Extracting Polymer Nanocomposite Samples from Full-Length Documents

Ghazal Khalighinejad, *Defne Circi, Cate Brinson, Bhuwan Dhingra*
Duke University

INTRODUCTION

Research articles on PNC are rich sources of information about sample compositions, crucial for understanding their diverse properties. The extraction of structured data from these articles is vital, yet challenging due to the dispersed nature of this data across texts, figures, and tables.

Contribution

- We develop PNCExtract, the first benchmark aimed at extracting sample lists from full-length PNC articles, emphasizing the detailing of unique N-ary relations.
- We create a novel evaluation metric for assessing the performance of state-of-the-art language models on PNCExtract.
- We evaluate various LLMs in different scenarios, including self-consistency checks and condensation of papers, to minimize hallucination and improve accuracy.
- Our findings reveal that even advanced LLMs fail to identify more than 50% of the samples, highlighting the need for further research in this area.

Enhancing data accessibility by extracting N-ary tuples from PNC articles, each tuple representing a unique sample.

End-to-End Prompt:

Instruction: *Identify the matrix, filler, and composition.*

JSON Template:

Matrix:
Filler:
Filler Mass:
Filler Volume:

Article:

LLM

PNC Sample #1:

Matrix: Poly(vinylidene fluoride)
Filler: Calcium copper titanate
Filler Mass: null
Filler Volume: 1.01 vol.%

⋮

PNC Sample #N:

Matrix: Poly(vinylidene fluoride)
Filler: Calcium copper titanate
Filler Mass: null
Filler Volume: 1.2 vol.%

RESULTS

- GPT-4 along with self-consistency to reduce hallucination outperforms all the other language models.
- But even GPT-4 fails to capture more than 50% of the PNC samples in the articles.

Model	Strict			Partial		
	P	R	F_1	P	R	F_1
Condensed Papers						
LLaMA2 C	21.7	0.6	1.2	60.0	1.5	3.0
Vicuna	5.8	2.6	3.6	49.9	19.5	28.1
Vicuna-16k	17.7	5.9	8.9	60.4	19.9	29.9
LongChat	6.6	3.5	4.6	47.3	24.4	32.2
GPT-4	43.6	32.0	36.9	64.5	**47.7**	54.8
Full Papers						
Vicuna-16k	18.4	1.5	2.7	65.7	4.6	8.5
LongChat	5.4	4.2	4.7	36.6	29.6	32.7
GPT-4	44.8	30.2	36.0	64.9	43.8	52.3
GPT-4 (NR)	28.4	**37.2**	32.2	-	-	-
GPT-4 + SC	**51.6**	31.1	**38.8**	**73.5**	43.8	**54.9**

- All the models perform better when the input article is condensed.