Microsoft

Generating and evaluating simulated medical notes: Getting a Natural Language Generation model to give you what you want

Robert Horton, Maryam Tavakoli Hosseinabadi, Alexandre Vilcek, Wolfgang M. Pauli & Mario E. Inchiosa

Introduction

Restrictions on sharing healthcare data

Using Language Models to generate History of Present Illness (HPI)

A new software paradigm

- Software 1.0: Write a program
- Software 2.0: Curate a data set
- Software 3.0: Compose a prompt

Tuning language generation: diversity vs. consistency

Quality metrics

Experimental Setup

Wolfgang M. Pauli, PhD –
Principal AI Developer

Prompt Design

```
first_name: George
surname: Charleson
sex: male
age: 55
chief complaint: shortness of breath
dx: hypertension
rx: amlodipine
procedures: chest x-ray
---
George Charleson is a 55 y/o male who presented to urgent care complaining of shortness of
breath for the past 2 weeks. He was initially seen by his primary care physician who ordered
a chest x-ray and a stress test. His chest x-ray revealed a minor amount of congestion but
otherwise showed no abnormalities. The stress test showed a normal ECG and no angina. He was
told that he had "benign essential hypertension" but that he should try to lose weight and
exercise more. Mr. Charleson follows up with Dr. Smith every 3 months for his hypertension
and has had a good response to amlodipine and gym membership. For the past 3 months, Mr.
Charleson has developed a dry cough which has worsened over the past week.

===
first name: Amber
last name: Hand
age: 32
gender: female
marital status: married
race: black
ethnicity: nonhispanic
dx: Fracture of forearm
---
```

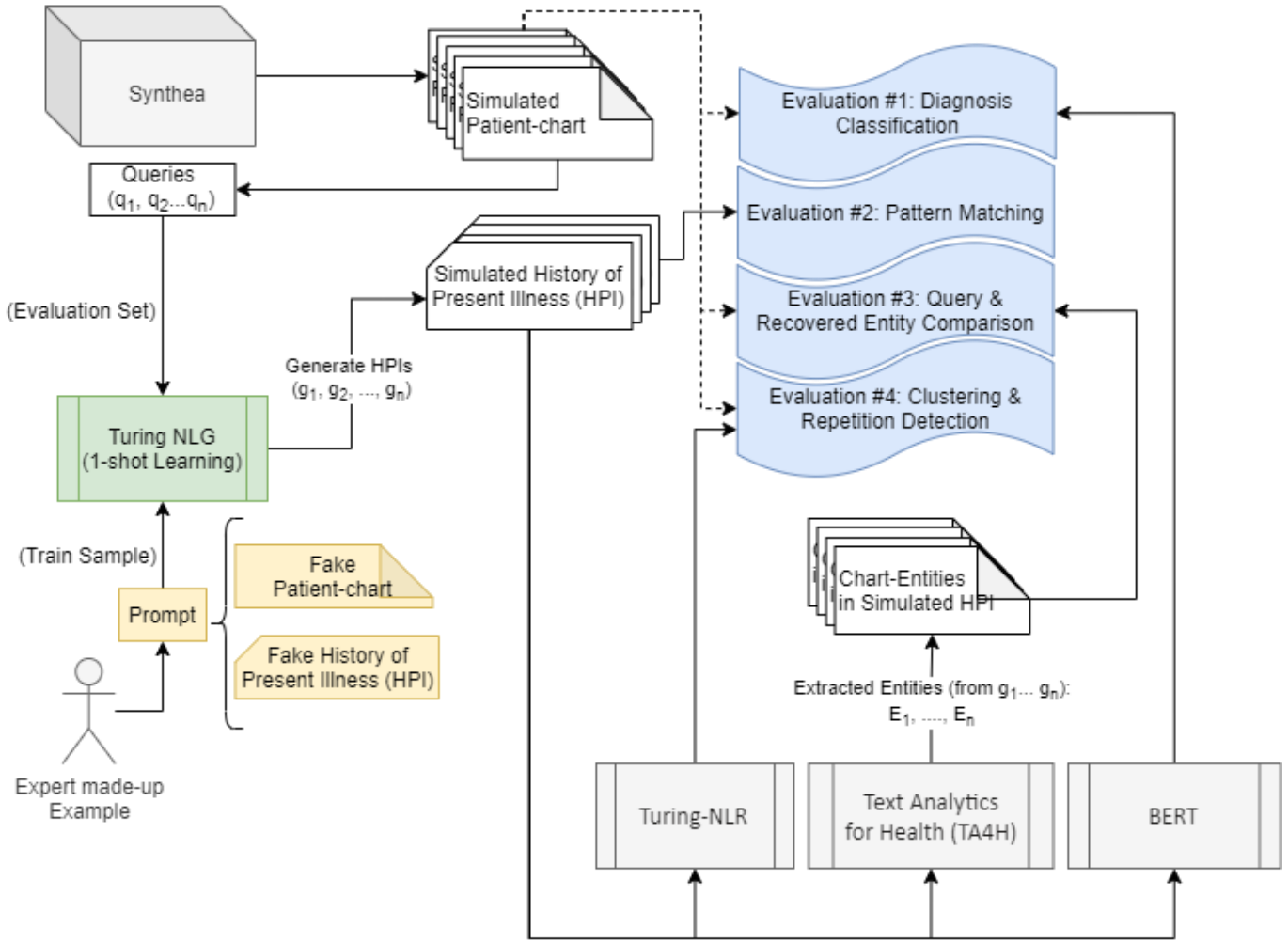
- Pair of:
  - Structured Facts
  - Example HPI passage
- Set of structured facts w/o HPI passage

Example Output

```
first name: Amber
last name: Hand
age: 32
gender: female
marital status: married
race: black
ethnicity: nonhispanic
dx: Fracture of forearm
---
```

Amber Hand is a 32 y/o female who was seen in the ED for a suspected broken arm. She reported to the ED that she was working at her job when she fell on the job. She had an acute onset of pain and swelling of her right forearm. The patient reported that she was a nursing assistant and was working in the operating room. She was given pain medication and placed in a splint. She was sent for an x-ray and her fracture was confirmed. She was then given pain medication and sent home.

Data Generation and Evaluation Flow

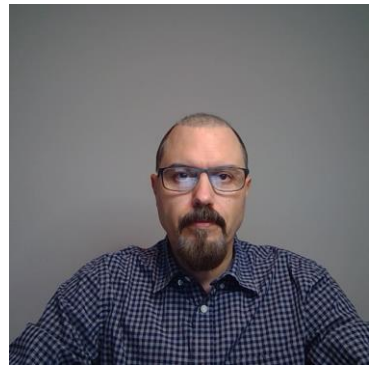


Diagnosis Classification

Alexandre Vilcek, Senior Data & Applied Scientist

Diagnosis Classification

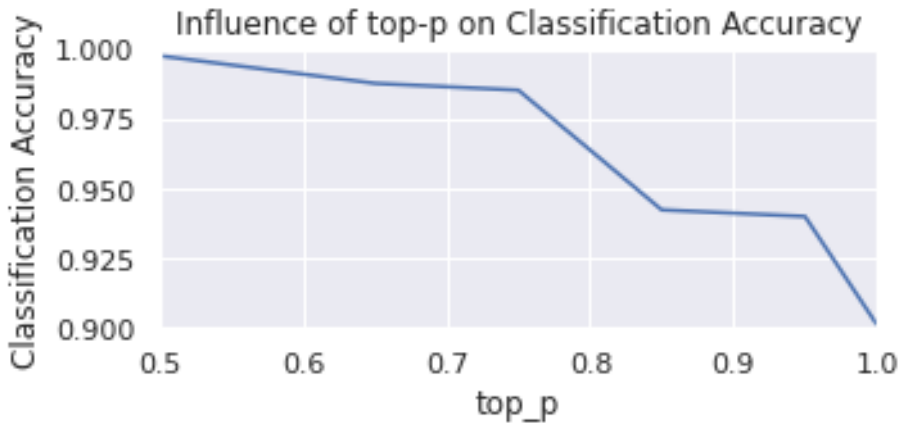
- Trying to predict the diagnosis, given by the structured facts, from the generated HPI text as a general evaluation of consistency.
- Qualitative evaluation: Classification error analysis allows us to categorize the major types of inconsistencies found.
- Quantitative evaluation: We also measured the influence of top-p on the classification accuracy.
- Experimentation setup:
  - Fine-tune a BERT model for text classification
  - Binary classification tasks: predicting diagnosis as “fracture of ankle” or “fracture of forearm”
  - Multiclass classification task: predicting diagnosis as “fracture of ankle”, “fracture of forearm”, or “sprain of ankle”
  - Training dataset comprised of 4,800 examples; evaluation dataset comprised of 1,200 examples
  - Each example has the generated narrative text as features and corresponding diagnosis as the label



Diagnosis Classification

- Three major inconsistency patterns found from the classification error analysis:
  - Facts described in generated text unrelated to any of the possible diagnosis. Here the generated text doesn't mention anything that could be used to infer the possible diagnosis.
  - Facts in generated text related to one (or more) of the possible diagnosis, but don't completely describe it. Here the generated text mentions indirect facts that would allow an expert to suspect of one of the possible diagnosis.
  - Facts in the generated text describing more than one of the possible diagnosis. Here the generated text contains facts directly related to more than one of the possible diagnosis, making all of them correct classifications.

- As expected, we noticed a drop (about 10%) in classification accuracy when allowing the text generation model to be more "creative" by increasing top-p:

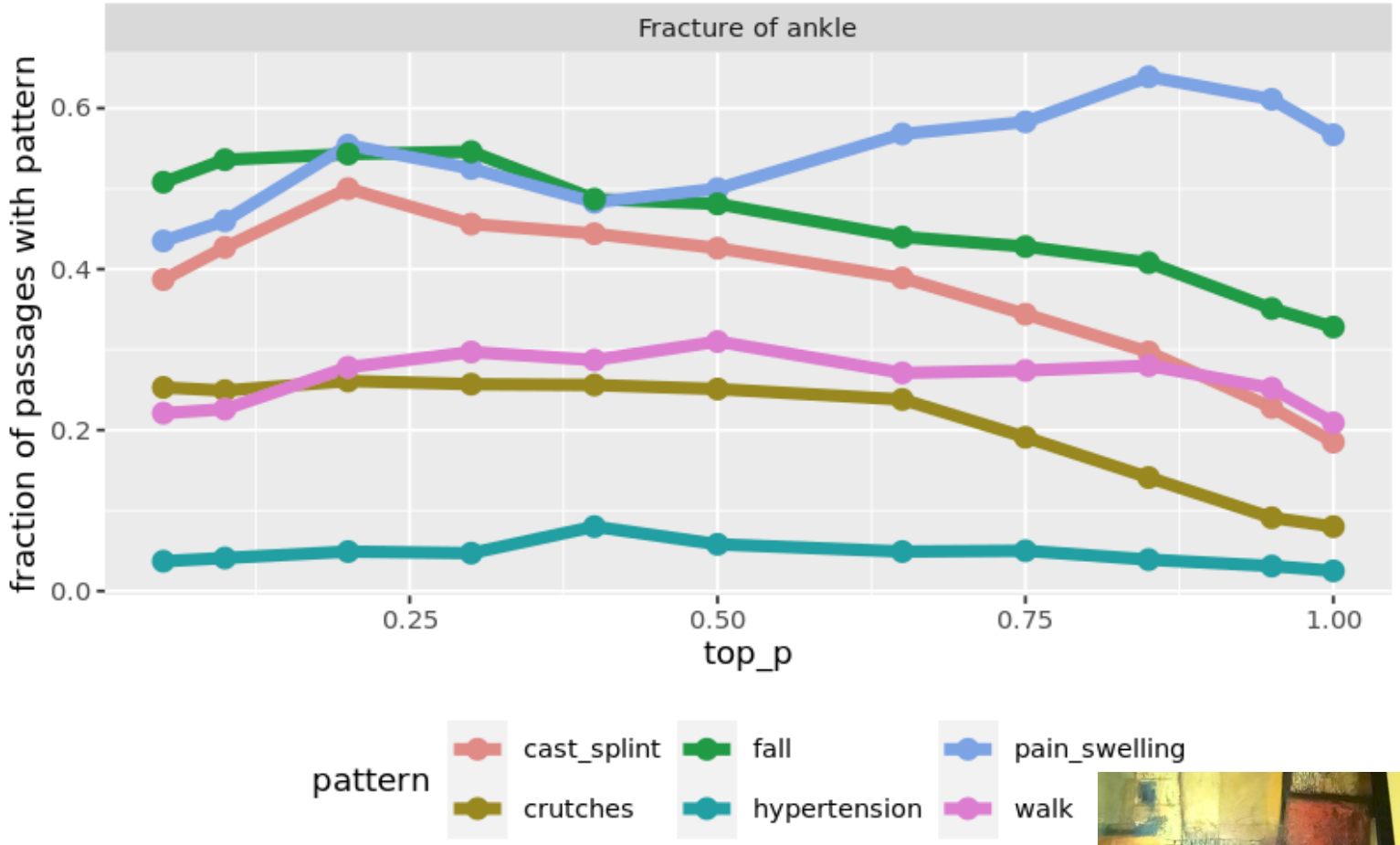


Pattern Matching & Entity Recovery

Maryam Tavakoli Hosseinabadi, PhD – Data & Applied Scientist

Pattern Matching

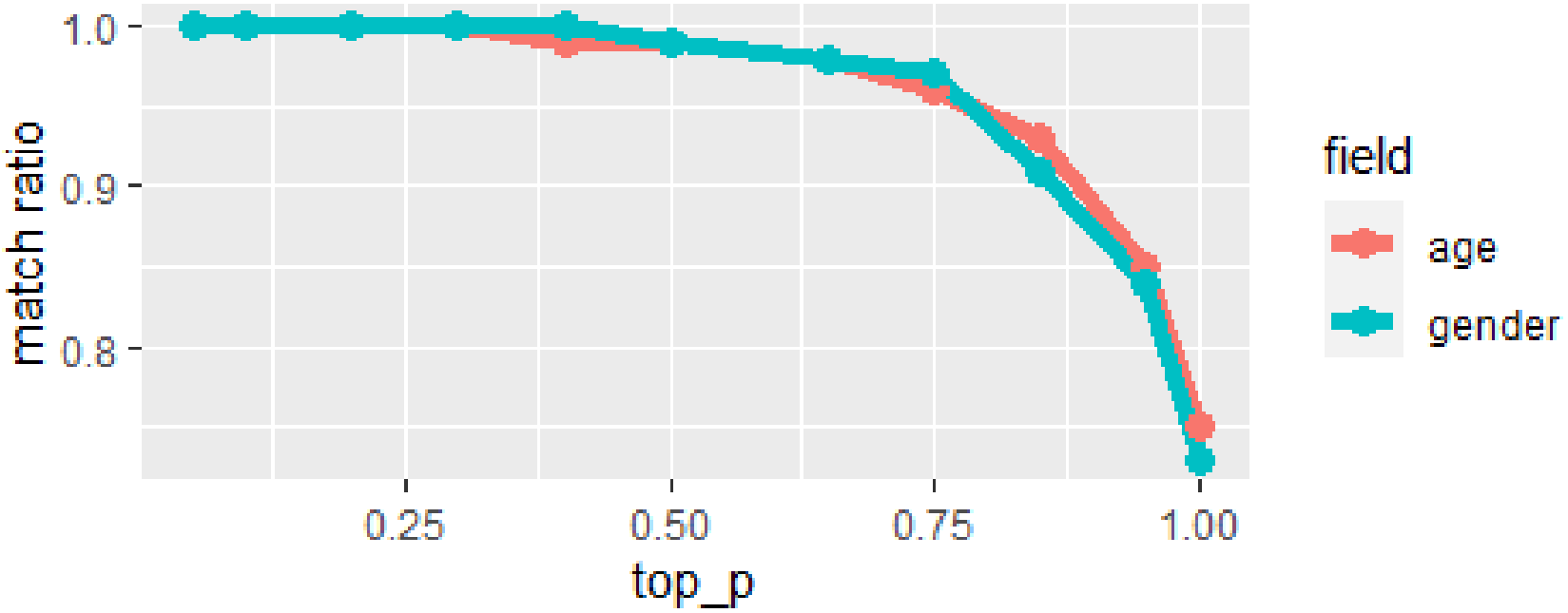
```
first_name: George
surname: Charleson
sex: male
age: 55
chief complaint: shortness of breath
dx: hypertension
rx: amlodipine
procedures: chest x-ray
---
George Charleson is a 55 y/o male who presented to urgent care com
breath for the past 2 weeks. He was initially seen by his primary
a chest x-ray and a stress test. His chest x-ray revealed a minor
otherwise showed no abnormalities. The stress test showed a normal
told that he had "benign essential hypertension" but that he shoul
exercise more. Mr. Charleson follows up with Dr. Smith every 3 mon
and has had a good response to amlodipine and gym membership. For
Charleson has developed a dry cough which has worsened over the pa

===
first name: Amber
last name: Hand
age: 32
gender: female
marital status: married
race: black
ethnicity: nonhispanic
dx: Fracture of ankle
---
```



- Regular expression & keyword Search
- [pain/swelling vs. hypertension]
- Can be extended to more complex patterns if a use-case expects (e.g., following a template)

Entity Recovery

- NER: Common practice in clinical NLP
  - *To what extent do the entities in simulated notes agree with the prompted information?*
    - NER API -- Text Analytics for Health (TA4H) Azure Cognitive Service
    - Entity extraction -- gender, age

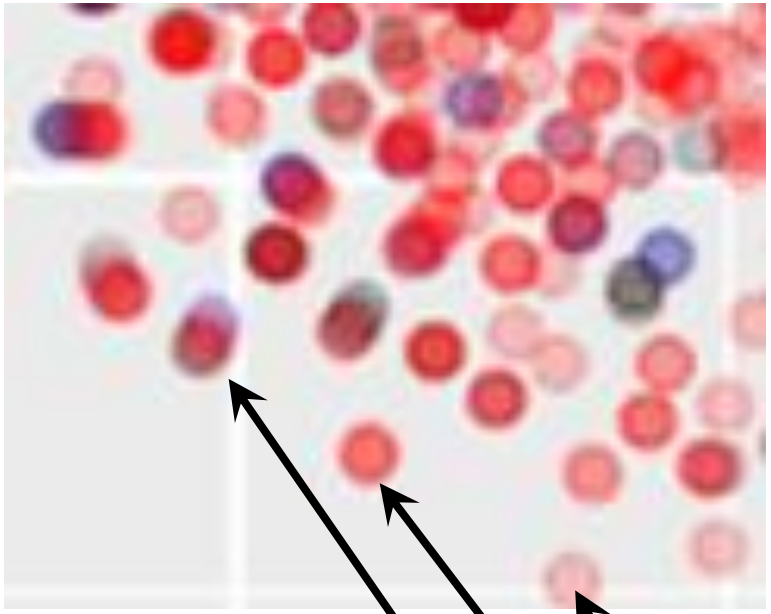


  - What are we missing with lower top_p?

Detecting Repeats

Robert Horton, PhD – Senior Data & Applied Scientist

Clusters show repetition



```
from scipy.cluster.hierarchy import ward, fcluster
from scipy.spatial.distance import pdist

X = sentence_info['vector'].tolist()
y = pdist(X, metric='cosine')
z = ward(y)

sentence_info['cluster_small'] = fcluster(z, 0.01, criterion='distance')
sentence_info['cluster_medium'] = fcluster(z, 0.1, criterion='distance')
```

Individual points are semi-transparent

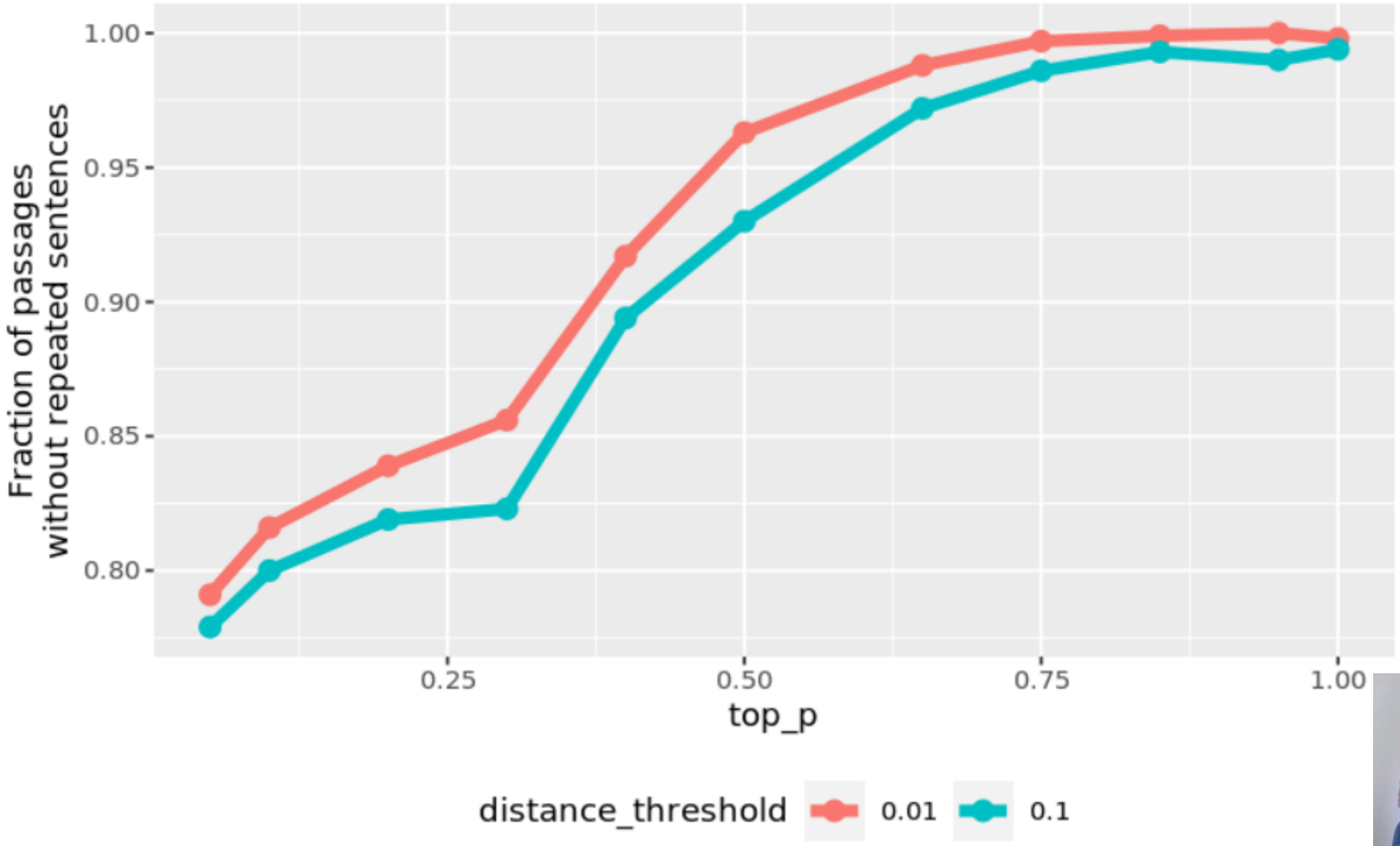
Brighter colors represent multiple points piled up; these are repeated sentences

Approximate repeats are close together, but not exactly on top of one another.

Clustering finds repetition within an encounter

Terrance Barton is a 1 year old male who was seen at urgent care for a fractured right forearm. He was taken by ambulance to the local hospital where he was treated and released to his mother. She reported that he had been seen at the local hospital 3 weeks earlier with a left femur fracture. The fracture occurred in the same location as the previous fracture. Mr. Barton was also seen 2 months ago for a fractured clavicle. This fracture occurred in the same location as the current fracture. Mr. Barton was also seen 2 months ago for a bruised hip. This fracture occurred in the same location as the current fracture. Mr. Barton was also seen 1 month ago for a broken wrist.

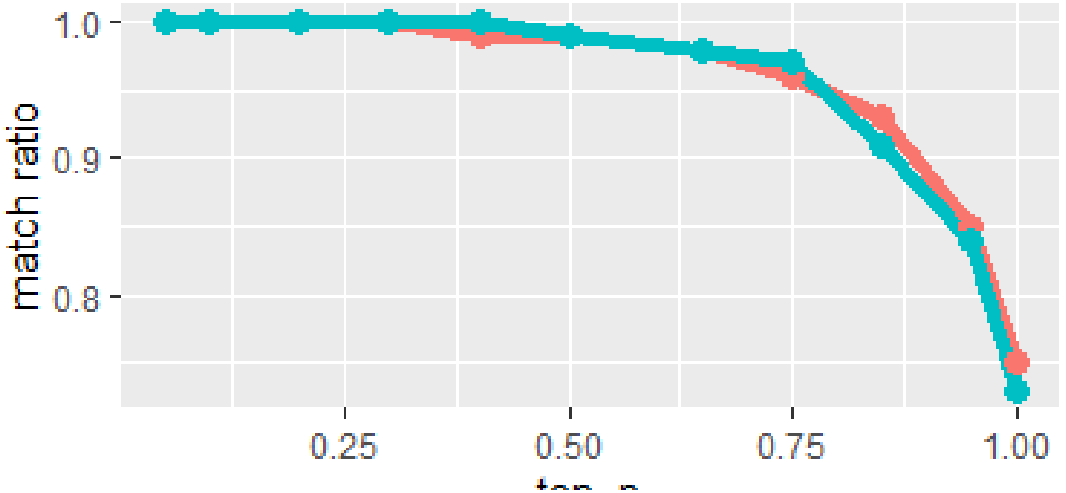
Higher values of top_p give fewer repeats

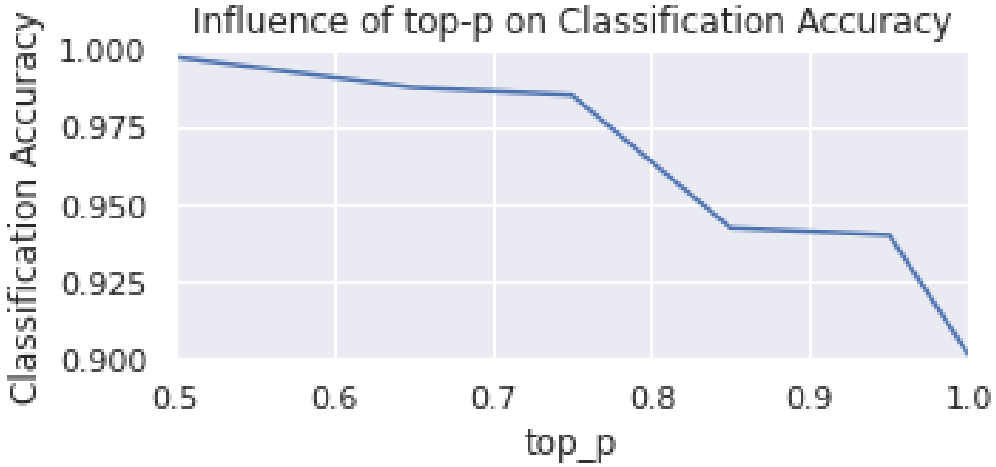


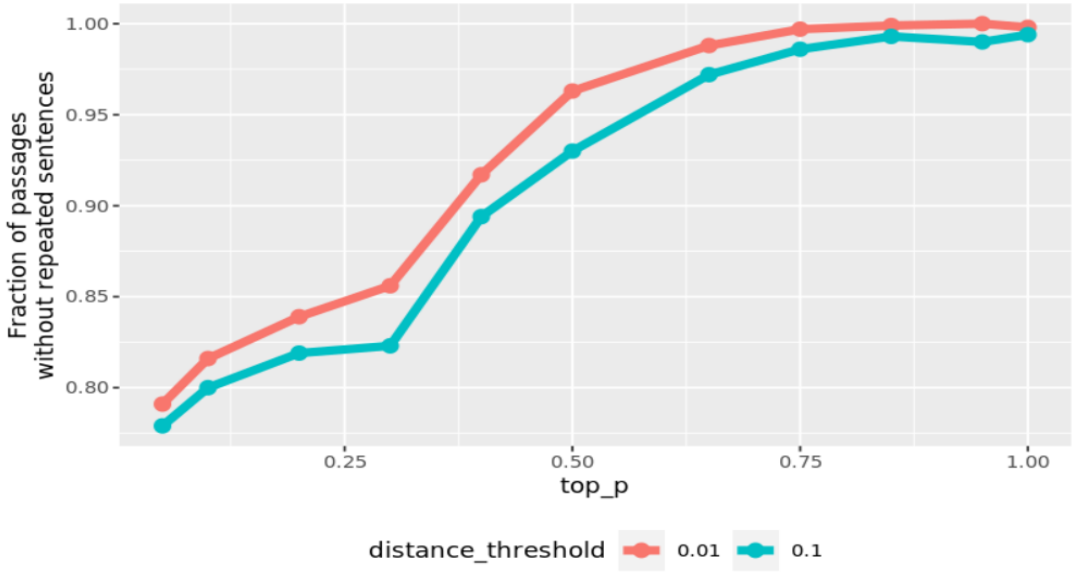
Discussion

Robert Horton, PhD – Senior Data & Applied Scientist

Finding the balance







Semantic errors

"... a 59 year old male, native Caucasian male ... "

"Mr. Nikolaus was a 16 y/o male ..."

"Mrs. Oretha Flatley is a 5 y/o female ..."

Domain-knowledge errors

"Melita is on the waiting list for an elbow replacement."

"... was given a sling and crutches for her ankle."

Microsoft

Thank you