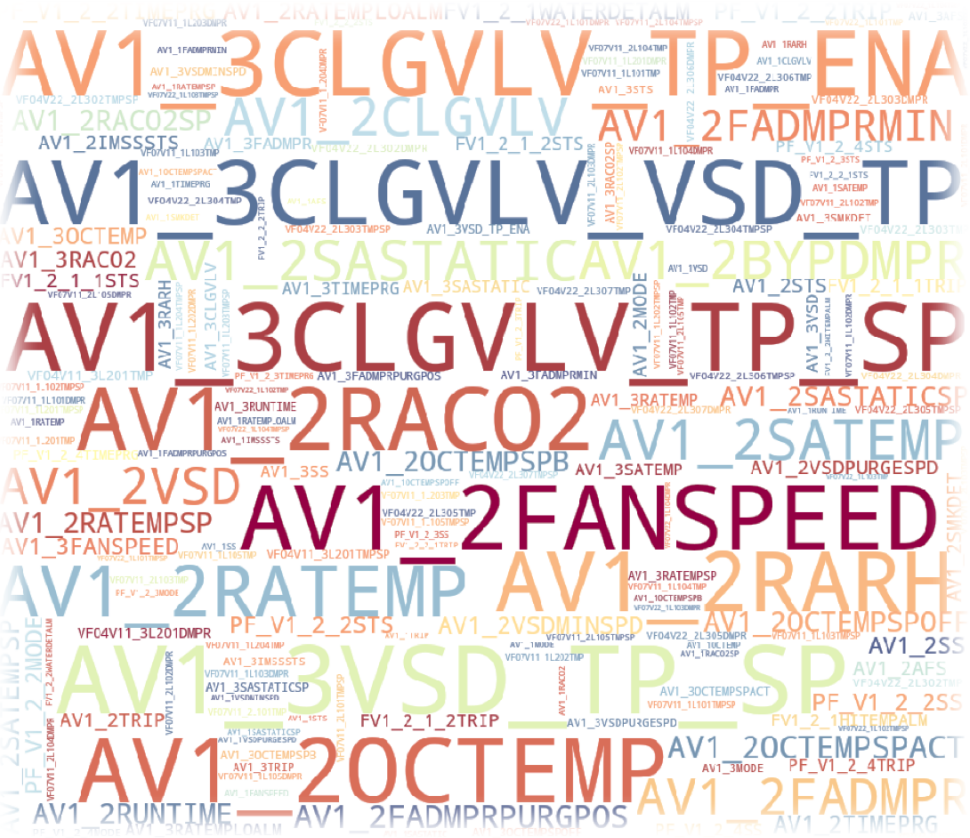
SN	OPC Tag
1	/FACILITY/VENTUS/VENTUS_BLK1/AV1_3CLGVLV_TP_SP.POINTVALUE
2	/FACILITY/VENTUS/VENTUS_BLK1/AV1_3VSD_TP_SP.POINTVALUE
3	/FACILITY/VENTUS/VENTUS_BLK1/AV1_3CLGVLV_TP_ENA.POINTVALUE
4	/FACILITY/VENTUS/VENTUS_BLK1/AV1_3CLGVLV_VSD_TP.POINTVALUE
5	/FACILITY/VENTUS/VENTUS_BLK1/AV1_2FANSPEED.POINTVALUE
6	/FACILITY/VENTUS/VENTUS_BLK1/AV1_2OCTEMP.POINTVALUE
7	/FACILITY/VENTUS/VENTUS_BLK1/AV1_2RACO2.POINTVALUE
8	/FACILITY/VENTUS/VENTUS_BLK1/AV1_2RARH.POINTVALUE
9	/FACILITY/VENTUS/VENTUS_BLK1/AV1_2RATEMP.POINTVALUE
10	/FACILITY/VENTUS/VENTUS_BLK1/AV1_2SASTATIC.POINTVALUE
11	/FACILITY/VENTUS/VENTUS_BLK1/AV1_2SATEMP.POINTVALUE
12	/FACILITY/VENTUS/VENTUS_BLK1/AV1_2BYPDMPR.POINTVALUE
13	/FACILITY/VENTUS/VENTUS_BLK1/AV1_2CLGVLV.POINTVALUE
14	/FACILITY/VENTUS/VENTUS_BLK1/AV1_2FADMPR.POINTVALUE
15	/FACILITY/VENTUS/VENTUS_BLK1/AV1_2VSD.POINTVALUE
16	/FACILITY/VENTUS/VENTUS_BLK1/AV1_2FADMPRMIN.POINTVALUE
17	/FACILITY/VENTUS/VENTUS_BLK1/AV1_2FADMPRPURGPOS.POINTVALUE
18	/FACILITY/VENTUS/VENTUS_BLK1/AV1_2OCTEMPSPACT.POINTVALUE
19	/FACILITY/VENTUS/VENTUS_BLK1/AV1_2OCTEMPSPB.POINTVALUE
20	/FACILITY/VENTUS/VENTUS_BLK1/AV1_2OCTEMPSPOFF.POINTVALUE
21	/FACILITY/VENTUS/VENTUS_BLK1/AV1_2RACO2SP.POINTVALUE
22	/FACILITY/VENTUS/VENTUS_BLK1/AV1_2RATEMPSP.POINTVALUE
23	/FACILITY/VENTUS/VENTUS_BLK1/AV1_2RUNTIME.POINTVALUE
24	/FACILITY/VENTUS/VENTUS_BLK1/AV1_2SASTATICSP.POINTVALUE
25	/FACILITY/VENTUS/VENTUS_BLK1/AV1_2SATEMPSP.POINTVALUE
26	/FACILITY/VENTUS/VENTUS_BLK1/AV1_2VSDMINSPD.POINTVALUE
27	/FACILITY/VENTUS/VENTUS_BLK1/FV1_2_1WATERDETALM.POINTVALUE
28	/FACILITY/VENTUS/VENTUS_BLK1/PF_V1_2_2MODE.POINTVALUE
29	/FACILITY/VENTUS/VENTUS_BLK1/PF_V1_2_2STS.POINTVALUE
30	/FACILITY/VENTUS/VENTUS_BLK1/PF_V1_2_2TRIP.POINTVALUE

⋮

BACKGROUND

/FACILITY/VENTUS/VENTUS_BLK1/VF07V11_2L102TMP.POINTVALUE

Network interface	Location	Equipment	Point type



- Hundreds and thousands of data points in each database
- Randomly customized naming rules (e.g. chilled water supply temperature: ‘SWT’, ‘Supply Temp’, ‘BMFWDTMP’)
- Very troublesome and error-prone manual mapping process
- An **AUTOMATED SOLUTION** is required!

BACKGROUND

Bioinformatics & NLP

- Methods for Gene Name Entity Recognition (e.g. morphological rules [1], dictionary based inexact string matching [2], SVM based classification [3], etc.)

x Requires intensive contextual information

BIM application

- IFC ↔ other schemas (TF.IDF [4], instance based rule [5], etc.)

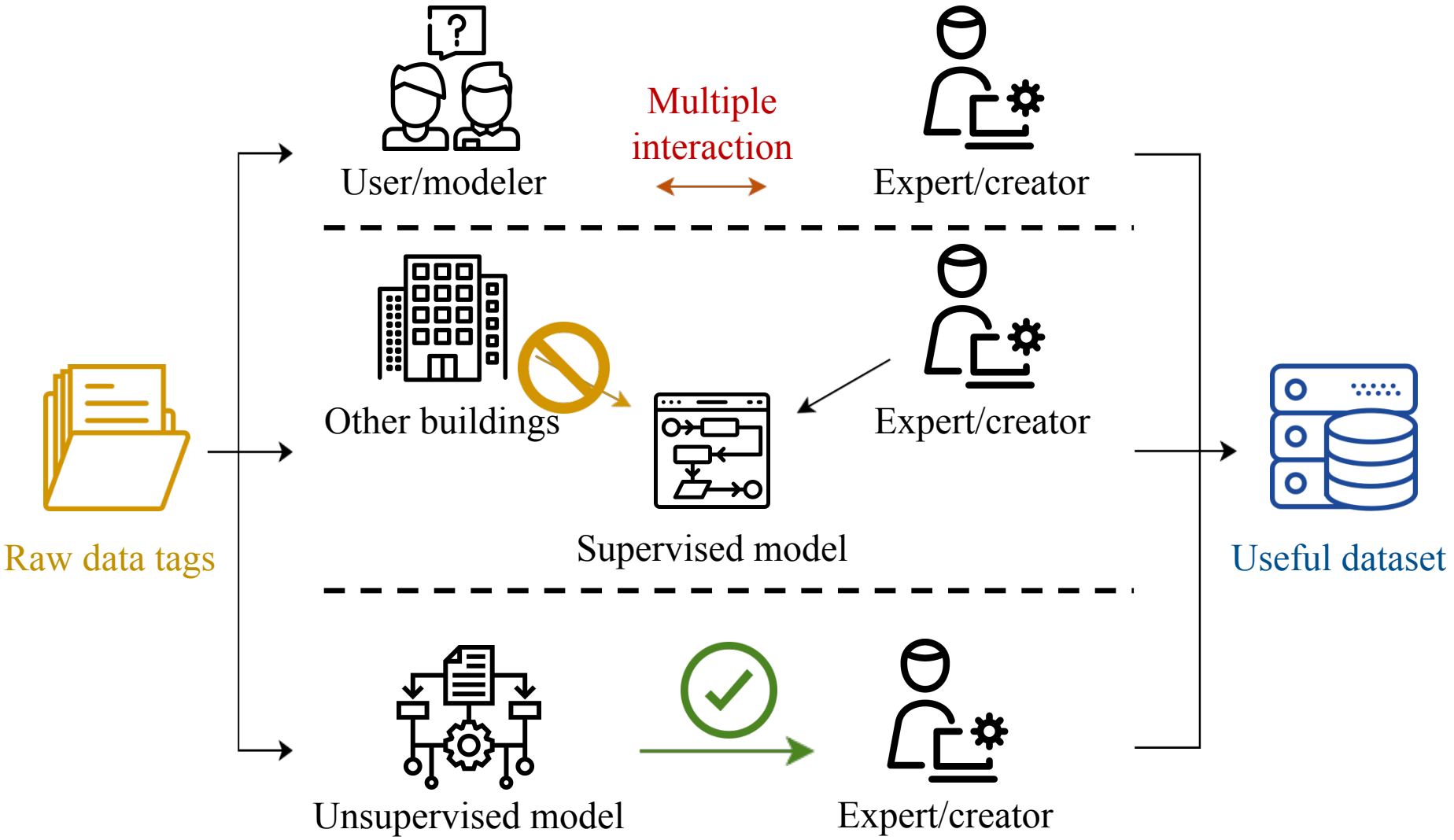
x Mainly focused on geometry properties; between standard schemas

Metadata normalization

- Multi-layer perceptron, conditional random fields model to understand BMS tags[6]

x Supervised learning is not generalizable for non-similar buildings

BACKGROUND


User/modeler
Multiple interaction
Expert/creator
Other buildings
Expert/creator
Raw data tags
Supervised model
Useful dataset
Unsupervised model
Expert/creator

METHODOLOGY

Pre-processing

- Trimming
- Vectorization

Unsupervised Learning

- DBSCAN
- Representative string extraction

Fuzzy String Matching

- X-gram segmentation
- Matching E+ dictionary

Test dataset: BMS point names of a office building in Singapore
Explanatory sample: " *RMTEMP* " (room air temperature)

METHODOLOGY

Pre-processing

raw tag S_K: "/FACILITY/VENTUS/VENTUS_BLK3/FV3_2_1RMTEMP.POINTVALUE"

Trimming with delimiter

core segment of S_K: "FV3_2_1RMTEMP"

Calculate Levenshtein distance

feature vector of S_K: [5.0 3.0 13.25 ... 0.0 8.75 ... 14.5]

distance to S_1 | distance to S_2 | ... | distance to S_K | ... | distance to S_N

Levenshtein distance (edit distance): minimum number of character changes needed to alter a string into another one, e.g. D('BIM','BEM') = 1, D(itself) = 0

METHODOLOGY

Unsupervised Learning

Important observation: in large scale buildings, there are multiple points with the same measurement type, following similar naming rules.

Sensor type	Count	Sensor type	Count
Room temperature	78	Return air temperature	18
Damper position	78	Off coil temperature	18
Room temperature setpoint	78	Chiller water temperature	18
Power meter KWHR	26	Supply air temperature	6
Power meter KW	26	Supply air CO2 level	6
Control status	18	High temperature alarm	2
…	…	…	…

METHODOLOGY

Unsupervised Learning

Color stands for sensor type

METHODOLOGY

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Clustering result in 2D dimension

METHODOLOGY

Fuzzy String Matching

Representative substring

"RMTEMP"

Generate X-gram

X-gram lists

[["R","M","T","E","M","P"]
["RM","T","E","M","P"]
["RMT","E","M","P"]
...
["RM","TEMP"]
...
["RMTEMP"]]

EnergyPlus based dictionary

[{"RM","ZONE","ROOM","ZN","SPC"},
{"AIR","A"},
{"TEMP","TMP","T"}]

$$R_{matched} = \frac{2}{\frac{L_{total}}{L_{matched}} + \frac{N_{total}}{N_{matched}}}$$

["R","M","T","E","M","P"]: $R_{match} = 0.22$
["RM","T","E","M","P"]: $R_{match} = 0.5$
["RMT","E","M","P"]: $R_{match} = 0$
...
["RM","TEMP"]: $R_{match} = 0.8$
...
["RMTEMP"]]: $R_{match} = 0$

Proposed matches:

#1,"RMTEMP", $R_{match} = 0.8$

#2,"SATEMP", $R_{match} = 0.74$

#3, "RATEMP", $R_{match} = 0.74$

#4, "RMTEMPSP", $R_{match} = 0.7$

#5, "ATEMPSP", $R_{match} = 0.69$

#6, "RMTEMPSPB", $R_{match} = 0.67$

#7, "RATEMPSP", $R_{match} = 0.64$

RESULT & ANALYSIS

Unsupervised Learning

Clustering result

-1: ['AV1_2BYPDMPR', 'AV3_1OCTEMPOFF',
'FAF_V3_2_2MODE', 'FAF_V3_2_2STS', 'FAF_V3_2_2TRIP',
'VENTUSBMS_RMHITEMP', 'VF11_ZM1_OFFSETSP'],

0: ['AV1_3CLGVLV_TP_SP', 'AV1_1CLGVLV_TP_SP', ...
'AV2_1CLGVLV_TP_SP', 'AV3_1CLGVLV_TP_SP'],

1: ['AV1_3VSD_TP_SP', 'AV1_1VSD_TP_SP', ...
'AV2_1VSD_TP_SP', 'AV3_1VSD_TP_SP'],

...

44: ['VF04V11_3L201TMP', 'VF04V22_2L302TMP', ...
'VF11V33_1L303TMPSP', 'VF11V33_1L305TMPSP']
...

107: ['VL3FCU301BMDIFFTMP', 'VL3FCU302BMDIFFTMP']

Extract representative substring

Representative substrings

0: 'CLGVLV_TP_SP'

1: 'VSD_TP_SP'

2: 'CLGVLV_TP_ENA'

3: 'CLGVLV_VSD_TP'

4: 'FANSPEED'

5: 'OCTEMP'

6: 'RACO2'

...

61: 'BMFWDTMP'

F-measure = 0.872 (recall = 0.869, precision = 0.876)

RESULT & ANALYSIS

Fuzzy String Matching

	Proposed X-gram	Baseline N-gram
Representative list of "RMTEMP"	[["R","M","T","E","M","P"], ["RM","T","E","M","P"], ["RMT","E","M","P"], ... ["RM","TEMP"] ... ["RMTEMP"]]	["R","M","T","E","M","P", "RM","MT","TE","EM","MP", "RMT","MTE","TEM","EMP", "RMTE","MTEM","TEMP", "RMTEM","MTEMP", "RMTEMP"]
Proposed match of "zone mean air temperature"	**#1,"RMTEMP", R_{match} = 0.8;** #2,"SATEMP", R_{match} = 0.74; #3, "RATEMP", R_{match} = 0.74; #4, "RMTEMPSP", R_{match} = 0.7; #5, "ATEMPSP", R_{match} = 0.69; #6, "RMTEMPSPB", R_{match} = 0.67; #7, "RATEMPSP", R_{match} = 0.64;	#1,"TMP", $S_{jaccard}$ = 0.22; #2,"STS", $S_{jaccard}$ = 0.22; #3, "AFS", $S_{jaccard}$ = 0.22; #4, "SATEMP", $S_{jaccard}$ = 0.17; **#5, "RMTEMP", $S_{jaccard}$ = 0.17;** #6, "RATEMP", $S_{jaccard}$ = 0.17; #7, "TRIP", $S_{jaccard}$ = 0.15;
Overall accuracy	**0.869**	**0.587**

DISCUSSION

- ✓ The framework successfully eliminate the human effort required to apply BMS data for BEM relevant application
- ✓ The framework can be applied to various other scenarios such as BIM by adjusting the dictionary
- ✓ More information such as sensor location and sensor data can be used as input to achieve mapping at higher LoD (Level of Detail)
- ✓ Ultimately, to fully exploit the information contained in BMS and to help understand the building energy performance

REFERENCE

[1] Gaizauskas R., et al. 2003. Protein structures and information extraction from biological texts: the PASTA system. Bioinformatics. 19(1), 135-143.

[2] Altschul S. F., et al. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res.Vol. 25(17), pp. 3389–3402.

[3] Chang J. T., Schütze H., & Altman R. B. 2004. GAPSCORE: finding gene and protein names one word at a time. Bioinformatics. 20(2), 216-225.

[4] Cheng J. C., Deng Y., & Anumba C. 2015. Mapping BIM schema and 3D GIS schema semi-automatically utilizing linguistic and text mining techniques. Journal of Information Technology in Construction (ITcon). 20(14), 193-212.

[5] Deng Y., Cheng J. C., & Anumba C. 2016. Mapping between BIM and 3D GIS in different levels of detail using schema mediation and instance comparison. Automation in Construction. 67, 1-21.

[6] Koh, J., et al. 2018. Scrabble: transferrable semi-automated semantic metadata normalization using intermediate representation. In *Proceedings of the 5th Conference on Systems for Built Environments* (pp. 11-20). ACM.

NUS National University of Singapore | A2018 sim

THANK YOU!