


Identifying Missing Information in Text Using Graph-Based Approaches

Hustin Cao, Samika Gupta, Aylin Gunal





1.

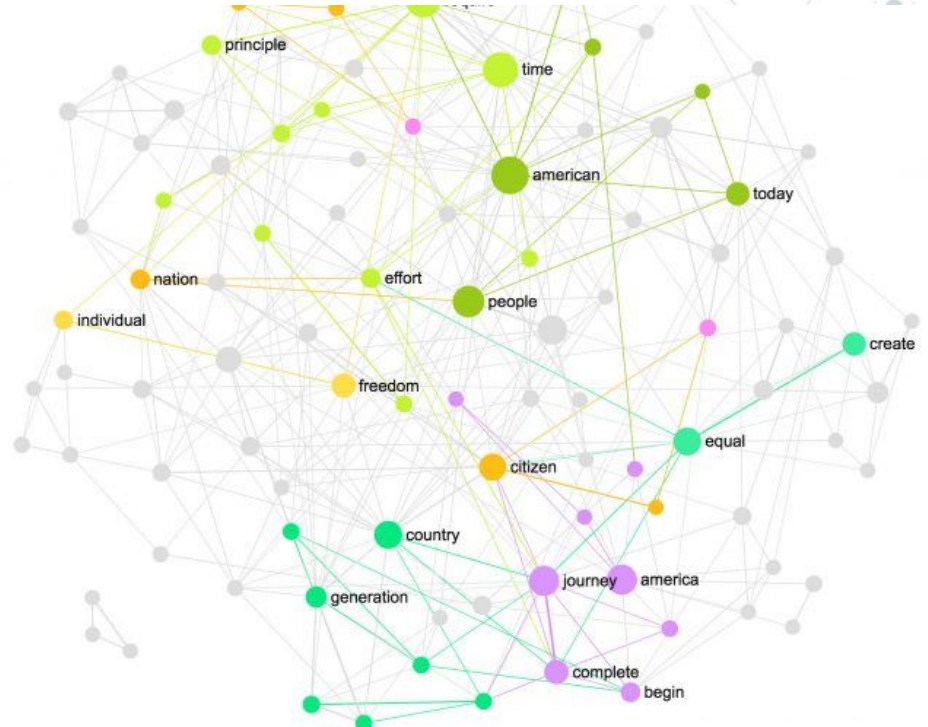
Problem Statement

How can we identify missing information in text?

Introduction

Motivation

Information retrieval tasks that rely on a knowledge base can be hindered by missing or partial information.



Example Scenarios

Clarification Questions (Ambiguity)

"What are the languages used to create the source code of Midori?"

Is the question referring to the **web-browser Midori** or the **operating-system Midori**?

Knowledge Gap / Reading Comprehension

Which of these would let the most heat travel through?

- A) a new pair of jeans.
- B) a steel spoon in a cafeteria.
- C) a cotton candy at a store.
- D) a calvin klein cotton hat.

Text Summarization

Titles, abstracts, conclusions, article previews...

Summarizing an article or a corpus without losing the important aspects of the text.

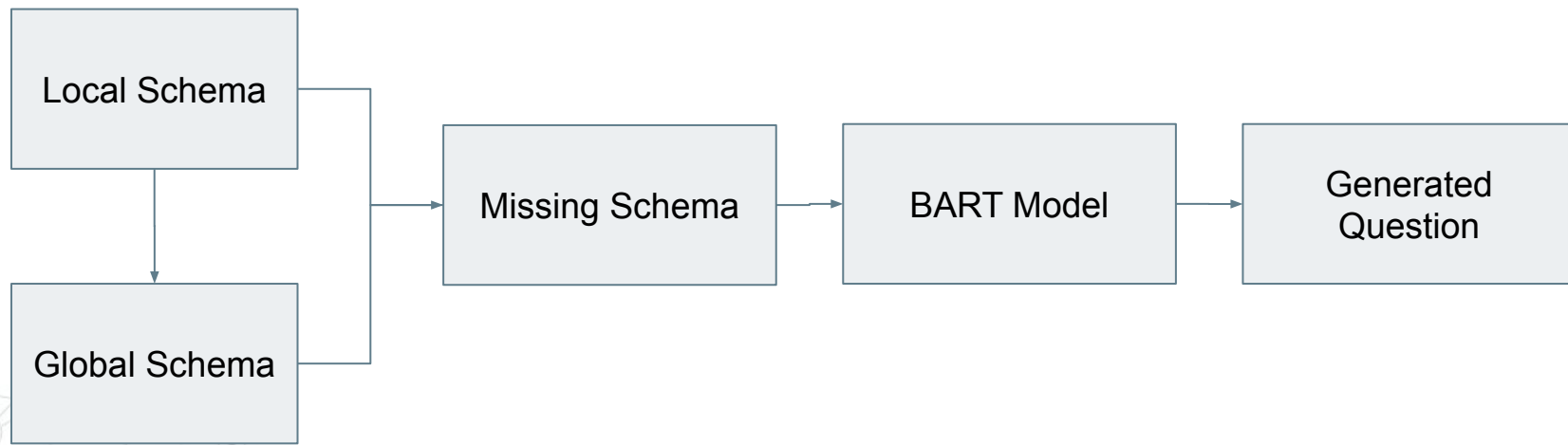
Goal: Can we identify the most important missing context?

A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. The nodes are represented by small circles, some of which are larger and have concentric circles, suggesting different levels of connectivity or importance. The lines are thin and gray, creating a mesh-like structure.

2. **Approach**

Method Overview

- Extract local and global schema
- From the local and global schemas, extract missing schema
- Feed missing schema to question generation model



Data

- **Data source**
 - Amazon Reviews and Question-Answer Datasets
- **Data filtering**
 - Select the Groceries and Gourmet Food category
 - Total dataset: 385 product description-question pairs
 - 90% training, 10% validation

QA Dataset: <https://jmcauley.ucsd.edu/data/amazon/qa/>

Reviews Dataset: <https://nijianmo.github.io/amazon/index.html>

Example || Great Lakes Select Honey, 32-Ounce Bottles (Pack of 3)



Description // Local Schema

- Great Lakes Select Honey is a U.S. Grade A , True Source Certified, Extra Light Amber, honey. This honey has a light floral taste reminiscent of the floral smells of its namesake region. This honey is great for mixing in tea, spreading on your morning toast, or using in a recipe. Honey is a natural sweetener for your healthy lifestyle. Check out all our Great Lakes honeys on this site.

User Questions // Gold Standard Labels

- Is this 100% pure honey?
- Where in Michigan is this made (I'm guessing that would be near where the honey is harvested)?
Is this by any chance Raw Honey?

All Descriptions Combined // Global Schema





3.

Baseline Method

Baseline Method

Paper: “Ask what’s missing and what’s useful: Improving Clarification Question Generation using Global Knowledge”

Schema Definition

The paper defines the schema of a sentence s :

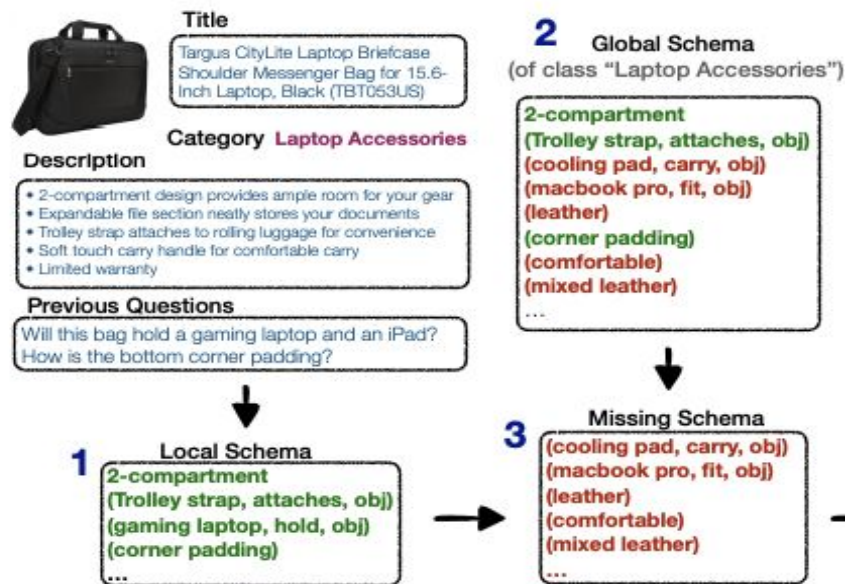
$$\text{schema}_s = \{ \text{element} \}$$
$$\text{element} \in \{ (\text{key-phrase}, \text{verb}, \text{relation}), \text{key-phrase} \}$$

Identifying Missing Schema

The global schemas are created by combining all of the local schemas.

The missing schema for each local schema is defined by the set difference of the global and local schema.

Literature paper: <https://aclanthology.org/2021.naacl-main.340.pdf>





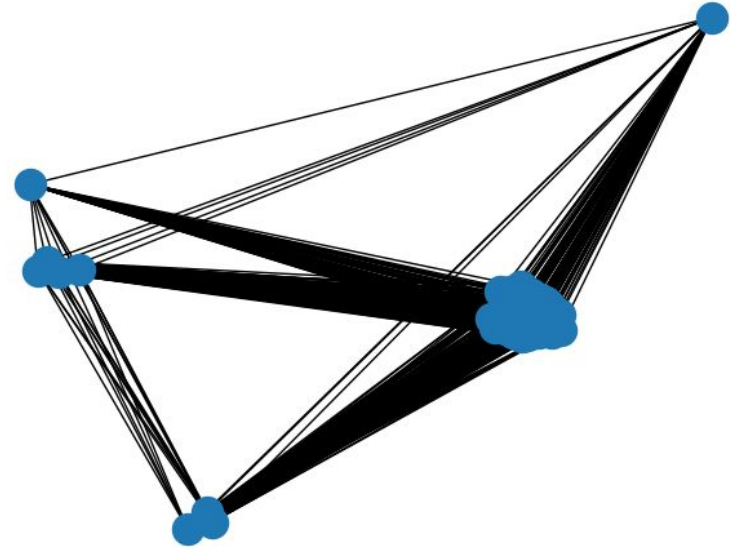
4.

Using a Graph-Based Approach

Graph Based Approach

Graph Based Method

- © Tokenize all local descriptions by sentence and assign each sentence to a node
- © Cluster similar nodes together using cosine similarity to form global clusters
- © Compare local description to global clusters; the global clusters that are **least similar to the local description** comprise the missing schema.



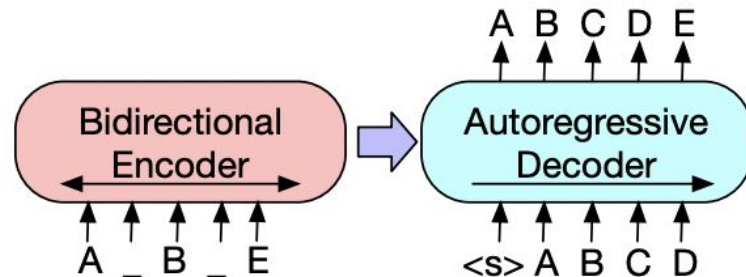
Core Clustering Algorithm

1. **Compute the density variation sequence**
 - a. Ex: $[a, b, c, d] \rightarrow [a, c] \rightarrow [c] \rightarrow []$
2. **Identify core nodes**
 - a. Iterations from DSV calculation that fit certain requirements
3. **Partition core nodes**
4. **Expand core clusters**
 - a. Assign non-core nodes to clusters

Compare local descriptions to global clusters; global clusters that are least similar to a local description will make up the missing schema to be fed into the BART model as input.

BART Model

- ◎ A denoising autoencoder for pretraining sequence-to-sequence models.
- ◎ BART is trained by
 - (1) Corrupting text with an arbitrary noising function
 - (2) Learning a model to reconstruct the original text.
- ◎ It uses a standard Transformer-based neural machine translation architecture
- ◎ We used the Hugging Face Transformer library, BartforConditionalGeneration using bart-base with input of (question, missing info) pairs for output of questions.



A decorative network diagram in the top-left corner, featuring a complex web of interconnected nodes and lines. The nodes are represented by small circles, some of which are highlighted with a double-circle outline. The lines are thin and gray, creating a mesh-like structure.

5.

Results & Evaluation

Automatic Quantitative Evaluation

	BLEU-4	METEOR	Google BLEU
Baseline	0.918	0.951	0.944
Graph-based	0.975	0.989	0.998

*Calculations based on 90 product descriptions

Results

Model Input: Net **Carbs** 2gTotal **Carbs**: 19gNon-Impact **Carbs**: -17gThe -17 non-impact **carbs** are derived from maltitol (a sugar alcohol) and dietary fiber. Unsweetened Chocolate, Maltitol (from wheat but gluten free), Cocoa Butter, Cream (Milk), Cocoa Mass, Soy Lecithin, Natural Vanilla, may contain traces of various nuts. Statements regarding dietary supplements have not been evaluated by the FDA and are not intended to diagnose, treat, cure, or prevent any disease or health condition.

- Baseline Generated Question: How many **carbs** per serving?
- Graph Based Generated Question: How many calories per packet?

Model Input: Unsulfured blackstrap molasses is more than just a sweetener. Its full of health promoting vitamins and minerals and makes a great **nutritional** supplement. Taken from the second pressing of the cane, all of the white sugar has been removed from this molasses leaving all the trace elements, especially all of those valuable minerals. Delicious mixed in herbal teas or in hot water with lemon and milk. There are 3 different grades of molasses. The first and second grades are fairly sweet. Blackstrap molasses is the third grade its not very sweet, but it does have the highest vitamin and mineral content of all three grades. 100% certified organic blackstrap molasses. Some foam at the top is normal. Statements regarding dietary supplements have not been evaluated by the FDA and are not intended to diagnose, treat, cure, or prevent any disease or health condition.

- Baseline Generated Question: What is the shelf life of this product?
- Graph Based Generated Question: What is the **nutritional** content of this product?

Future Work

Human Annotation

Since the questions generated are subjective to human knowledge, it is important to also incorporate human evaluation.

Therefore, if time permits, we hope to do larger scale comparisons between the questions resulting from the two methods to decipher if there is a significant difference.

Additional Clustering Methods

Reinforcing our hypothesis that graph-clustering methods work well.

PPLM (Plug and Play Model)

In baseline method, the major improvements that were made were using BART + PPLM, which incorporates the results of human annotation. We would like to run our model after adding PPLM into our pipeline and observing the results.

More and Diverse Data

We trained our models on a relatively small set. Larger datasets and more diverse category sets may improve our question generation.

The background of the slide features a complex, light gray network pattern. It consists of numerous small circles, some of which are solid gray and others are hollow with a gray outline. These circles are interconnected by a web of thin, light gray lines, creating a dense, interconnected mesh that resembles a molecular structure or a data network.

Thank you!

Reach out with questions or for codebase!
{gunala, samika, hustin} @ umich . edu

Github repo*:

https://github.com/aylingunal/EECS592_CourseProject

*full and commented codebase will be uploaded by Friday Dec. 9th

References:

Core clustering: <https://ieeexplore.ieee.org/document/4760954>

Main Clarification paper: <https://aclanthology.org/2021.naacl-main.340.pdf>

QA Dataset: <https://jmcauley.ucsd.edu/data/amazon/qa/>

Reviews Dataset: <https://nijianmo.github.io/amazon/index.html>

Text segmentation paper: <https://aclanthology.org/S16-2016/>

PPLM paper: <https://arxiv.org/abs/1912.02164>

BART paper: <http://arxiv.org/abs/1910.13461>

BLEU paper: <https://aclanthology.org/P02-1040.pdf>

METEOR paper: <https://aclanthology.org/W14-3348/>