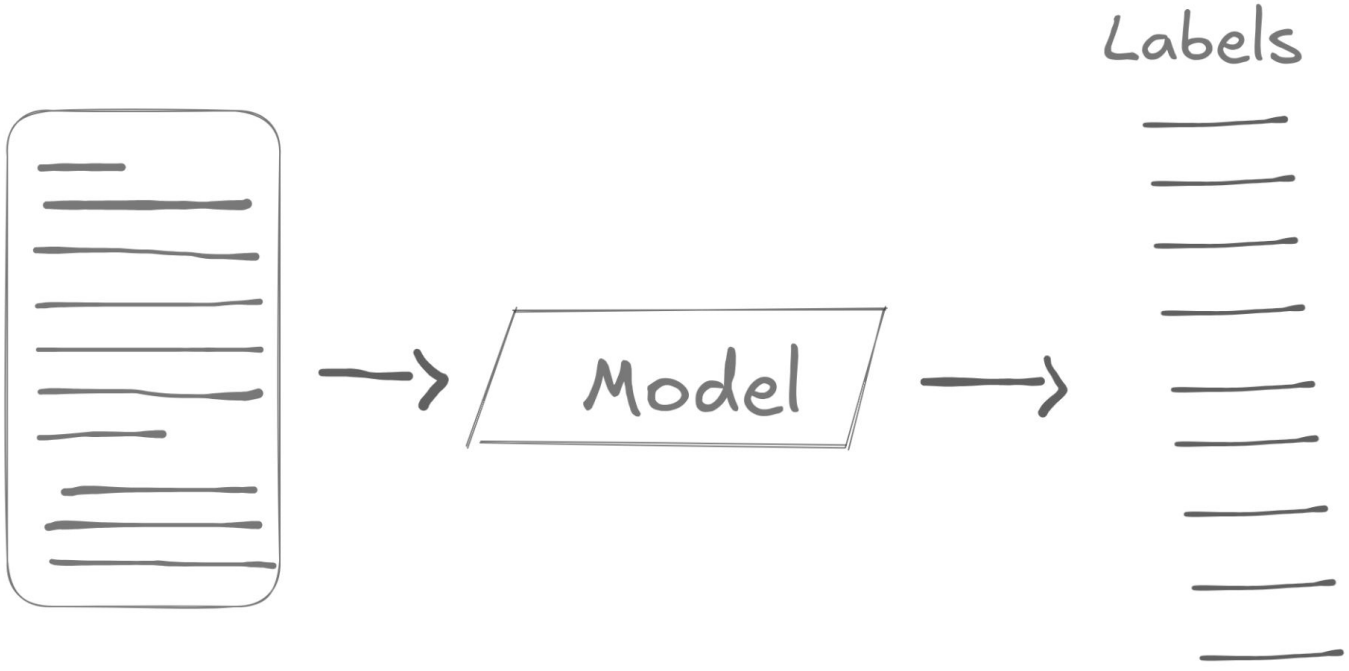
UNIVERSITÉ D'ARTOIS

centre de recherche en informatique de lens

Towards Better Concept Representations for Document Classification

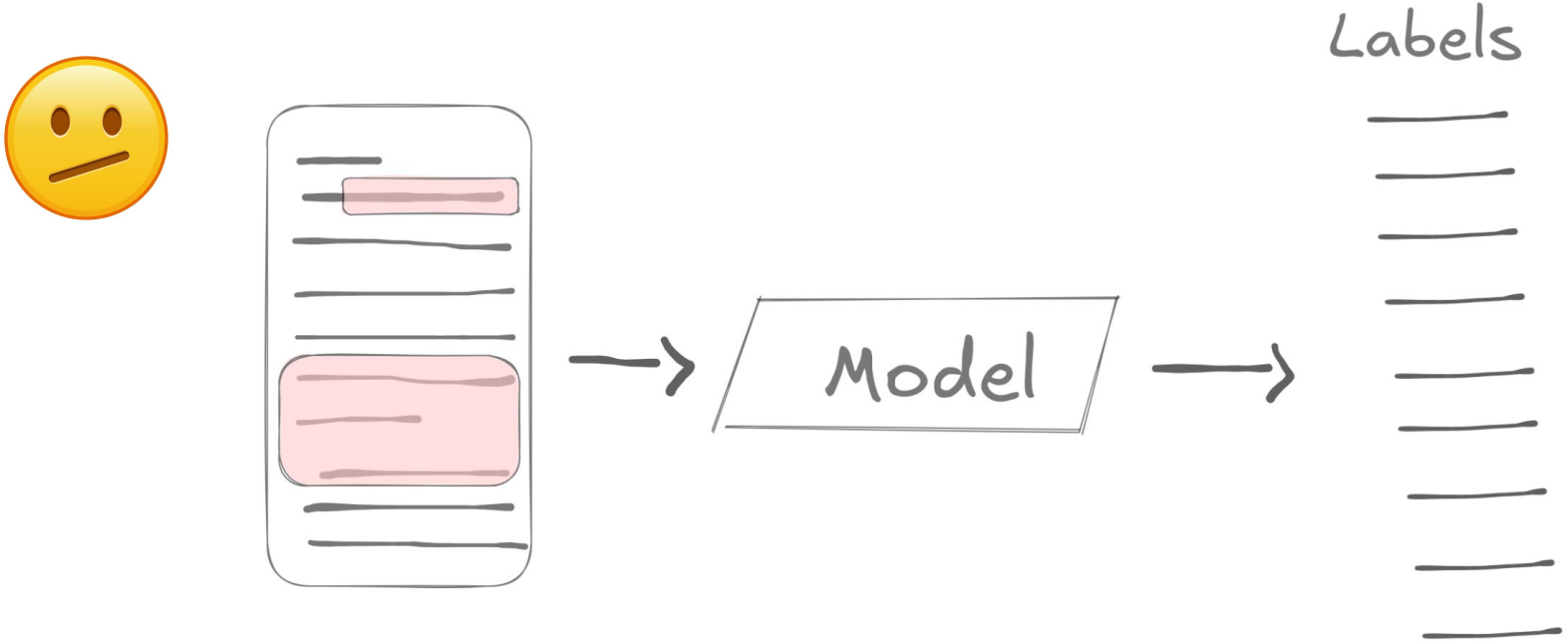
Context

The Challenge of Medical Document Classification



Context

The Challenge of Medical Document Classification



Motivation - Conventional Approaches

[0.56, 0.91, 0,03 …]

He withdrew money from the bank.

[0.74, 0.16, 0,23 …]

The bank is closed today.

[0.32, 0.64, 0,77 …]

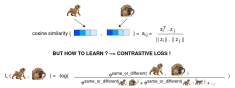
There are houses scattered along the river bank.

Improving Concept Embeddings with Fine-Tuned BERT

He withdrew money from the bank.

The bank is closed today.

There are houses scattered along the river bank.



Contrastive loss

BERT LM

He withdrew money from the bank.

The bank is closed today.

There are houses scattered along the river bank.

Distilling semantic concept embeddings from contrastively fine-tuned language models. Li et al. SIGIR 2023.

Improving Concept Embeddings with Fine-Tuned BERT

He withdrew money from th

The bank is closed toc

There are houses scattered along

cosine similarity ( , ) = $s_{i,j} = \frac{z_i^T \cdot z_j}{\| z_i \| \cdot \| z_j \|}$

BUT HOW TO LEARN ? --> CONTRASTIVE LOSS !

$L($  ,  $) = -\log\left(\frac{e^{\text{same\_or\_different}(}$  ,  $)}{e^{\text{same\_or\_different}(}$  ,  $) + e^{\text{same\_or\_different}(}$  ,  $) + \ldots} \right)$

iver bank.

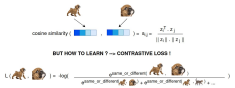
Distilling semantic concept embeddings from contrastively fine-tuned language models. Li et al. SIGIR 2023.

Improving Concept Embeddings with Fine-Tuned BERT

He withdrew money from the bank.

The bank is closed today.

There are houses scattered along the river bank.



Contrastive loss

BERT LM

He withdrew money from the bank.

The bank is closed today.

There are houses scattered along the river bank.

Distilling semantic concept embeddings from contrastively fine-tuned language models. Li et al. SIGIR 2023.

Problem Statement

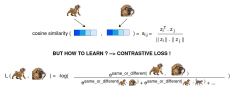
Input data → Language Model → Conclude concept representation → Task

Improving Concept Embeddings with Fine-Tuned BERT

He withdrew money from the bank.

The bank is closed today.

There are houses scattered along the river bank.



Contrastive loss

BERT LM

He withdrew money from the bank.

The bank is closed today.

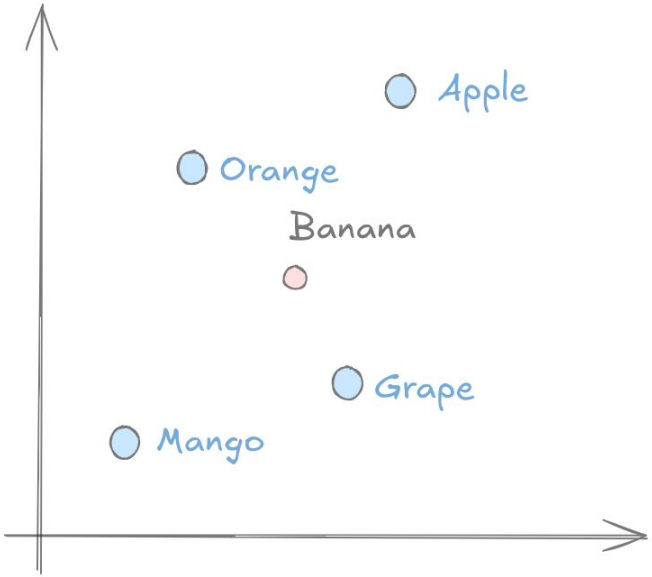
There are houses scattered along the river bank.

Distilling semantic concept embeddings from contrastively fine-tuned language models. Li et al. SIGIR 2023.

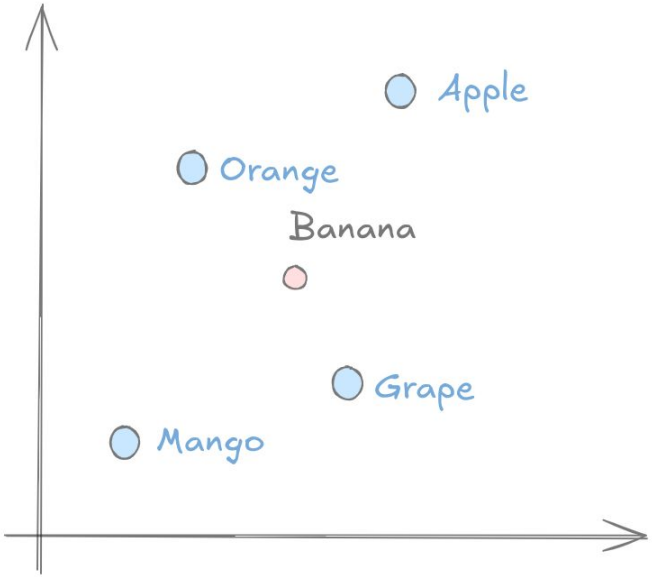
Problem Statement

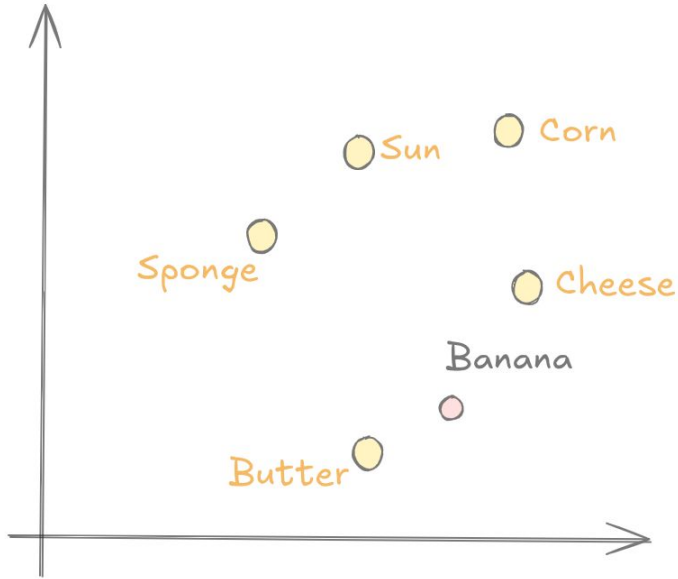
Cluster the words: banana, full moon, crescent moon, Lemon

Problem Statement

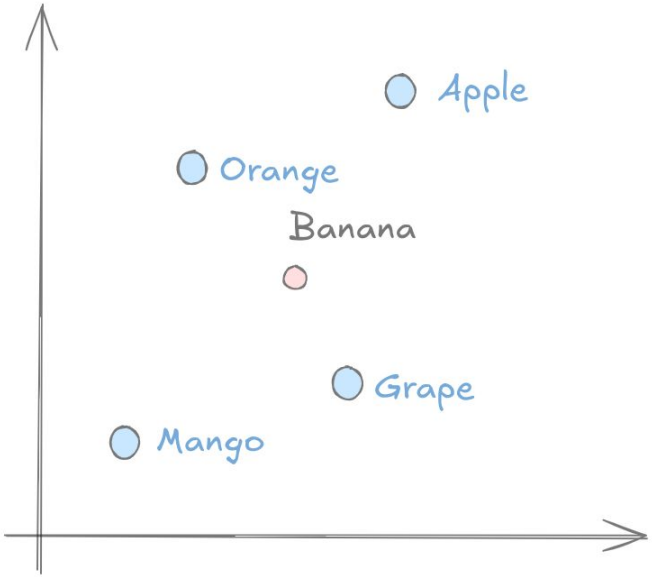


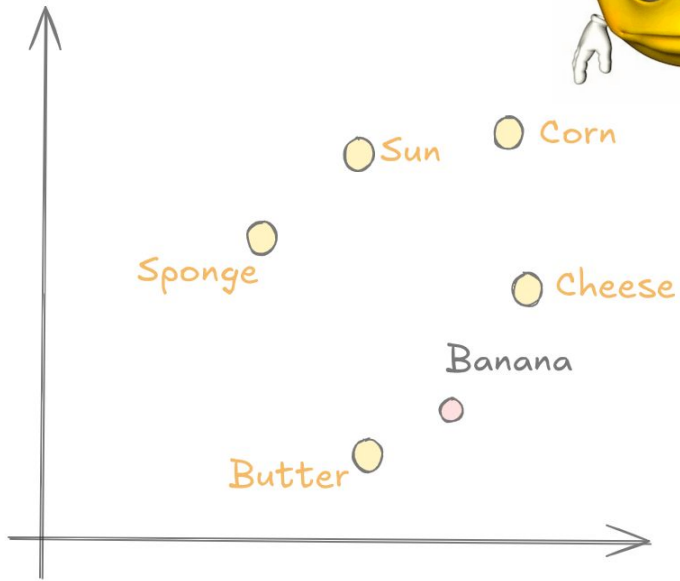
Problem Statement



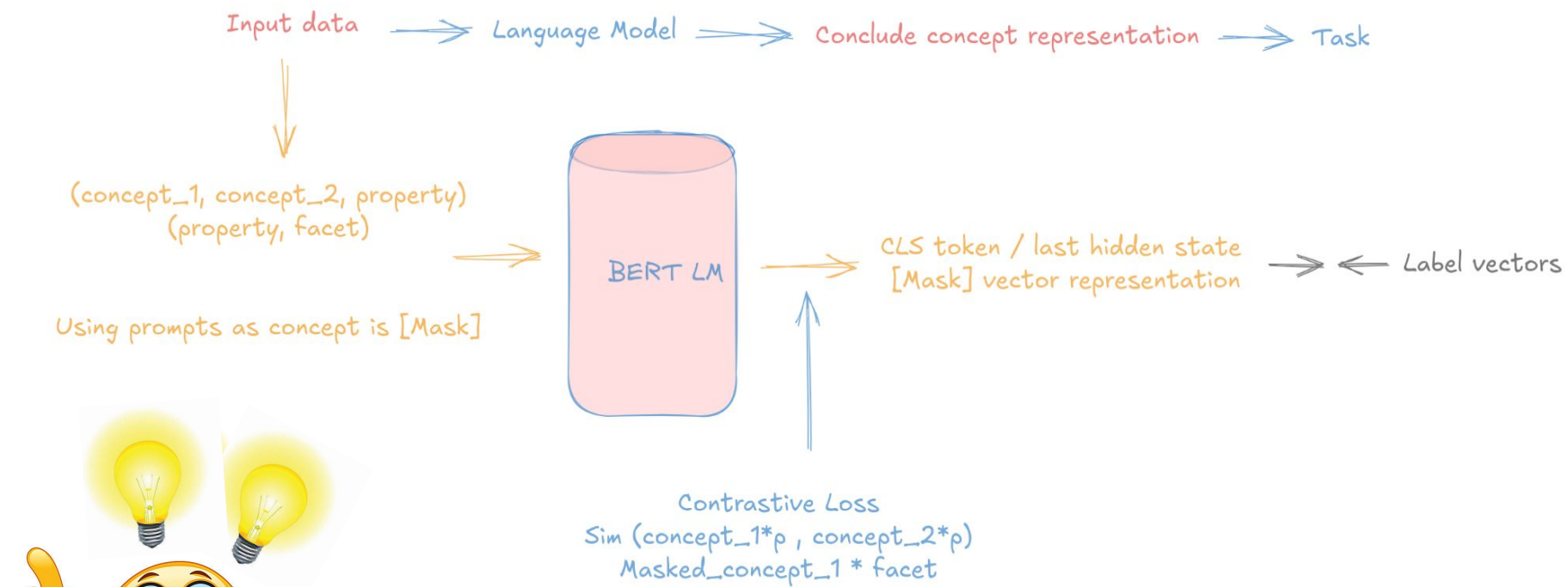


Problem Statement





Solution


Input data
Language Model
Conclude concept representation
Task
(concept_1, concept_2, property)
(property, facet)
Using prompts as concept is [Mask]
BERT LM
CLS token / last hidden state
[Mask] vector representation
Label vectors
Contrastive Loss
Sim (concept_1*p , concept_2*p)
Masked_concept_1 * facet

Multi-facets Concepts Embeddings

Training Dataset

concept-property judgements
banana has the property rich in potassium

property-facet judgments
rich in potassium refers to nutritional content

ChatGPT

I am interested in knowing common properties that are satisfied by different concepts.
1. Sound: loud "thunder, jet engine, siren"
2. Temperature: cold "ice, refrigerator, Antarctica"
3. Colour: orange "mandarin, basketball, clownfish"
4. Shape: round "sun, orange, ball"
5. Purpose: used for cleaning "broom, lemon, soap"
6. Location: located in the ocean "sand, whale, corals"

Please provide me with a list of 30 such examples.

In total, we obtained 828 unique facet-property pairs, covering 127 unique facets

ConceptNet

Triplet: (boat, at location, sea)
Concept-property pair: (boat, at location sea)
Property-facet pair: (at location sea, at location)

We specifically used: RelatedTo, FormOf, IsA, UsedFor, AtLocation, CapableOf, HasProperty, HasA, InstanceOf and MadeOf.

We thus end up with 884 distinct properties, 10 facets, 18505 concepts, 884 property-facet pairs and 36955 concept-property pairs.

Model formulation

$$\mathcal{L} = -\sum_{(c,p)\in D_{cp}} \log \sigma(\mathrm{Con}(c)\cdot \mathrm{Prop}(p)) - \sum_{(c,p)\in N_{cp}} \log(1-\sigma(\mathrm{Con}(c)\cdot \mathrm{Prop}(p)))$$

$$MC(c,p) = \frac{\mathrm{Con}(c)\odot \mathrm{Facet}(p)}{\|\mathrm{Con}(c)\odot \mathrm{Facet}(p)\|}$$

$$\mathcal{L}_1 = -\sum_{(c,p)\in D_{cp}} \log \sigma(MC(c,p)\cdot \mathrm{Prop}(p)) - \sum_{(c,p)\in N_{cp}} \log(1-\sigma(MC(c,p)\cdot \mathrm{Prop}(p)))$$

$$+ \quad \mathcal{L}_2 = -\sum_f \sum_{p,q\in P_f} \log \frac{\exp\left(\frac{\cos(F(p),F(q))}{\tau}\right)}{\sum_r \exp\left(\frac{\cos(F(p),F(r))}{\tau}\right)}$$

*A. Gajbhiye, L. Espinosa-Anke, and S. Schockaert. 2022. Modelling commonsense properties using pre-trained bi-encoders.

Extracting Facet-Specific Representations

Concept Neighbours

Frisbee
Facet 1: tricycle, surfboard, sports_ball, tennis_racket, snowboard, kite, doll, balloon, toy, pie
Facet 2: balloon, pie, sports_ball, cake, donut, teddy_bear, surfboard, kite, doll, toy
Facet 3: kite, doll, balloon, toy, tricycle, moth, pie, sports_ball, surfboard, football
Facet 4: tricycle, surfboard, sports_ball, tennis_racket, snowboard, kite, doll, balloon, toy, pie

Bureau
Facet 1: envelope, certificate, typewriter, doorknob, fence, cabinet, carpet, shelves, bookcase, gopher
Facet 2: desk, dining_table, table, shelves, bookcase, envelope, typewriter, gopher, escalator, peg
Facet 3: shelves, bookcase, envelope, desk, peg, gopher, tack, cabinet, handbag, hook
Facet 4: bookcase, cabinet, shelves, desk, doorknob, dining_table, envelope, typewriter, handbag, lamp

Experiments

	LM	Train properties ($\mathcal{D}_{cp}$)	Train facets ($\mathcal{D}_{pf}$)	McRae			CSLB		
				Con	Prop	C+P	Con	Prop	C+P
BiEnc*	BB	MSCG	-	79.8	49.6	44.5	54.5	39.1	32.6
BiEnc*	BL	MSCG	-	80.5	49.3	45.5	57.7	41.8	36.4
BiEnc*	RB	MSCG	-	75.6	42.4	38.1	49.9	36.4	24.3
BiEnc*	RL	MSCG	-	80.1	46.5	42.5	59.0	42.5	36.0
BiEnc	BL	CN	-	78.0	56.7	51.8	61.4	49.6	50.0
BiEnc	BL	ChatGPT	-	80.5	57.3	56.6	65.1	56.5	52.7
BiEnc	BL	ChatGPT+CN	-	81.7	62.1	59.5	67.8	59.6	53.1
BiEnc	BB	ChatGPT+CN	-	76.2	60.6	58.4	66.9	56.6	51.8
BiEnc	RB	ChatGPT+CN	-	75.8	60.1	58.2	66.1	56.3	51.7
BiEnc	RL	ChatGPT+CN	-	80.8	61.7	59.3	67.2	58.8	52.7
BiEnc-F	BL	ChatGPT+CN	CN	84.3	63.5	57.7	69.4	61.0	59.9
BiEnc-F	BL	ChatGPT+CN	ChatGPT	84.3	64.9	65.5	69.5	61.6	61.9
BiEnc-F	BL	ChatGPT+CN	ChatGPT+CN	**86.2**	**65.9**	**67.0**	**70.3**	**63.6**	**63.0**
BiEnc-F	BB	ChatGPT+CN	ChatGPT+CN	82.1	63.0	61.2	65.3	60.2	59.9
BiEnc-F	RB	ChatGPT+CN	ChatGPT+CN	81.5	62.3	60.8	65.0	59.6	61.3
BiEnc-F	RL	ChatGPT+CN	ChatGPT+CN	85.6	65.1	65.9	69.2	63.1	62.8

Table 1: Results for commonsense property prediction in terms of F1 (%). Results marked with * were taken from Gajbhiye et al. (2022). MSCG corresponds to the training set from Gajbhiye et al. (2022); ChatGPT and CN (ConceptNet) refer to the training sets that were described in Section 3.1. We evaluate using BERT-base-uncased (BB), BERT-large-uncased (BL), RoBERTa-base (RB) and RoBERTa-large (RL).

Conclusion

- Diverse concepts representations, allowing diverse clustering options
- Static concept vectors usable for tasks as Ultra fine entity typing and ontology complitions

- Try using larger models
- Modeling the relations between concepts for entity linking and retrieval tasks
- Assessing the importance and relevance of a concept/sentence in relation to the document's topic

Tank You !

Any Question ?