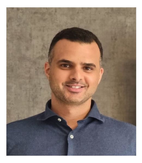
Life is a Circus and We are the Clowns 🤡: Automatically Finding Analogies between Situations and Processes

@oren_sultan
oren.sultan@mail.huji.ac.il



Oren Sultan and Dafna Shahaf

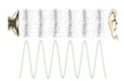
@HyadataLab
dshahaf@cs.huji.ac.il

Background and Motivation

Analogies in **human cognition:**

- Abstract information.
- Important role in many areas (e.g, education, politics, etc.).
- Inventions throughout history.

Chrysippus



NASA

Analogies in **AI:**

- Key to non-brittle systems that can adapt to new domains and form humanlike concepts and abstractions.

Analogies in **NLP:**

- Works focused on word analogies: "a to b is like c to d".

Our focus – analogies between **situations** and **processes:**

Structure Mapping Engine (SME)

- **Input**: two domains (how the heart works / how a pump works).
- **Goal**: map objects from **base** to **target** according to **relational structure** rather than **object attributes**. (e.g, **blood → water**)
- **Problem**: domain descriptions in a highly structured language.

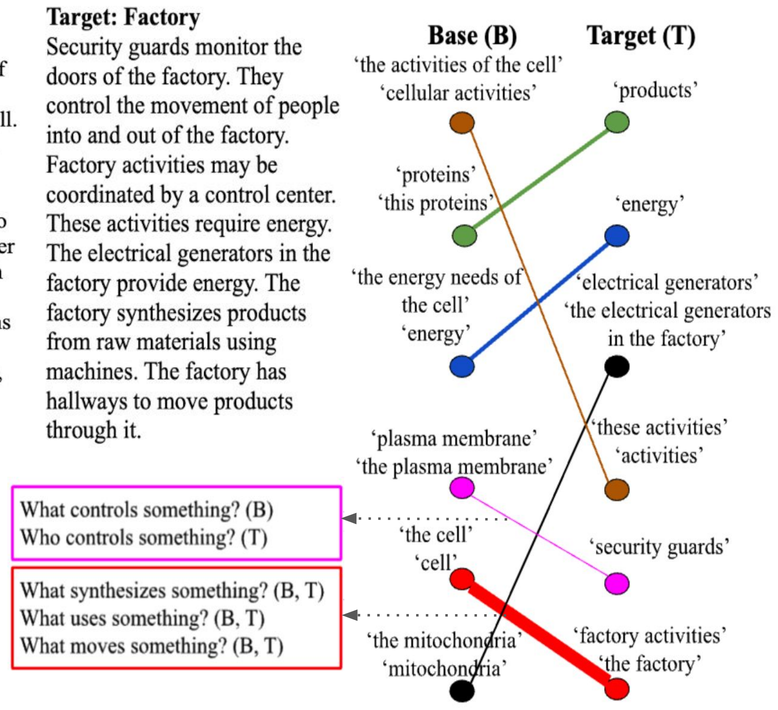
```
CAUSE(PULL(piston),CAUSE(GREATER(PRESSURE(water),
PRESSURE(pipe)),FLOW(water,pipe)))
```

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 $sim : 2^{\mathcal{R}} \times 2^{\mathcal{R}} \rightarrow [0, \infty)$ – 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 $\mathcal{M} : \mathcal{B} \rightarrow \mathcal{T} \cup \perp$

- 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 – Find Mappings by Questions (FMQ)

Text Processing → **Structure Extraction** → **Clustering Entities** → **Find Mappings**

Text Processing: chunking sentences & **coreference** (resolve **pronouns**)

Structure Extraction: How can we know **entities play similar roles**?

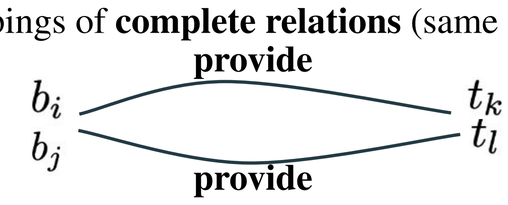
- **QA-SRL** model – **Input**: a sentence, **Output**: questions and answers.
- The spans identified as **answers** form the **entities**.
- **Similar questions** between domains, indicate entities **play similar roles**.

Clustering Entities: Agglomerative Clustering. For example:
{'animal cells','animal cell','cell', 'the cell', 'the animal cell'}

Find Mappings:

- **Problem 1**: Detection of relations across sentences, or complex references.
- **Problem 2**: QA-SRL mentions just one entity per question.
- **Solution**: A heuristic approach to approximate Equation 1.
  - **Similarity score between two entities in the domains** := **sum** of the **cosine distances** over their associated **questions' SBERT embeddings**.
  - Increasing the score for both mappings of **complete relations** (same verb)

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

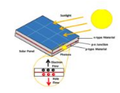


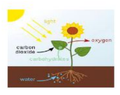
Results

Mining Analogies:

- **Goal:** To **mine analogies** from a large dataset of **procedural texts**, by **ranking** all pairs, s.t analogies rise to the top.
- **Dataset: ProPara**.
- **Methods**: **FMQ**, **FMV** (verbs)**, SBERT**
- **SBERT** – paragraphs on the **same topic**, **our method** – many **close** and **far** analogies (the more interesting ones).
- An example from the top of FMQ ranking:

Paragraph's Prompt:
How does a solar panel work?





Paragraph's Prompt:
What happens during photosynthesis?

- We also show that FMQ wins FMV in terms of **IR metrics** (P, AP, NDCG)

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

Evaluating the Mappings:

- **Goal:** To evaluate the **correctness of the mappings** produced by our method's solution between **pairs of texts**.
- **Dataset: ProPara, Stories.**
- **Methods**: **FMQ, FMV**.
- **Metrics**: **Precision**, **Recall**, **F1** score

(top-1 and top-3 solutions from beam search)

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**

Robustness to Paraphrases:

We show our method is **robust to paraphrasing** the input texts for both:

- **Automatic Paraphrases**: using **wordtune**.
- **The same prompt**: **different authors** writing on the same topic.

Conclusions & Future Work

- Analogies are important for **humans** and **AI**.
- We explored analogies between **natural language procedural texts**.
- Our method finds a **mapping** based on **relational similarity**.
- We show that our method can be used to **mine analogies,** produce the **correct mapping solution** and **robust to paraphrasing**.
- **Code repository**: https://github.com/orensul/analogies_mining
- **Future direction**: commonsense knowledge augmentation