Westat

Improving Lives Through Research®

Using Natural Language Processing to Help Develop a Frame of Energy Suppliers

Meghan Martin, Cindy Good, Michelle Amsbary (Westat),
Francisco Cifuentes (U.S. Energy Information Administration)

Agenda

Study Background

Matching Challenges

Natural Language Processing to the Rescue!

Reviewing the Output

Improving the Odds, Reducing False Positives

Assessing the Results

Further Implications and Applications

Study Background

Residential Energy Consumption Survey (RECS)

- Household Survey
  - 19,000 households
- Energy Supplier Survey (ESS)
  - Case = Household + Energy Source — 30,000 cases
  - Respondent = Energy Supplier — 3,000 suppliers

Assign each CASE to a SUPPLIER

Data Sources

Residential Energy Consumption Survey (RECS)

- Household Survey →
  - Self-administered web/paper
  - Supplier name, account number: open text fields
- Energy Supplier Survey (ESS)

↓

Reference list of energy suppliers from prior cycles

Matching Challenge

Step 1A Match supplier names from HH survey to suppliers on reference list

Reference List

- WASHINGTON GAS

Write-in Responses

- Washington Gas
- Washington Gas Light
- Washington Gas Light Company
- WGL
- Washington
- Wash Gas
- DC Gas
- …

Natural Language Processing to the Rescue!

- Search for variations on supplier names
- Python script
- Compare HH-provided supplier name against reference list
- Calculate Levenshtein distance between input text and reference list candidates
  - Value between 0 and 1
  - 0 = identical

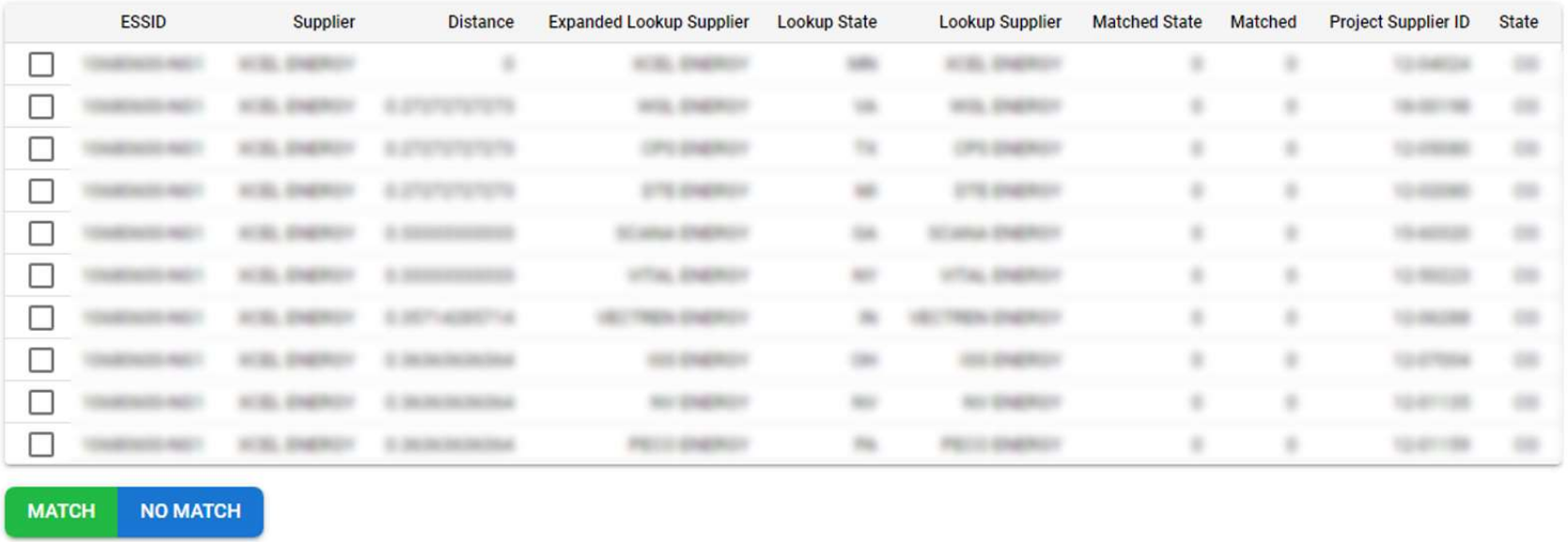
the number of single-character edits – including insertions, deletions, and substitutions – to transform the input by the respondent into a given candidate on the reference list

Natural Language Processing to the Rescue!

- Set threshold for *likely* matches
  - Score between 0.0 and 0.2: likely match
  - Score between 0.2 and 1.0: no likely match

- Set output guidelines
  - If there's a likely match: output 1 (best) candidate
  - If there's no likely match: output 10 candidates with lowest distance score

Reviewing the Output



	ESSID	Supplier	Distance	Expanded Lookup Supplier	Lookup State	Lookup Supplier	Matched State	Matched	Project Supplier ID	State
☐										
☐										
☐										
☐										
☐										
☐										
☐										
☐										
☐										
☐										

MATCH NO MATCH

Improving the Odds, Reducing False Positives

› Expand the reference list

- Manually: add known aliases
- Programmatically:
  - — Expand common abbreviations (e.g., "CO" to "COMPANY")
  - — Create acronyms or other shortened names (e.g., "Washington Gas Light" to "WGL")

› Add additional rules

- Use other data elements (e.g., check HH state against reference list state)

Assessing the Results

- Even with 100% review of output, still much faster than matching manually

- Category flag results

Category flag	# Cases	% Cases (all)	% Cases (1-3)
1. Likely match – confirmed	10,213	34%	42%
2. Possible match – confirmed	4,901	16%	20%
3. No confirmed match	9,456	32%	38%
4. No supplier name given	5,419	18%	

Further Implications and Applications

- How could we improve on our results?
  - Expand the reference list
  - Refine the rules

Category flag	# Cases
1. Likely match – confirmed	10,213
2. Possible match – confirmed	4,901
3. No confirmed match	9,456
4. No supplier name given	5,419

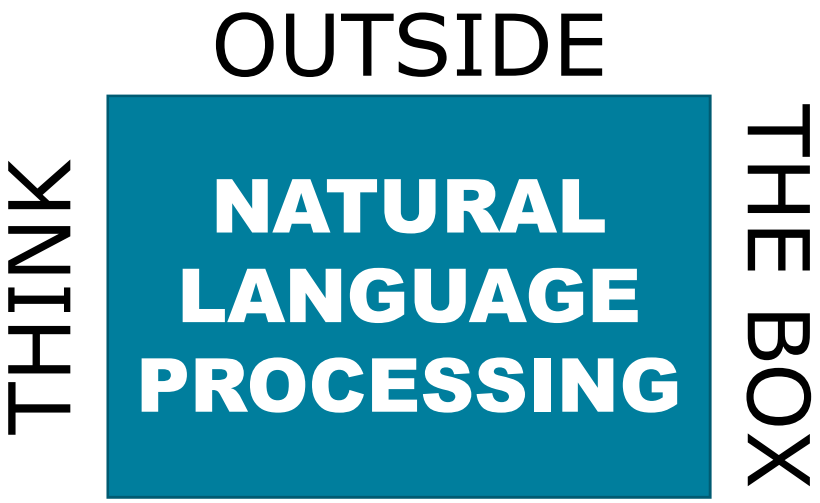
Further Implications and Applications

- Looking beyond the initial data
  - Do the results from this early stage in the study correlate to any results from the later stages? YES!

Category flag	% Disavowed
1. Likely match – confirmed	4%
2. Possible match – confirmed	5%
3. No confirmed match	8%
4. No supplier name given	22%

Category flag	% Completed
1. Likely match – confirmed	97%
2. Possible match – confirmed	95%
3. No confirmed match	92%
4. No supplier name given	84%

Further Implications and Applications



- Full sentences *or*
  Small chunks of text
- Complex analysis *or*
  Simple categorization
- End result *or*
  First step

Thank You!

MeghanMartin@westat.com

CindyGood@westat.com

MichelleAmsbary@westat.com

Francisco.Cifuentes@eia.gov