Learning Data Transformation Rules through Examples: Preliminary Results

Bo Wu, Pedro Szekely, Craig A.Knoblock
Information Science Institute
University of Southern California

Transforming Data

Original	Transformed
30/07/2010	2010-07-30
30/09/2010	2010-09-30
14/01/2011	2011-01-14

Transforming Data

Original	Transformed
1 Lombard Street,London	London,1 Lombard Street
1 Dominick Street,New York	New York, 1 Dominick Street
1 North Belmont Avenue,Richmond	Richmond, 1 North Belmont Avenue

Transforming Data by Example

Example

1 Lombard Street,London	**London,1 Lombard Street**

Original
1 Lombard Street,London
1 Dominick Street,New York
1 North Belmont Avenue,Richmond

Transformed
London, 1 Lombard Street
New York,1 Dominick Street
Richmond,1 North Belmont Avenue

Examples Are Ambiguous

Example

1 Lombard Street,London	**London,1 Lombard Street**

Original
1 Lombard Street,London
1 Dominick Street,New York
1 North Belmont Avenue,Richmond

Result 1
London ,1 Lombard Street
New,1 Dominick Street York
Richmond ,1 North Belmont Avenue

Result 2
London ,1 Lombard Street
New,1 Dominick Street York
, Avenue1 North Belmont Richmond

522 interpretations given this example

Objective

Minimize number of examples users have to give to produce the desired transformation program

Outline

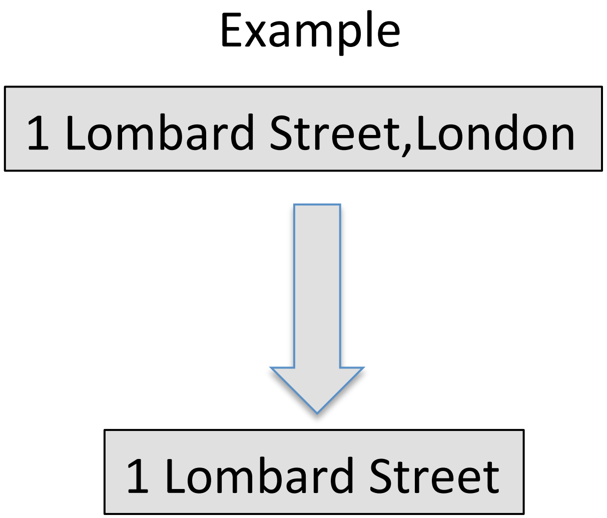
- Transformation Grammar
- System Overview
- Search spaces
- Searching
- Ranking
- Evaluation

Transformation Grammar

- program➔(ins|del|mov)+
- del➔DEL what ⋁ DEL range
- ins➔INS(token)+ where
- mov➔ MOV tokenspec where ⋁MOV range where
- what➔quantifier tokenspec
- quantifier ➔ANYNUM⋁NUM
- tokenspec➔singletokenspec⋁singletokenspec tokenspec
- singletokenspec➔token⋁type⋁ANYTOK
- type➔NUMTYP⋁WRDTYP⋁SYBTYP⋁BNKTYP
- range ➔start end
- scanningOrder➔FRM_BEG⋁FRM_END
- start➔scanningOrder posquantifier
- end➔scanningOrder posquantifier
- where➔scanningOrder posquantifier
- where➔scanningOrder posquantifier
- posquantifier➔INCLD? tokenspec⋁NUM

Transformation Grammar

- Specifying the target pattern(tokenspec)
  - any two tokens
  - “,”London
  - symbol word
  - “,” word
  - …
- Specifying the position(range)
  - [5,6]
  - after “,” before END
  - after 5, before END
  - …



Example

1 Lombard Street,London

1 Lombard Street

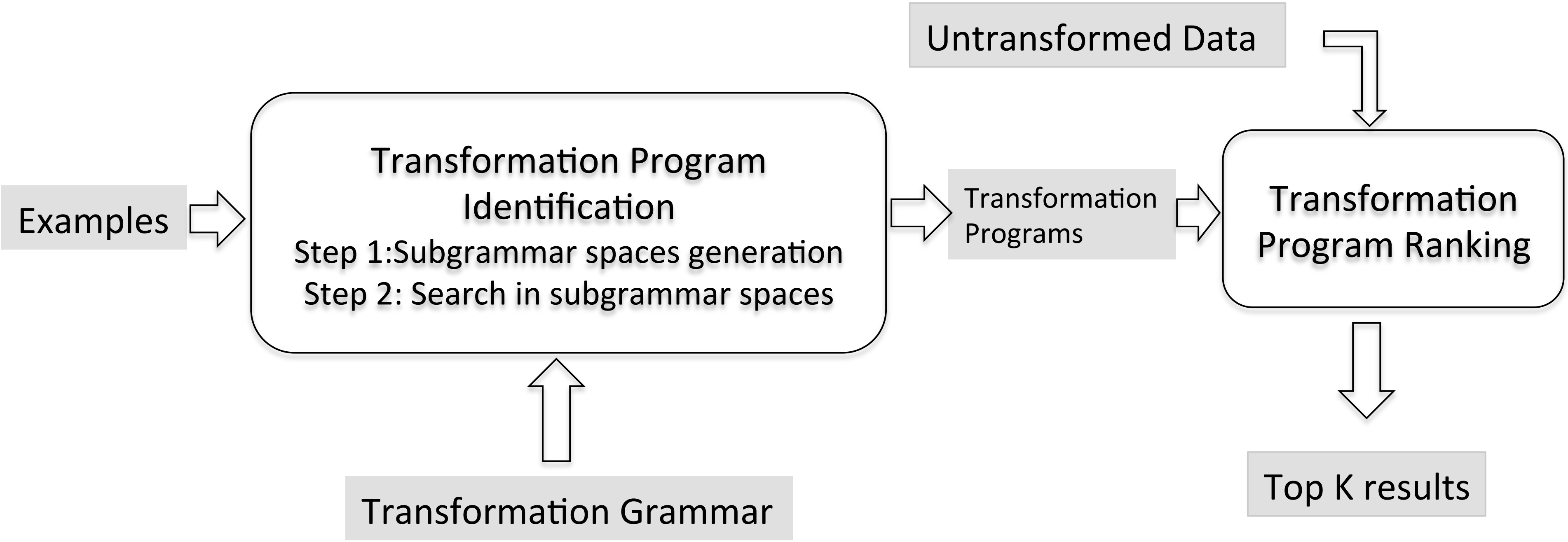
Challenges

- Large search space

$$G = (ins|mov|del)*$$

- Many interpretations

System Overview



Untransformed Data

Examples

Transformation Program Identification

Step 1:Subgrammar spaces generation

Step 2: Search in subgrammar spaces

Transformation Programs

Transformation Program Ranking

Transformation Grammar

Top K results

Subgrammar space

<START>1 Dominick Street,New York<END> || New York,1 Dominick Street

MOV	MOV
Tokenspec: • <S>1 Domininick Street • <S>NUM BNK WRD BNK WRD • ANYTOK ANYTOK ANYTOK ANYTOK ANYTOK ANYTOK • <S>NUM BNK Dominick BNK Street • … … Start: • 0 • START • NUM • …	Tokenspec: • , • SYB Start: • 0 • START • SYB

MOV	MOV
Tokenspec: • <S>1 Domininick Street • <S>NUM BNK WRD BNK WRD • ANYTOK ANYTOK ANYTOK ANYTOK ANYTOK ANYTOK • <S>NUM BNK Dominick BNK Street • … … Start: • 0 • START • NUM • …	Tokenspec: • New York<END> • WRD BNK WRD<END> • New BNK York<END> • WRD BNK York<END> • … Start: • 1 • WRD • SYB

<START>1 Dominick Street , New York<END>

<START>1 Dominick Street , New York<END>

Subgrammar space

Example 1

1 Dominick Street,New York New York,1 Dominick Street

Edit Sequences

[mov: 0,5,11[], mov: 0,0,5[]]

… …

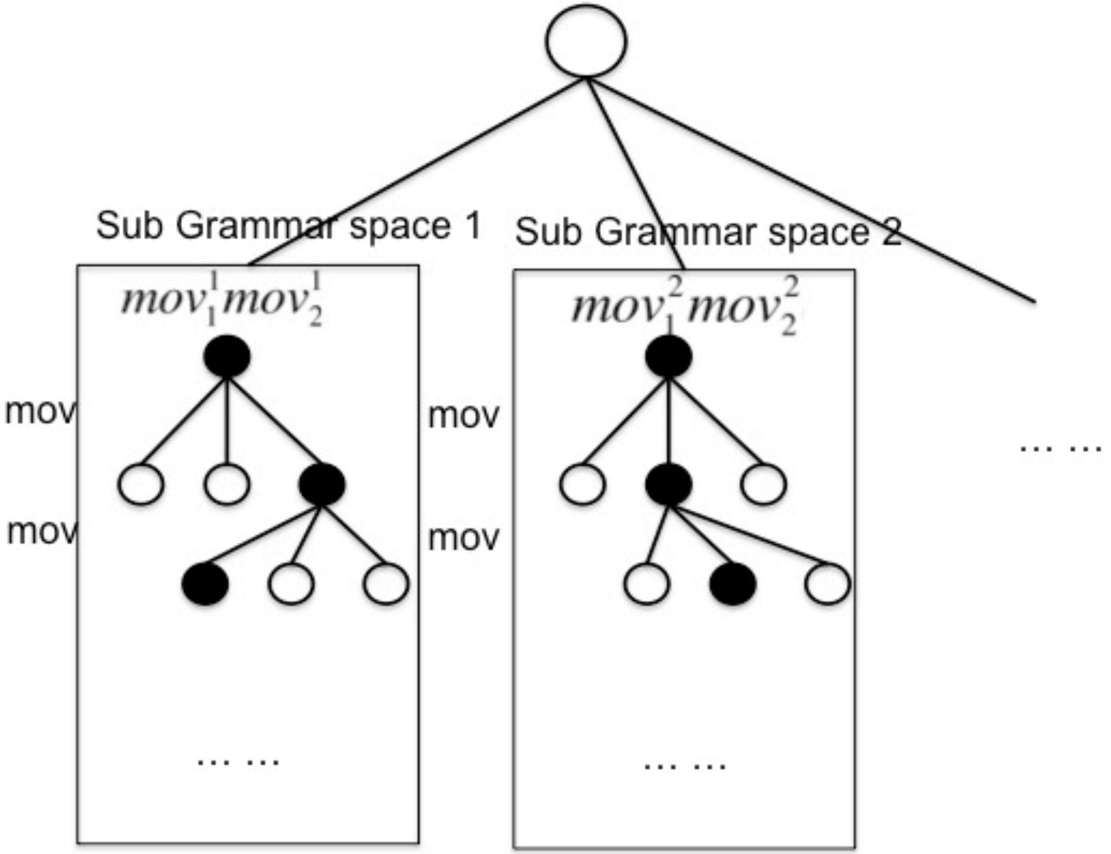
MOV	MOV
Tokenspec: • 1 Domininick Street • NUM BNK WRD BNK WRD • ANYTOK ANYTOK ANYTOK ANYTOK ANYTOK • NUM BNK Dominick BNK Street • … … Start: • 0 • START • NUM • …	Tokenspec: • , • SYB Start: • 0 • START • SYB

Search

Search Space is still large: do sampling-based search

1 Sample a subgrammar space to search

2 Do UCT (Levente Kocsis et al.) search in the sampled search space



Ranking

Result 1	/ count	Result 2	/ count
2010-07-30	0	2010-07-30	0
2010-09-30	0	/09/2010--30	2
2011-01-31	0	/03/2011--31	2

Assumption:
User wouldn't want to transform data into a noisy and irregular state

Features: capture the homogeneity

- enp_cnt_/: entropy of the distribution of the slash count
- enp_cnt_-: … …

… …

Approach:

- Build a logistic regression classifier
- Use confidence score as result's score

Evaluation

Editing Scenarios

Address 1

First row: Brankova&nbsp;13 , Brankova 13

Address2

First row: 1 Lombard Street,London , London,1 Lombard Street

Date1

First row: 2010-07-30 , 07/30/2010

Date2

First row: 13/05/2010 , 2010-05-13

Tel1

First row: Tel:</B> 020-7928 3131 , 020-7928 3131

Tel2

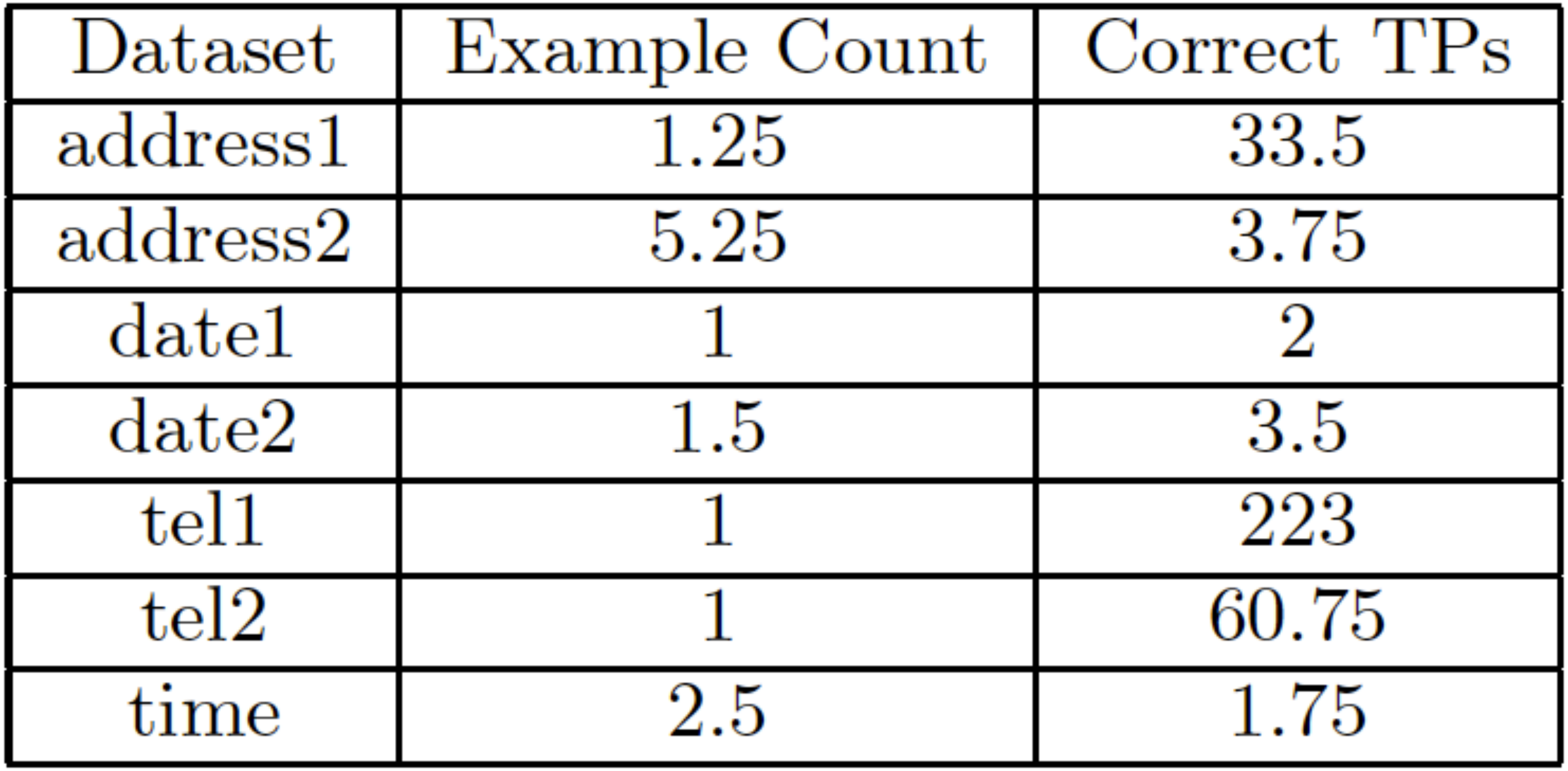
First row: 020-8944 9496 , (020)8944 9496

Time

First row:1 January 2007 4:48pm , January 1,2007 4:48pm

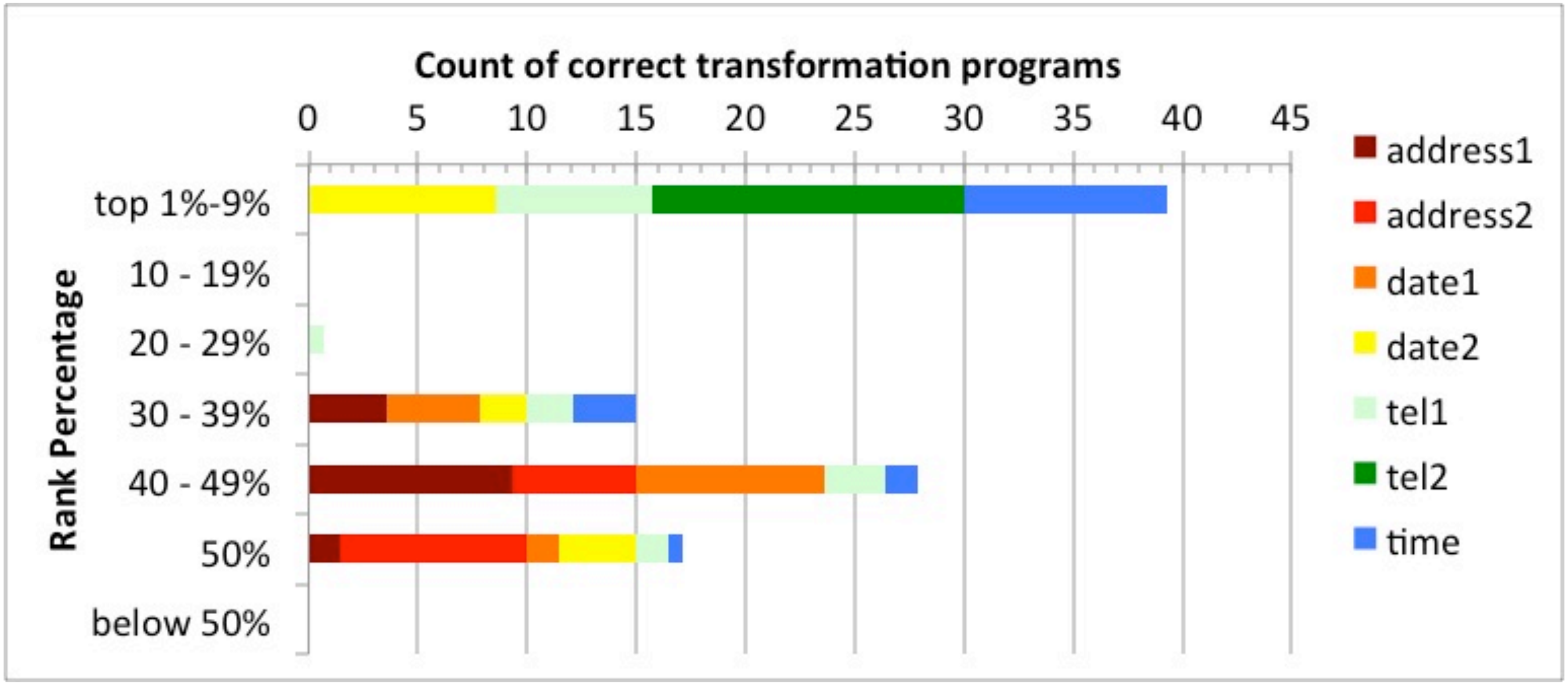
Results

Run experiment 20 times and average the result.



Dataset	Example Count	Correct TPs
address1	1.25	33.5
address2	5.25	3.75
date1	1	2
date2	1.5	3.5
tel1	1	223
tel2	1	60.75
time	2.5	1.75

Results


Count of correct transformation programs
0
5
10
15
20
25
30
35
40
45
Rank Percentage
top 1%-9%
10 - 19%
20 - 29%
30 - 39%
40 - 49%
50%
below 50%
address1
address2
date1
date2
tel1
tel2
time

- Thank You !