Zero-shot Medical Entity Retrieval without Annotation: Learning From Rich Knowledge Graph Semantics

Amazon AI, accepted to Findings, EMNLP 2021

Backgrounds

- Entity Retrieval is the task of linking mentions of named entities to concepts in a curated knowledge graph

- It allows medical researchers and clinicians to search medical literature easily using standardized codes and terms to improve patient care.

Problem Definition

- It is difficult to adapt quickly enough to those newly appeared medical conditions and drug treatments under a public health crisis

- Hence, a robust medical entity retrieval system is expected to have decent performance in a zero-shot scenarios

- Zero-shot retrieval is challenging due to the nature of medical domain: large numbers of ambiguous terms, acronyms and synonymous terms

Related works in entity retrieval

- (2005) String matching methods such as exact match, approximate match
- (2020) weighted keyword match such as BM25

=> Can be used as zero-shot, but difficult to handle synonyms and paraphrase with large surface form differences

- (2019) large scale pretrained motel such as Clinical BERT and BioBert

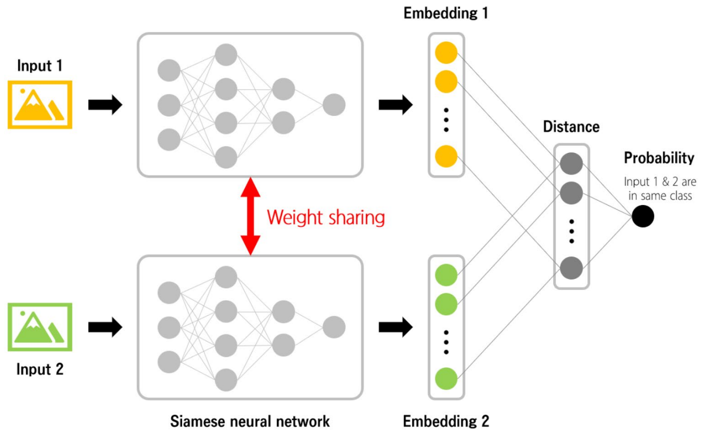
⇒Need fine-tuning process for final use

Contribution

- Proposing a framework which allows the information in medicalcal KGs to be incorporated into zero-shot entity retrieval models
- Applying the framework to major medical ontologies to show the effectiveness of the framework
- Showing that the proposed framework can be easily plugged into an existing supervised approach

Model Architecture

- Siamese architecture



Inner product: S(m,c) = <T(m), T(c)>

Experiment

- Learning by finding very similar or closely related textual descriptions and use them to construct (m,c) pairs

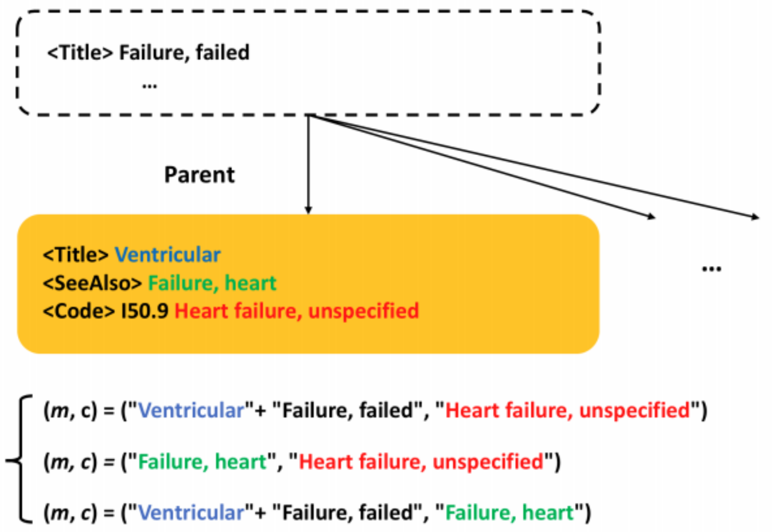


Figure 1: ICD-10 synonym-based task defined at an example node

KG	Task Type	Train	Dev
ICD-10	syn	113K	28K
	graph	33K	8K
SNOMED	syn	1.4M	374K
	graph	955K	238K
UMLS	syn	27M	7M
	graph	7M	2M
Comb (*by down-sampling*)		198K	488K

Table 1: **Task Description**: Number of (m, c) pairs in train and dev for all tasks.

Dataset	Split	KG	Test size
MedM.	-	UMLS	66,572
COMETA	SG	SNOMED	4,350
	SS		4,369
	ZG		3,995
	ZS		4,283
3DNotes	ICD	ICD-10	5,742
	SN	SNOMED	7,521

Table 2: **Test Set Size.**

Results

Dataset	Split	KG	BM25	Clinical BERT	Siamese architecture trained with KG learning tasks (**ours**) ICD-10 Syn	ICD-10 Graph	SNOMED Syn	SNOMED Graph	UMLS Syn	UMLS Graph	Comb
MedM.	-	UMLS	.04(.17)	.10(.30)	.31(.58)	.31(.56)	.32(.55)	.33(.61)	**.32(.53)**	**.30(.57)**	**.32(.60)**
COMETA	SG	SNOMED	.02(.10)	.01(.06)	.30(.52)	.30(.48)	**.43(.65)**	**.37(.58)**	.33(.50)	.32(.54)	**.37(.58)**
	SS		.02(.11)	.01(.06)	.28(.51)	.28(.47)	**.41(.62)**	**.36(.56)**	.31(.48)	.31(.52)	**.35(.56)**
	ZG		.02(.12)	.01(.07)	.32(.57)	.32(.54)	**.47(.71)**	**.39(.61)**	.36(.55)	.33(.57)	**.40(.62)**
	ZS		.02(.10)	.01(.07)	.30(.52)	.29(.47)	**.40(.64)**	**.35(.57)**	.31(.49)	.29(.53)	**.35(.57)**
3DNotes	ICD	ICD-10	.05(.22)	.11(.17)	**.28(.54)**	**.23(.46)**	.20(.45)	.20(.52)	.18(.39)	.21(.53)	**.30(.54)**
	SN	SNOMED	.07(.20)	.01(.05)	.20(.50)	.18(.45)	**.38(.63)**	**.25(.61)**	.25(.49)	.29(.55)	**.34(.59)**

Table 3: Retrieval performance R@1(25). Siamese architecture trained with our tasks are shown to significantly outperform benchmarks. Evaluation for *zero-shot on mentions only* is highlighted in **bold** the rest belongs to *zero-shot on mentions and concepts*. The former enjoys a bigger gain as expected.

Mention	Gold Concept	Syn	Graph
shortness of breath	dyspnea (finding)	✓	✗
GI hemorrhage	gastrointestinal hemorrhage (disorder)	✓	✗
coronary structure	coronary artery (body structure)	✗	✓
heart	heart structure (body structure)	✗	✓

Table 4: Prediction error of the model trained with SNOMED tasks evaluated on 3DNotes-SN.

- Synonym based task: Learning by synonyms
- Graph-based task: Learning by connections

Conclusion

- Using a medical KG to enable entity retrieval model to mine rich semantics from the KG

- The model can be used as an auxiliary task when annotations are available