Life is a Circus and We are the Clowns: Automatically Finding Analogies between Situations and Processes

Oren Sultan

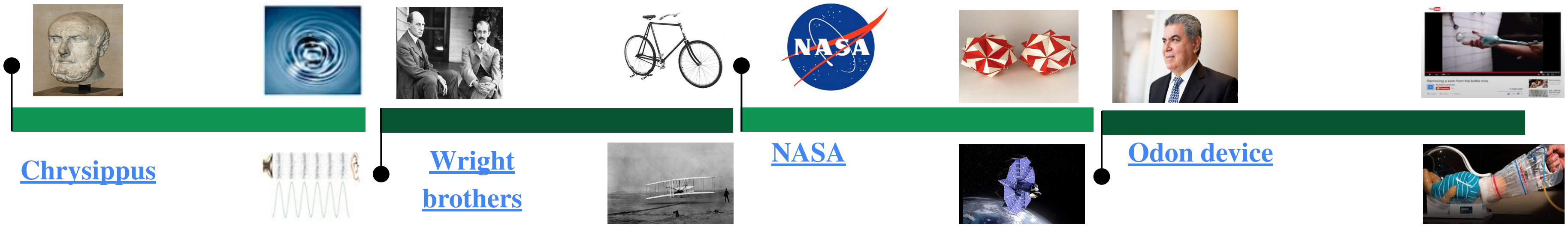
Advisor: Prof. Dafna Shahaf

School of Computer Science & Engineering, The Hebrew University of Jerusalem

*Proceedings of EMNLP 2022 (NLP Applications Track)

Introduction & Motivation – Analogies in **human cognition**

- Analogy-making is a **central** part of **Human Cognition** *(Minsky, 1988; Hofstadter and Sander, 2013; Holyoak, 1984)*
    - Abstract information.
    - Important role in many areas (e.g, education, politics, etc.).
    - Inventions throughout history.



Analogies in **Artificial Intelligence (AI)**

- Analogies are **essentials** for **Artificial Intelligence (AI)** *(Mitchell, 2021)*
  - Key to non-brittle AI systems that can adapt to new domains, and form humanlike concepts and abstractions.

- Analogies in **Natural Language Processing** (**NLP**)
  - Most works focused on word analogies - “a to b is like c to d” *(Mikolov, 2013)*

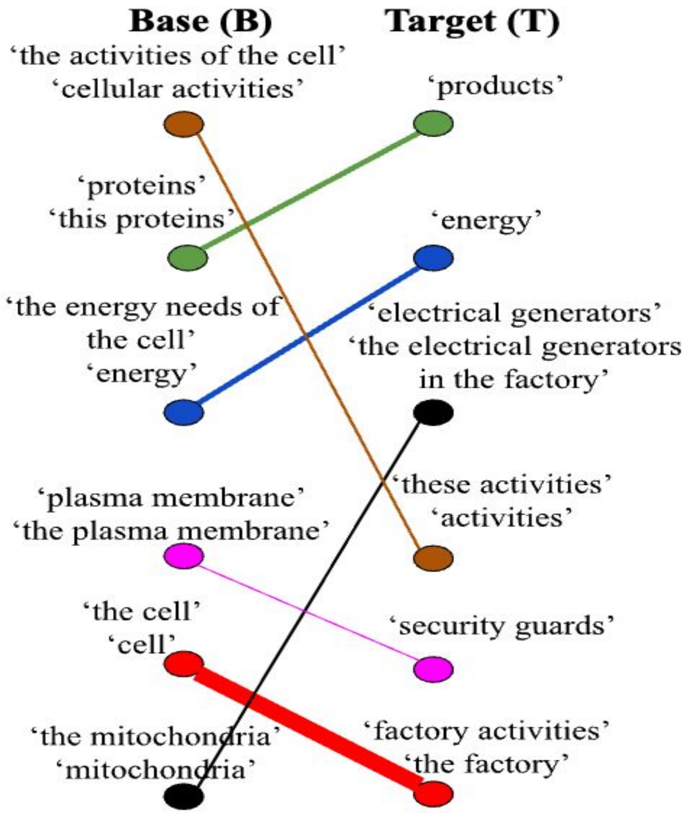
- **Our focus**: Analogies between **situations** and **processes**
  - Structure Mapping Engine (SME) *(Gentner, 1983; Falkenhainer, 1989; Turney, 2008; Forbus, 2011)*
    - **Input**: two domains (e.g., how the heart works / how a pump works).
    - **Goal**: map objects from **base** to **target** according to **relational structure** rather than **object attributes**.
    - **Problem**: the domain descriptions in a highly structured language.

```
CAUSE(PULL(piston), CAUSE(GREATER(PRESSURE(water),PRESSURE(pipe)),FLOW(water, pipe)))
```

Our work: we tackle a more **realistic** setting – **analogies** between **natural language procedural texts** describing **situations or processes**

Base: Animal Cell
The plasma membrane encloses the animal cell. It controls the movement of materials into and out of the cell. The Nucleus controls the activities of the cell. These cellular activities require energy. The Mitochondria extract energy from food molecules to provide the energy needs of the cell. Animal cells must also synthesize a variety of proteins and other organic molecules necessary for growth and repair. Ribosomes produce these proteins. The cell may use these proteins or move them out of the cell for use in other cells. To move organic molecules, the cell contains a complex system of membranes that create channels within the cell. This system of membranes is called the endoplasmic reticulum.

Target: Factory
Security guards monitor the doors of the factory. They control the movement of people into and out of the factory. Factory activities may be coordinated by a control center. These activities require energy. The electrical generators in the factory provide energy. The factory synthesizes products from raw materials using machines. The factory has hallways to move products through it.



Problem Formulation

Entities: Let $\mathcal{B} = \{b_1, ..., b_n\}, \mathcal{T} = \{t_1, ..., t_m\}$ – entities in the domains (**nouns**).

Relations: Let $\mathcal{R}$ – set of relations – a set of **ordered** entity pairs.

- We focus on **verbs**. (e.g, *"mitochondria* **provides** *energy"*)
- Let $\mathcal{R}(e_1, e_2) \subseteq 2^{\mathcal{R}}$ – set of relations between two entities.

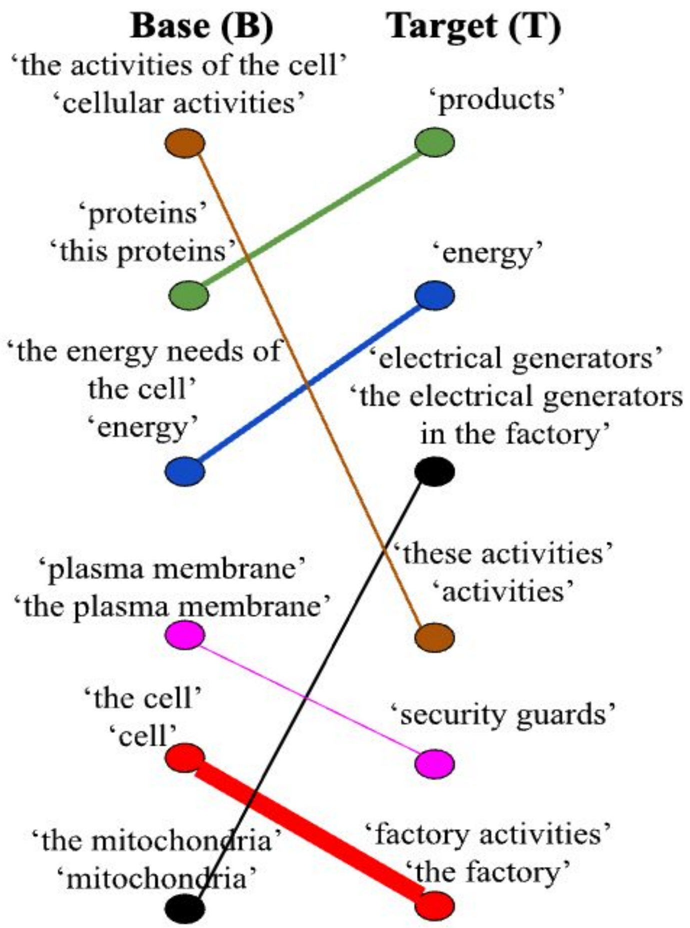
Similarity: Let $sim : 2^{\mathcal{R}} \times 2^{\mathcal{R}} \to [0, \infty)$– similarity metric between two sets of relations. High **Similarity** ↔ two sets **share many distinct** relations.

$$sim^*(b_i, b_j, t_k, t_l) = sim(\mathcal{R}(b_i, b_j), \mathcal{R}(t_k, t_l)) + sim(\mathcal{R}(b_j, b_i), \mathcal{R}(t_l, t_k)) \quad (1)$$

Objective: find a **consistent mapping** function $\mathcal{M} : \mathcal{B} \to \mathcal{T} \cup \perp$

- We look for a mapping that maximizes the **relational similarity** between mapped entities:

$$\mathcal{M}^* = \arg\max_{\mathcal{M}} \sum_{\substack{j \in [1, n-1] \\ i \in [j+1, n]}} sim^*(b_j, b_i, \mathcal{M}(b_j), \mathcal{M}(b_i)) \quad (2)$$



Our Method – Analogous Matching Algorithm

- Text processing
- Structure Extraction
- Clustering Entities
- Find Mappings

Text Processing

- **Chunking** the sentences in the input.

- Resolve **pronouns**
  - Apply **co-reference model** *(Kirstain, 2021)* which generates clusters
    (e.g, “the plasma membrane“, “plasma membrane“, “**it**“)
  - Replace all **pronouns** by a **representative** from the cluster – the shortest string which isn’t a **pronoun** or a **verb**.

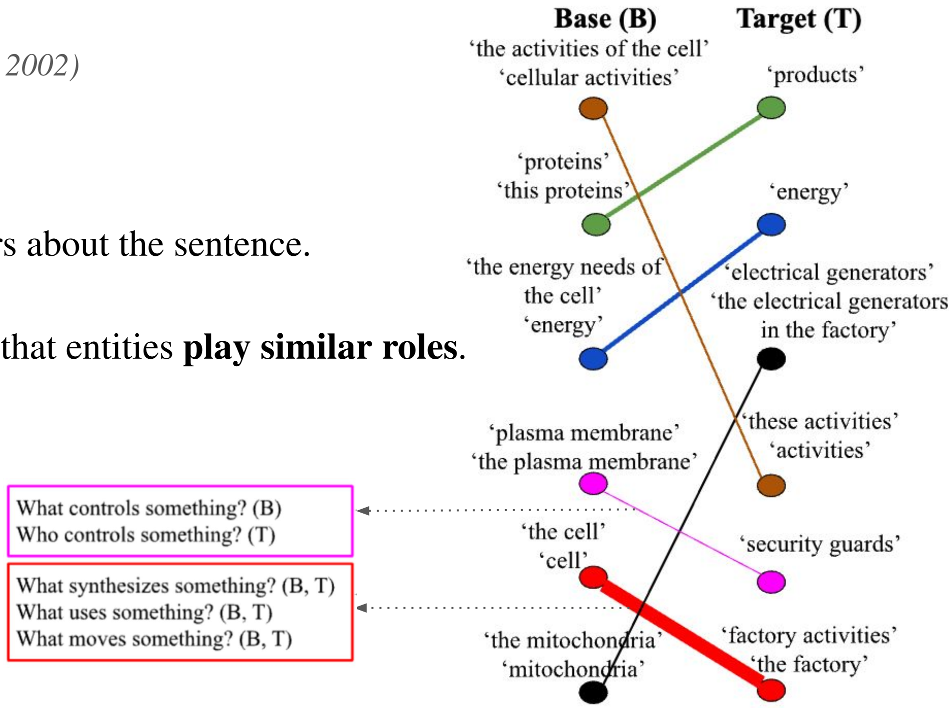
Structure Extraction

- **Q**: How can we know that **entities** in the domains **play similar roles**?
- **A**: We need to **extract the structure** in the texts (**entities** and their **relations**).

- **Semantic Role Labeling (SRL)** *(Gildea and Jurafsky, 2002)*

- **QA-SRL** model *(FitzGerald, 2018)*
  - **Input:** A sentence. **Output**: questions and answers about the sentence.
  - The **answers** form the **entities**.
  - **Similar questions** between the domains, indicate that entities **play similar roles**.

- **Considerations for extracting useful relations:**
  - Filter “When”, “Where”, “Why” questions.
  - Filter “Be” verbs.
  - Filter questions and answers with low probability.



Clustering Entities – **Agglomerative Clustering** *(Zepeda-Mendoza and Resendis, 2013)*

The animal cell

1) ’system of membranes’, ’a complex system of membranes’, ’this system of membranes’, ’membranes’, ’complex system of membranes’.
2) ’food molecules’, ’organic molecules’.
3) ’energy’, ’the energy needs of the cell’.
4) ’these proteins’, ’proteins’.
5) ’animal cells’, ’animal cell’, ’the animal cell’.
6) ’these cellular activities’, ’cellular activities’, ’the activities of the cell’.
7) ’the cell’, ’cell’.
8) ’nucleus’, ’the nucleus’.
9) ’endoplasmic reticulum’, ’the endoplasmic reticulum’.
10) ’mitochondria’, ’the mitochondria’.
11) ’channels’.
12) ’the plasma membrane’, ’plasma membrane’.
13) ’the activities’.
14) ’ribosomes’.
15) ’movement of materials’, ’the movement of materials’.

The factory

1) ’factory activities’, ’the factory’.
2) ’the electrical generators’, ’the electrical generators in the factory’, ’electrical generators’.
3) ’the doors’, ’the doors of the factory’.
4) ’these activities’, ’activities’.
5) ’hallways’.
6) ’machines’.
7) ’raw materials’.
8) ’energy’.
9) ’the movement of people’.
10) ’products’.
11) ’security guards’.
12) ’a control center’.

Find Mappings

- **Problem 1**: QA-SRL cannot detect relations across sentences, or using complex references.

"Animal cells must also produce proteins and other organic molecules necessary for growth and repair. Ribosomes are used for this process" / "The factory synthesizes products from raw materials using machines"

- We would like to infer that *ribosomes* **produce** *proteins* and *machines* **synthesize** *products*.
- QA-SRL gives us partial information: both *proteins* and *products* are associated with similar questions (**what is produced?**, **what is synthesized?**), hinting they might play similar roles.

- **Problem 2**: QA-SRL mentions just one entity per question.

- **Solution**: we propose a heuristic approach to approximate Equation 1.

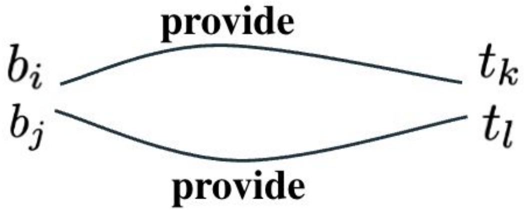
Find Mappings

- Intuitively, the **similarity score between two entities in the domains** is high if the similarity between their **associated questions** is high (e.g, **cell** and **factory** have multiple distinct similar questions).

- We define **similarity score between two entities in the domains :=** the **sum** of the **cosine distances** over their associated **questions' SBERT embeddings**.
  - We filter distances below a **similarity threshold** (manually fine-tuned)

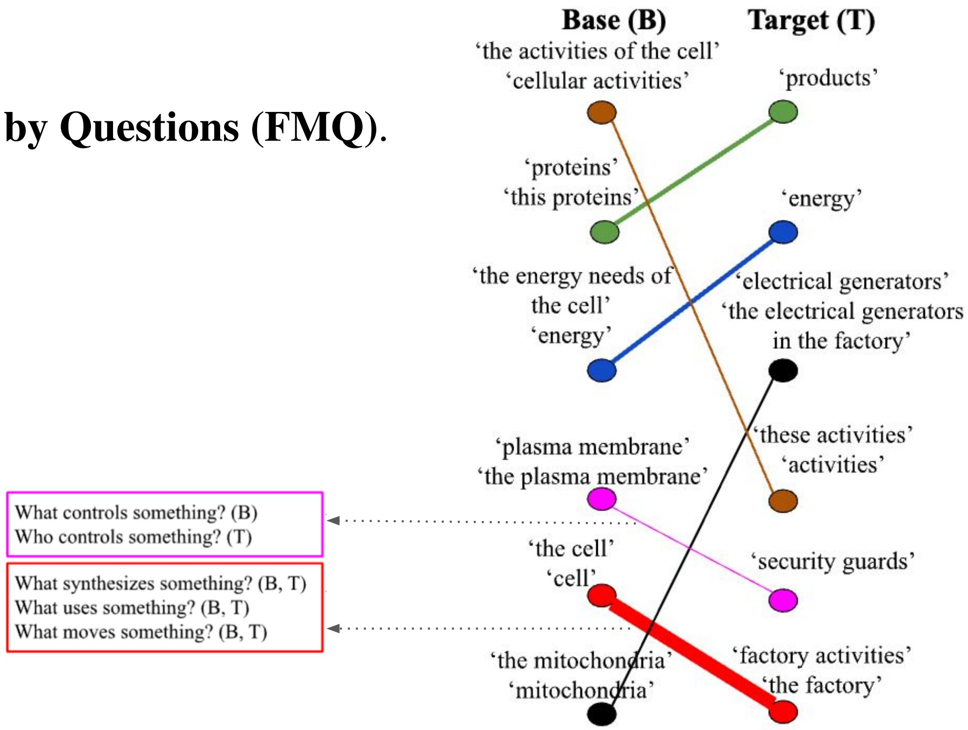
- Increasing the score for both mappings of **complete relations** (same verb)

What **provides** something?
What does something **provide**?



Beam Search

- After computing all similarities, we use beam search to find the mapping $\mathcal{M}^*$
- The mapping should be **consistent**.
- Our method is **interpretable**.
- We call our method: **Find Mappings by Questions (FMQ)**.



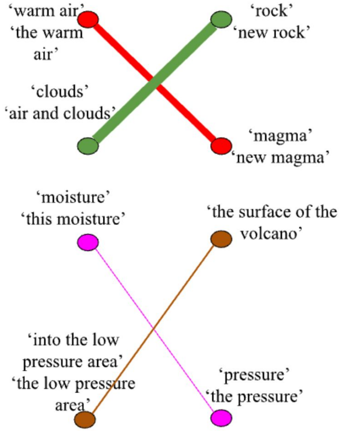
Experiments

- **RQ1**: Can we leverage our algorithm for **retrieving analogies** from a large dataset of procedural texts?
- **RQ2**: Does our algorithm produce the **correct mapping solution**?
- **RQ3**: Is our algorithm **robust to paraphrasing** the input texts?

- We tested our ideas on **ProPara** dataset *(Dalvi, 2018)* of crowdsourced paragraphs describing **processes**. (e.g, “What happens during photosynthesis?”) were given to 1-6 workers each.

PROMPT: Describe the process by which hurricanes form
Warm water floats up from the ocean.
A pocket of low pressure air is created near the surface of the ocean.
Warm air from outside areas pushes into the low pressure area.
The warm air keeps rising and forms clouds. The wind is getting fed moisture that evaporates from the ocean. This moisture causes the swirling air and clouds to grow bigger.
As it gets bigger the wind rotates faster.

PROMPT: What causes a volcano to erupt?
Magma rises from deep in the earth. The magma goes into volcanos. The volcanos pressure the magma upwards.
The pressure causes the magma to push through the surface of the volcano. The lava cools. The lava forms new rock. New magma is pressured to the surface of the volcano.
The volcano bursts through the rock that formed after the last eruption.



Experiment I: **Mining Analogies** – Setup

- **Goal**: Find analogies in the ProPara dataset.
  - Rank all **76K** possible pairs of paragraphs, so that analogies rise to the top.

- **Ranking formula**: multiplying #mappings by the median similarity. $|M| * \text{median}(\text{scores}(M))$

- **Baselines**: **to the best of our knowledge, there is no baseline that solves our task**.
  - **SBERT** *(Reimers and Gurevych, 2019)*
  - **Find Mappings by Verbs (FMV)**

- **Annotation:** top **100** pairs, as well as **40** pairs from all quartiles (bottom, middle, 25% and 75%)
  - **260** annotated pairs for each method's ranking list (**702** unique).

Experiment I: **Mining Analogies** – Labels

Label	**Description**
Not analogy	The texts are not analogous to each other.
Self analogy	Entities and their roles are identical (paragraphs on the same topic).
Close analogy	A close topic, entities from a similar domain.
Far analogy	Unrelated topics with different entities.
Sub analogy	Only a part of one process is analogous to a part of the other (>=2 similar relations).

Experiment I: **Mining Analogies** – Examples

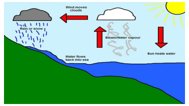
Examples for analogies mined by our method (FMQ):

B1 Prompt: Describe how oxygen reaches cells in the body

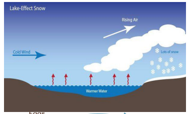
Self analogy

T1 Prompt: What do lungs do?

B2 Prompt: How does rain form?

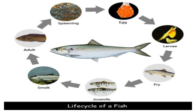


Close analogy

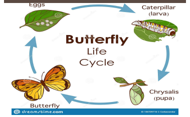


T2 Prompt: How does snow form?

B3 Prompt: Describe the life cycle of a fish.

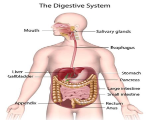


Far Analogy



T3 Prompt: Describe the life cycle of a butterfly.

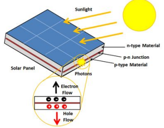
B4 Prompt: How does the digestive system work?



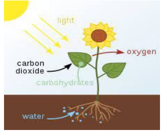
Far analogy

T4 Prompt: How does weathering cause rocks to break apart?

B5 Prompt: How does a solar panel work?

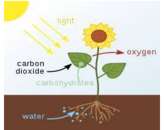


Far analogy

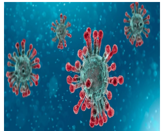


T5 Prompt: What happens during photosynthesis?

B6 Prompt: What happens during photosynthesis?



Not analogy



T6 Prompt: How does a virus infect an animal?

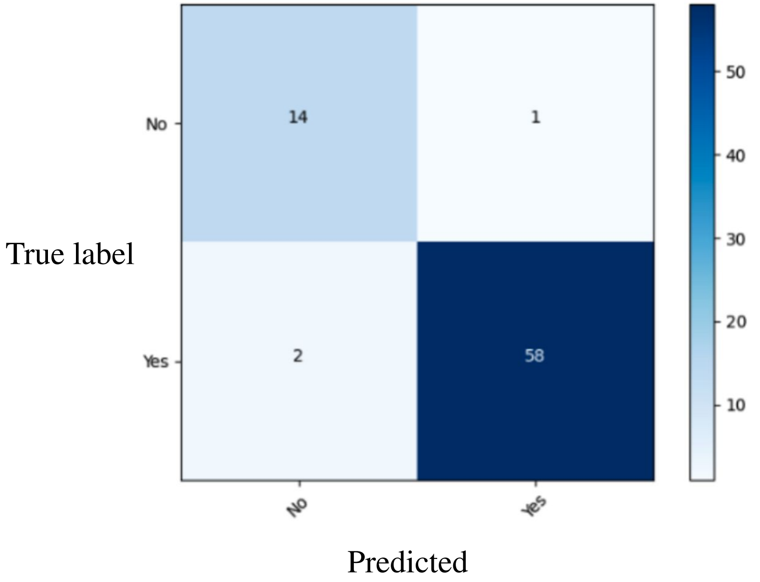
Experiment I: **Mining Analogies** – Annotation Process

- Our annotator (member of our team) annotated **702** unique pairs of paragraphs.
- **Goal**: to assess the clarity and consistency of our annotation scheme.
  - Our annotator (**=GT**), annotators (**=Predictions**)

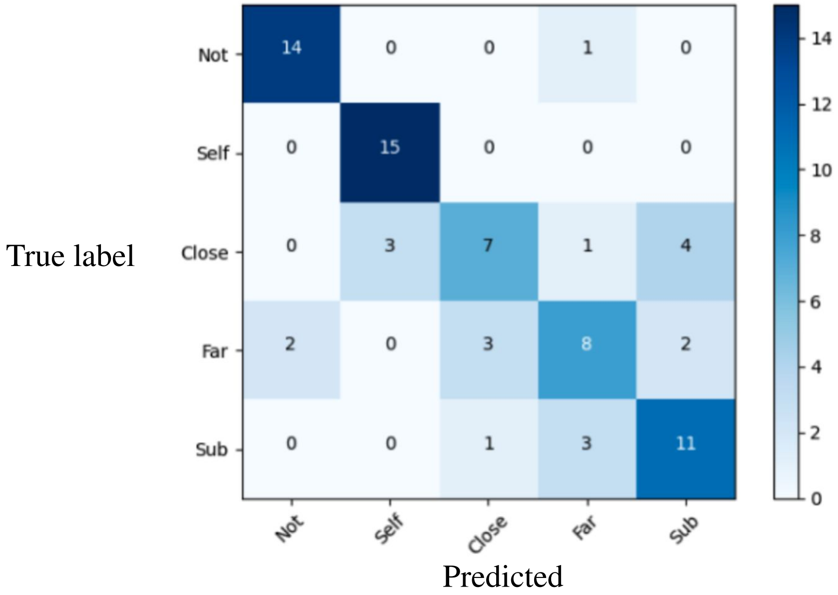
- **Check 1: agreement with another annotator of our team** (**10** pairs, **2** for each label)
  - **90%** agreement.
  - **Cohen's Kappa** of **0.74** for 2-labels and **0.88** for 5-labels. (mismatch in sub vs not analogy)

- **Check 2: 15** volunteer annotators
  - **Training – two** examples **for each label** with the **correct label** and **explanation**.
  - **Test** – we sampled from our annotator **5** pairs for each label, resulting in **25** pairs of paragraphs.
  - Each volunteer annotator received **5** pairs, s.t each pair is assigned to **3** annotators.

Experiment I: **Mining Analogies** – Annotation Process



Accuracy: 0.96, **Fleiss Kappa**: 0.82 (almost perfect)



Accuracy: 0.73, **Fleiss Kappa**: 0.58 (good)

- We conclude from the two sanity checks that our annotation schema is overall effective.

Experiment I: **Mining Analogies** – Results

- All methods had **zero** analogies in the 25%, middle, 75%, and bottom samples.
- **At the top**: SBERT was able to find almost only paragraphs on the same topic (self-analogies), our method was able to find many close and far analogies.

Method	Not	Sub	Self	Close	Far
SBERT	0	0	89	11	0
FMV	28	15	26	20	11
FMQ	21	16	29	18	16

Top-100 of the ranking

- **Analogies prevalence in data**: **~3%**

- We also show that FMQ wins FMV in terms of **IR metrics** (P, AP, NDCG)
  - Supporting our intuition that questions are more useful than verbs alone.

Experiment I: **Mining Analogies** – Results

Method	P	AP	NDCG
FMV (@25)	0.68	0.36	0.4
(@50)	0.72	0.37	0.41
(@75)	0.71	0.36	0.43
(@100)	0.72	0.36	0.43
FMQ (@25)	**0.96**	**0.5**	**0.57**
(@50)	**0.84**	**0.43**	**0.52**
(@75)	**0.77**	**0.39**	**0.47**
(@100)	**0.79**	**0.4**	**0.49**

- For NDCG we defined gains of 0, 1, 2, 3, 4 for not, sub, self, close, and far respectively.
- FMQ > FMV in all 3 metrics, supporting our intuition that questions are more useful than verbs alone.

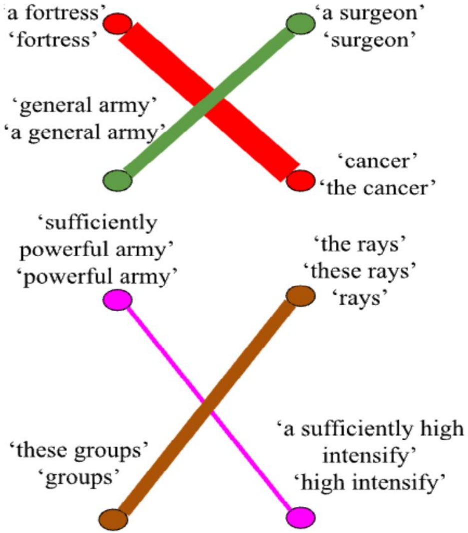
Experiment II: **Evaluating the Mappings** – Setup

- We chose **15** analogous pairs of paragraphs from **ProPara**
  - Equally divided between close, self, and far analogy.
  - We assigned one pair for every annotator, and asked them to find the correct mapping between the entities.

- Different kind of data – **14** pairs of **analogous stories** from cognitive-psychology literature.
  - We assigned these stories to 14 annotators, and asked them to do the same.

Base: The general
A general was trying to destroy a fortress which was situated at the center of a country with roads leading to it, by using his army. He needed to use his army as a complete group in order to destroy the fortress. However, he could not march his army down a road to the fortress because the roads were mined to explode when large groups of men passed over them. After considerable thought, he knew just what to do. He divided his army up into small groups of men, and by sending these groups, simultaneously, from a number of different directions, they converged on the fortress, making up a sufficiently powerful army to destroy it.

Target: The surgeon
A surgeon was trying to destroy a cancer which was situated in the central region of a patient's brain, by using a type of ray.
He needed to use these rays at a high intensity in order to destroy the cancerous tissue. However, at such an intensity the healthy brain tissue will also be destroyed. After considerable thought, he knew just what to do. He divided the rays up into batches of low-intensity rays, and then by sending them, simultaneously, from a number of different directions, they converged on the cancer, making up a sufficiently high intensity to destroy it.



Experiment II: **Evaluating the Mappings** – Annotation process

- We showed them **two** examples of correct mappings with explanations.§
- We emphasized that the mappings should be **consistent** and based on **roles entities play** in the texts.

- We consider the annotator’s mappings as **GT** and the algorithm’s mappings as **predictions**.

Experiment II: **Evaluating the Mappings** – Results

Dataset	Method	P	R	F1
ProPara	FMV (@1)	0.48	0.33	0.39
	FMQ (@1)	**0.82**	**0.64**	**0.72**
	FMV (@3)	0.58	0.40	0.47
	FMQ (@3)	**0.87**	**0.67**	**0.76**
Stories	FMV (@1)	0.64	0.46	0.54
	FMQ (@1)	**0.88**	**0.68**	**0.77**
	FMV (@3)	0.73	0.52	0.61
	FMQ (@3)	**0.94**	**0.76**	**0.84**

- Our method (**FMQ**) achieves a very high precision on **both datasets**!
- **FMQ > FMV**: Richer information provided by the questions.

Experiment III: **Robustness to Paraphrases** – Automatic paraphrases

- We chose **10** paragraphs which are **not analogous** to each other.
- For each paragraph, we generated **4 paraphrases** using **wordtune – two long** and **two short** versions (**50** paragraphs, or **1225** possible pairs).
- We labeled the **100** pairs that came from the same original paragraph with the label **True**, and the rest as **False**.
- We ranked all pairs via SBERT, FMV and FMQ.

Original paragraph: How do lungs work?
You breathe air in. Air enters bronchial tubes. Air is then split into the bronchioles of each lung. Bronchioles have alveoli which are tiny air sacs. Alveoli is surrounded by many small blood vessels. Oxygen passes through alveoli into blood vessels. Blood leaves the lungs as it travels through the body. Blood carries carbon dioxide back to the lungs. Carbon dioxide released when you exhale.

Wordtune expand:
When you breathe in, you are taking in air. Through your bronchial tubes, air enters your lungs. After the air has passed through the bronchial tubes, it is divided into the bronchioles of each lung. Alveoli, which are tiny sacs of air, are situated in the bronchioles. The alveoli are surrounded by a big number of small blood vessels. It is through these blood vessels that oxygen moves into the alveoli. In the course of its journey through the body, the blood enters through the lungs. When blood returns to the lungs, it takes carbon dioxide along with it. It is this carbon dioxide that is released when you breathe out.

Wordtune short:
Breathing air in. Bronchial tubes obtain air. Lungs split air into bronchioles. Alveoli are tiny air sacs in the bronchioles. Small vessels nearby alveoli. Alveoli grab oxygen to blood vessels. As blood passes through the body, it leaves the lungs. CO2 is carried by blood to the lungs. CO2 is discharged when you breathe out.

Experiment III: **Robustness to Paraphrases** – Responses to the same prompt

- We chose **10 non-analogous** paragraphs, and randomly chose **5** authors for each (**50** paragraphs, **1225** pairs, **100** analogous).
- Now texts are more **natural**, but can be **non-paraphrasing** anymore. (authors can focus on different aspects or granularity)

Original paragraph: Describe the life cycle of a human
A human baby develops in the womb of the mother. After 9 months in the womb the baby is born. It is an infant. The infant is dependent on its parents for everything. It drinks mother's milk for nourishment. From 3-8 years old the child is in early childhood. Adolescence is from roughly 9-18 years old. During adolescence the child is growing rapidly and maturing sexually. At 18 years, the child becomes an adult. Adults can reproduce and have babies.

V1:
A human is born. The human is an infant. The infant grows into a toddler. The toddler grows into a child. The child grows into a teenager. The teenager grows into an adult. The adult grows old. The human dies.

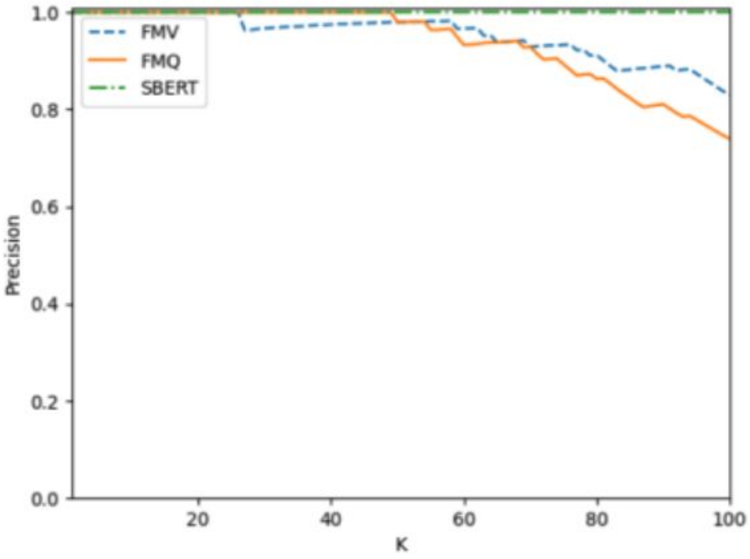
V2:
A human is born. The human is a child and learns. The human child grows into an adult. The adult uses its skills to survive. The human starts a new family and propogates. The human grows old. The human dies.

V3:
A zygote is formed via sexual reproduction. This zygtoe grows in the womb to become a fetus. After a typical 9-month period, a human is born. The human is an infant at this stage. The infant becomes a toddler, and learns to walk and speak. The toddler becomes a child. The child becomes a teenager after undergoing puberty. The teenager grows into an adult. The adult hits a peak, and development stops. Old age and eventually death occur.

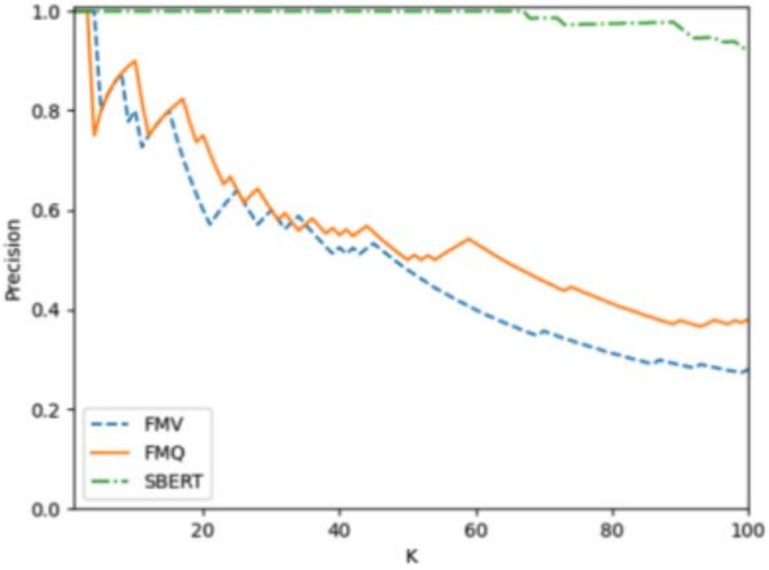
V4:
A sperm fertilizes an egg. The egg forms into a fetus. 9 months passes as the fetus grows into an infant. The infant is born. The baby begins to grow into an adolescent. The adolescent turns into a young adult. The young adult learns and grows into a fully mature adult.

Experiment III: **Robustness to Paraphrases** – Results

Automatic paraphrases



Responses to the same prompt



Experiment III: **Robustness to Paraphrases** – Error Analysis

- **False-Positives (FP)**
  - Non-analogous texts with similar verbs
  - QA-SRL handling of phrasal verbs (“take care”, “take off”)
  - Repeating verbs
  - Extraction issues (e.g, “Water, ice, and wind hit rocks” lead to singleton entities and “water, ice, and wind”, resulting in **double counting**).

- **False-Negatives (FN)**
  - Mistakes by wordtune (e.g., expanding “the water builds up” to “Nitrates build up in the body of the water”)
  - Mistakes in the GT – pairs of paragraphs describing the same topic from different points of view.

V1: how does internal combustion engine work?
Air and fuel are **used** in the internal combustion engine. In an enclosed chamber, a mixture of air and fuel is **injected**. The mixture **ignites** and **turns** a piston that **pumps** up and down. This piston is **connected** to a crankshaft which **rotates** to **provide** the power. The burned gas is **pushed** out of the chamber.

V2: how does internal combustion engine work?
The piston **moves** down. Gasoline and air **go** into the engine. The piston **moves** back up. The gasoline and air are **compressed**. The spark plug **emits** a spark. The gasoline **explodes**. The explosion **forces** the piston down. The exhaust valve **opens**. Exhaust **goes** to the tailpipe.

Real World Applications

- **Education**: analogies can help a teacher explain a complex concept.

- **Computer-assisted creativity**: engineers and designers could find inspiration in distant domains.

Future Work

- **Retrieve analogies from an existing dataset**
  - By combining our method's score and SBERT score (on a pair of paragraphs from ProPara), we see:

Our Score	SBERT Score	Label
high	low	far
high	medium	close
high	high	self
low	low	not

- **Our vision – create the first dataset of analogies between paragraphs of processes**
  - There is a demand for a dataset labeled with analogy / not analogy and the correct mappings.
    - Specifically, researchers are more interested in data of far analogies.
  - Can be used as a benchmark to train and evaluate models on this task.
  - One option is to use **ChatGPT** to assist in creating the dataset.

Future Work

- **Commonsense knowledge augmentation** – account for relations that do not appear in the text (e.g, we all know that *earth* **revolves** around the *sun*)
    - Study how much commonsense knowledge is missing in our data and could contribute to find more mappings.
    - **Retrieve relations**: How to retrieve these relations (data sources, GPT3, etc.).
    - **Candidates selection**: Choose the best candidates from the relations for both texts.
    - **Enrich the texts**: Generate new sentences in the texts by adding these new relations.

Future Work- Collaboration with Ayal Klein (BIU-NLP, Prof. Ido Dagan)

- **Potential incremental improvements:**
  - QASem (includes **QANom**) integration instead of QASRL + ablation study

Both were **shot** in the **confrontation** with police and have been **recovering** in hospital since the **attack** .			
QA-SRL	1	When was someone **shot?**	in the confrontation ; the attack
	2	Who was **shot?**	Both
	3	Who **shot** someone?	police
	4	Where has someone been **recovering?**	in hospital
	5	How long was someone **recovering** from something?	since the attack
	6	Who was **recovering** from something?	Both
	7	What was someone **recovering** from?	shot
QANom	8	Who **confronted** with something?	Both
	9	What did someone **confront** with?	police

  - Replacing SBERT questions similarity by semantic role alignment
  - **Broader alignment**: 2 entities in base vs. 2 entities in target.

Future Work- Collaboration with Ayal Klein (BIU-NLP, Prof. Ido Dagan)

- **Our vision – develop methods for other similar tasks:**
  - **Schema induction**. Example for what happens during photosynthesis/ how do solar panel works:
    - **Process**: photosynthesis / converting sunlight into electricity
    - **Actor**: plants, bacteria / solar panels, photovoltaic cells
    - **Materials**: carbon dioxide, water / silicon
    - **Output**: chemical energy / electricity

Conclusions

- Analogies are important for **humans** and **AI**.
- We explored analogies between **procedural texts** expressed in **natural language**.
- We develop a **scalable**, **interpretable** method to find **mappings** based on **relational similarity**.

- Our method was able to **mine different type of analogies** (in contrast to **SBERT**).
- Our method produced the correct mappings on both **ProPara** and the **Stories**.
- We showed our method is **robust to paraphrasing**.

Website

Thank You!

Data & Code

Paper

Video