An NLP-Based Approach to Record Linkage

FCSM 2023

October 24, 2023

Lilian Huang, NORC at the University of Chicago

In collaboration with:
Brandon Sepulvado, NORC
Dean Resnick, NORC
Jennifer Taub, NORC
Brenda Betancourt, NORC

NORC at the University of Chicago

The challenges of using text data in record linkage

What is record linkage?

- The science of bringing together records from multiple datasets
- Identifying and connecting records representing the same entity
- Usually individuals

Conventional approach to text: string distance

Levenshtein distance

- Minimum number of edits

Jaro-Winkler

- Weighted prefix

teacher
peacher *(substitution)*
preacher *(insertion)*
proeacher *(insertion)*
profeacher *(insertion)*
profescher *(substitution)*
professher *(substitution)*
professoer *(substitution)*
professor *(deletion)*
professor

How can NLP be integrated?

String similarity — Conceptual similarity

Occupations

- “teacher”, “instructor”, “professor”

Alternative to compiling lists

Embeddings: an alternative representation of words

Words as numbers

Another way to conceptualize and capture similarity

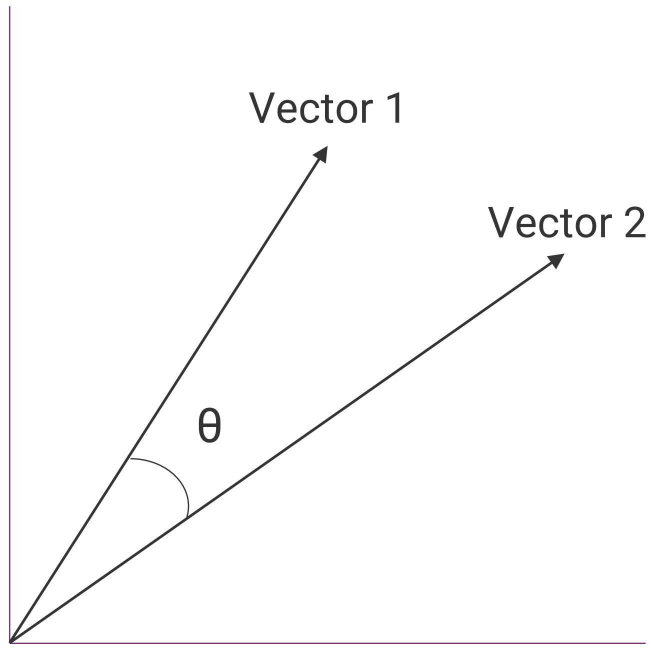
- Similar meaning = closer

```
[-4.23835516e-02,  2.33865883e-02, -3.06187198e-02,  3.74618955e-02,
 -6.81095496e-02, -9.35101211e-02,  1.15498930e-01,  7.19876215e-02,
 -4.08478007e-02, -3.51659884e-03, -5.85965328e-02, -8.57147276e-02,
 -4.34399620e-02,  1.27284229e-01,  3.09126135e-02,  2.40499619e-02,
 -8.09645001e-03, -4.56378348e-02,  7.84999877e-02,  6.91278055e-02,
  2.04219529e-03,  3.36282402e-02,  8.92838016e-02,  3.49337496e-02,
 -4.52726061e-04, -3.52332555e-02,  2.04937309e-02,  4.36704606e-03,
  5.04380092e-02,  1.12535976e-01, -2.93528736e-02, -8.07721391e-02,
 -1.66531075e-02,  2.15687007e-02, -1.94442440e-02,  2.32651979e-02,
  1.54493814e-02, -7.13979751e-02, -3.38817909e-02, -7.44279288e-03,
 -1.80203523e-02, -1.95950456e-02, -1.09561875e-01,  1.67215660e-01,
 -9.54278633e-02, -2.25519184e-02,  3.98423150e-02,  3.24025787e-02,
```

Comparing embeddings

Cosine similarity

- Angle between vectors



Different libraries for embedding production

fastText

- By Facebook

Universal Sentence Encoder

- By Google
- Context-sensitive

Sentence Transformers

- By Hugging Face
- Context-sensitive

Examples of the approach

Embeddings reflect the high similarity we intuitively expect

SIMILARITIES	“teacher” vs “professor”	“ceo” vs “founder ceo”
Levenshtein	0.111	0.273
Jaro-Winkler	0.503	0.449
fastText	0.539	0.860
Universal Sentence Encoder	0.748	0.733
Sentence Transformers	0.619	0.908

The experiment: Datasets used

Database on Ideology, Money in Politics, and Elections (DIME)

- 137,633 records from Maine

Federal Election Commission (FEC)

- 52,827 records from Maine

Variables in common
First name
Last name
Middle initial
Zip code
Employer name
Occupation title

The experiment: Three different methods

Standard (Jaro-Winkler for all; employer excluded)
First name
Last name
Middle initial
Zip code
Occupation title (string)

NLP 1 (NLP for occupation; employer excluded)
First name
Last name
Middle initial
Zip code
Occupation title (embed)

NLP 2 (NLP for occupation; Jaro-Winkler for employer)
First name
Last name
Middle initial
Zip code
Employer name
Occupation title (embed)

The experiment: Probabilistic record linkage process

NORCLink

- Proprietary record linkage software

Fellegi-Sunter model

- EM algorithm
- Overall match probability created
- Links above a threshold are accepted

The experiment: Results

NLP-enhanced models identified more links

- Probability cutoff of 0.9

Precision-recall tradeoff

- Issue of false positives

	Number of links
Standard (Jaro-Winkler for all)	48,944 unique links
NLP 1 (employer excluded)	55,696 unique links
NLP 2 (employer included)	57,535 unique links

Conclusion

Precision and recall must be balanced

- More work required
- Optimal balance depends on use case

NLP may only enhance process under specific conditions

Future enhancements

Fine-tuning embeddings

- Using our own datasets

Longer text fields

- e.g. self-identifications

Investigate correlated variables

- Affects implementation of Fellegi-Sunter method

Thank you.

Lilian Huang
Statistician
huang-lilian@norc.org

Research You Can Trust™

NORC at the University of Chicago

Questions?