University of Southern California

USC Viterbi School of Engineering

Minimizing User Effort in Transforming Data by Example

Bo Wu, Pedro Szekely and Craig A. Knoblock

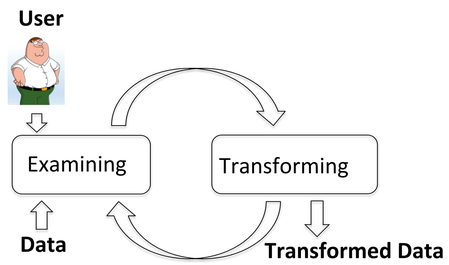
USC INFORMATION SCIENCES INSTITUTE

Previous Work

Input	Output
Brooker, Alexander	Alexander Brooker
T.C. Steele	T.C. Steele
Paxton, William Mcgregor	
…	
Hugh M. Poe	

Examples

```
if classify(input) == "class1":
    pos1 = (","BNK, UWRD)
    pos2 = (LWRD, END)
    pos3 = (",", BNK)
    pos4 = (BNK, UWRD)
    pos5 = (START, UWRD)
    pos6 = (UWRD ",")
    output = substring(input, pos1, pos2)+ substring(pos3, pos4)+substring(pos5, pos6)
elif classify(input) == "class2":
    pos1 = (START, UWRD)
    pos2 = (LWRD, END)
    output = substring(input, pos1, pos2)
...
```



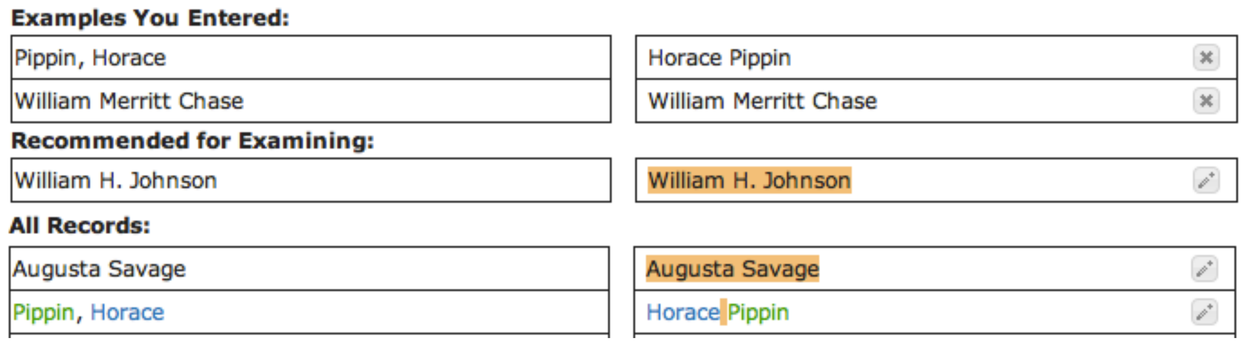
Problem

Input	Output
Brooker, Alexander	Alexander Brooker
T.C. Steele	T.C. Steele
Paxton, William Mcgregor	William Mcgregor Paxton
1000 rows	
Wagenhals, Katherine H.	Wagenhals, Katherine H.

User should select a row from a large amount of rows to provide an example

System available at http://www.isi.edu/integration/karma/

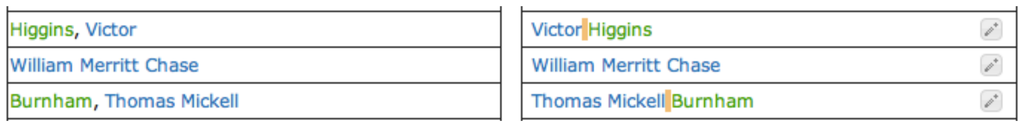
Our Approach



- Recommending Input
  - Recommend informative record
    - Sort record based on number of failed position expressions

    Ex: forsyth, william j. ➔ failed execution
  - Recommend questionable record (if no failed position expression)
    - Convert the records into feature vectors and sort the records based on their distances to examples

    Ex: Morris, George Lovett Kingsland ➔George Lovett Kingsland Morris
- Color-Coding Transformation



User Study Results

Scen	Without Extensions (group 1)		With Extensions (group 2)	
	Avg time (sec.)	Success rate	Avg time (sec.)	Success rate
s1	30	1.0	35	1.0
s2	119	0.6	41	1.0
s3	110	0.6	40	1.0
s4	unsolved	0.0	unsolved	0.0
s5	201	0.2	95	1.0
s6	unsolved	0.0	142	1.0
s7	unsolved	0.0	unsolved	0.0
s8	191	0.4	95	1.0
All	130.2	0.35	74.6	0.75

Future Work

- Provide users better understanding of the result
  - Visualize partitions
  - Show percentage of failed inputs
  - Show histogram of result
  - …