SurfCon: Synonym Discovery on Privacy-Aware Clinical Data

Zhen Wang*, Xiang Yue*, Soheil Moosavinasab†, Yungui Huang†, Simon Lin†, Huan Sun*

*The Ohio State University, †Abigail Wexner Research Institute at Nationwide Children's Hospital

THE OHIO STATE UNIVERSITY

NATIONWIDE CHILDREN'S — When your child needs a hospital, everything matters.

Introduction

Synonym Discovery in Clinical Data.

- Clinical texts in Electronic Medical Records (EMRs) contain lots of synonyms.

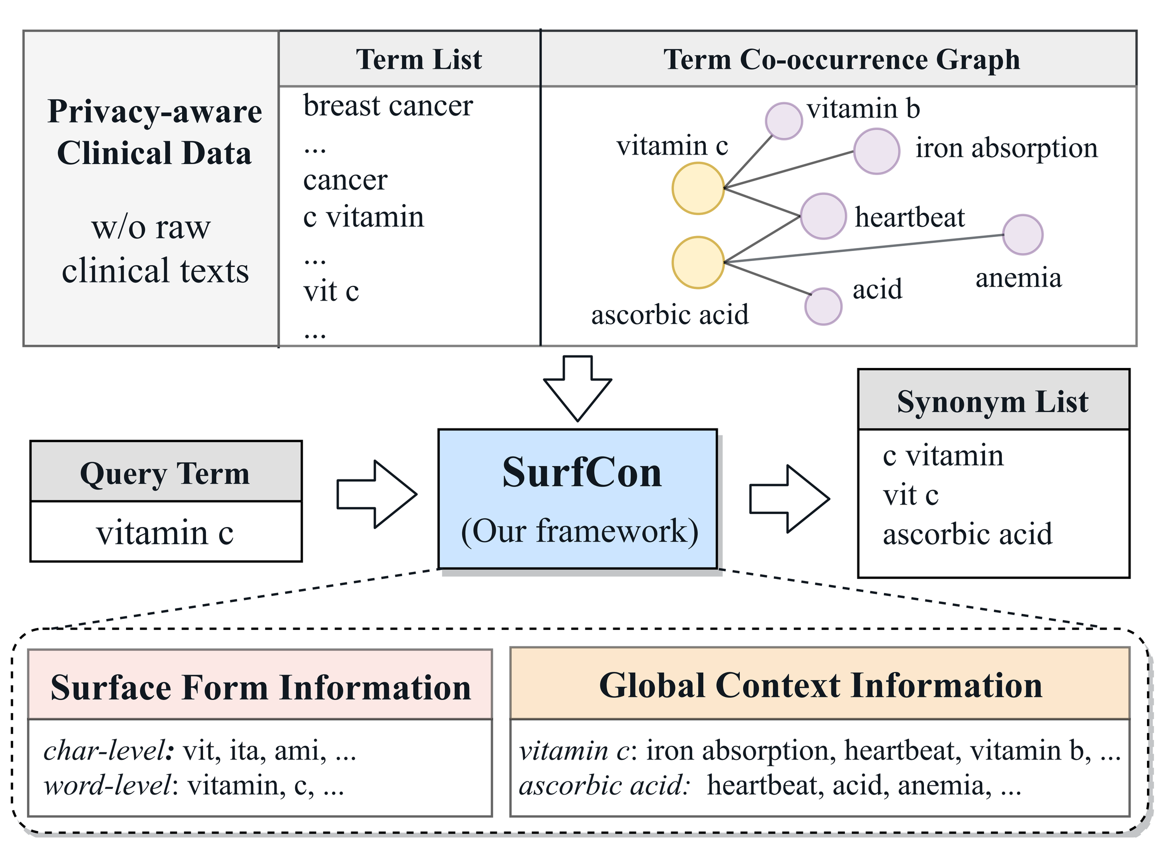
Medical Term	Synonyms
vitamin c	vit c; c vitmin; ascorbic acid; …
copper deficiency	copper low; copper decreased; hypocupremia; …
large kidney	enlarged kidneys; nephromegaly; renomegaly; …
hiv disease	hiv infection; human immunodeficiency virus; …

Privacy-Aware Clinical Data

- Due to the privacy concern for patients, large-scale clinical text corpora are rarely publicly available.
- Medical terms and their aggregated co-occurrence counts extracted from raw clinical texts are becoming a popular (although not perfect) substitute for raw clinical texts for the research community to study EMR data.
- In this work, we refer to the given set of **medical terms and their co-occurrence statistics** in a clinical text corpus as privacy-aware clinical data, and investigate synonym discovery task on such data.

Problem Formulation

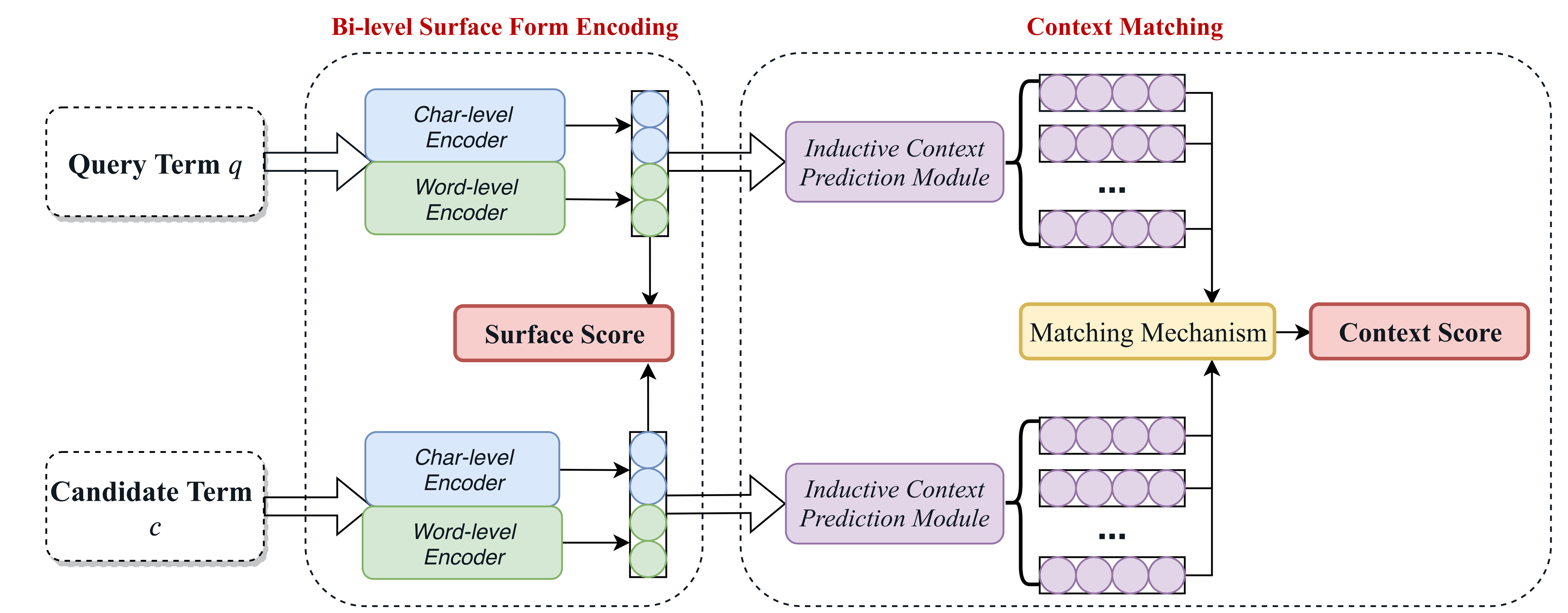
- *Given a set of terms extracted from clinical texts as well as their global co-occurrence graph, recommend a list of synonyms for a query term.*
- We observe and propose to leverage two important types of information: surface form and global contexts as follows.



Challenges

- How to balance Surface Form and Global Context information?
  - Previous methods (e.g., siamese recurrent networks [2]) are good at capturing string-level association but rarely model the semantic meaning together.
- How to handle the In-the-Vocabulary (InV) and Out-Of-Vocabulary (OOV) query terms at the same time?
  - OOV terms do not have any global contexts in the graph that traditional graph representation learning methods (e.g., [3]) cannot deal with.

SurfCon Framework



Framework overview

For each query term, a list of candidate terms will be ranked based on both the surface and context scores.

Methodology

- **Bi-level Surface Form Encoding**
  - Combining character-level and word-level information to measure the similarity in surface forms.
- **Inductive Context Prediction Module**
  - Given a co-occurrence graph, the context predictor is pre-trained to predict global contexts of terms by estimating how likely term u_j appears in the context of u_i by the following conditional probability:

$$p(u_j|u_i) = \frac{exp(\nu_{u_j}^T \cdot s_{u_i})}{\sum_{k=1}^{|V|} exp(\nu_{u_k}^T \cdot s_{u_i})} \quad (1)$$

where s_{u_i} is the surface form representation from previous component and v_{u_j} is context embedding vector.

- **Context Matching Mechanism**
  - Measuring semantic similarity by context association.
  - Dynamic matching: weighing query term's contexts by their matching degree with candidate term's contexts, and vice versa.

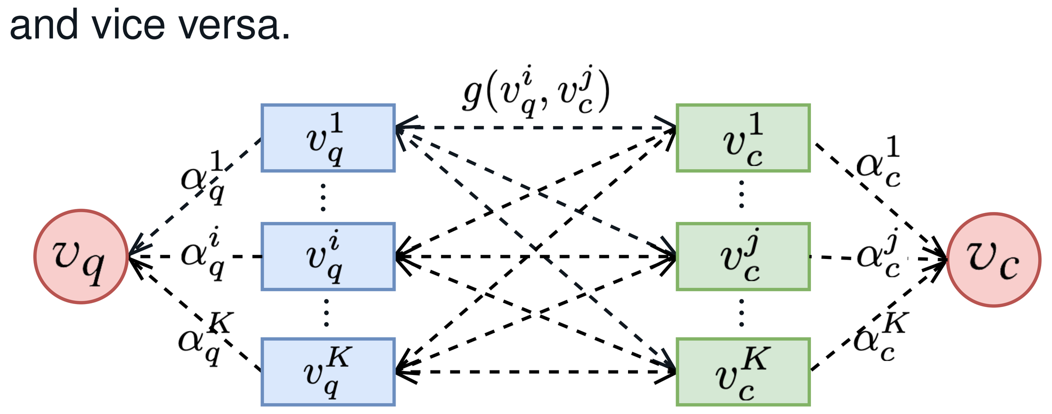


Fig. 4: Dynamic Context Matching Mechanism

$$\alpha_q^i = \frac{e^{match[v_q^i, \Phi(c)]}}{\sum_{k=1}^{K} e^{match[v_q^k, \Phi(c)]}} \quad (2)$$

$$match[v_q^i, \Phi(c)] = Pooling[g(v_q^i, v_c^1), ..., g(v_q^i, v_c^K)] \quad (3)$$

Experiments

Experimental Setup

- **Datasets.** Two publicly available medical term-term co-occurrence graphs extracted by Finlayson et al. [1] from 20 million clinical notes.
- **Synonym labels.** Grouping medical terms under a same concept from UMLS.
- **Testing scenarios.** InV (query) testing and OOV (query) testing as well as a subset, Dissim, with string-dissimilar synonyms for both InV and OOV sets.

		1-day dataset	All-day dataset
# Nodes		52,804	43,406
# Edges		16,197,319	50,134,332
Average # Degrees		613.5	2310.0
# Train Terms		9,451	7,021
# Dev Terms		960	726
# InV Test Terms	All	960	726
	Dissim	175	152
# OOV Test Terms	All	2,000	2,000
	Dissim	809	841

Fig. 5: Datasets Statistics

- **Evaluation.** Mean Average Precision (MAP)

Main Results

Method Category	Methods	InV Test All	InV Test Dissim	OOV Test All	OOV Test Dissim
Surface form based	CHARAGRAM [5]	0.8507	0.5504	0.7609	0.5142
Global context based	DPE-NoP [4]	0.6107	0.4855	-	-
Hybrid (surface+context)	Planetoid [6]	0.8514	0.5612	0.731	0.4714
Our model and variants	SurfCon (Surf-Only)	0.9053	0.6145	0.8228	0.5829
	SurfCon (Static)	0.9151	0.6542	0.8285	0.5933
	SurfCon	**0.9176**	**0.6821**	**0.8301**	**0.6009**

Experiments (Cont'd)

Case Studies

Query Term	"unable to vocalize" (InV)	"marijuana" (OOV)
SurfCon Top Ranked Candidates	"does not vocalize" "aphonia" "loss of voice" "vocalization" **"unable to phonate"**	**"marijuana abuse"** **"cannabis"** "cannabis use" "marijuana smoking" "narcotic"
Labeled Synonym Set	"unable to phonate"	"cannabis" "marijuana abuse" "marihuana abuse"

Bold terms are synonyms in our labeled set while underlined terms are new synonyms that our model discovers.

For more results with different settings and parameter sensitivity, please refer to our paper.

Takeaways

- By leveraging surface form and global context information, SurfCon model can discover synonyms from privacy-aware clinical data effectively.
- Medical term co-occurrence graph as privacy-aware clinical data can preserve privacy effectively and be used for solving many data mining tasks.
- Interesting future work includes extending the framework to other structured knowledge mining problems or exploring more tasks on the privacy-aware clinical data.

Contact & Code

Contact:
wang.9215@osu.edu
Arxiv:
https://arxiv.org/abs/1906.09285
Code:
https://github.com/zhenwang9102/SurfCon

References

[1] S. G. Finlayson, P. LePendu, and N. H. Shah. "Building the graph of medicine from millions of clinical narratives". In: *Scientific data* 1 (2014), p. 140032.

[2] J. Mueller and A. Thyagarajan. "Siamese Recurrent Architectures for Learning Sentence Similarity." In: *AAAI*. 2016.

[3] B. Perozzi, R. Al-Rfou, and S. Skiena. "Deepwalk: Online learning of social representations". In: *KDD*. 2014.

[4] M. Qu, X. Ren, and J. Han. "Automatic synonym discovery with knowledge bases". In: *KDD*. 2017.

[5] J. Wieting et al. "Charagram: Embedding words and sentences via character n-grams". In: *EMNLP*. 2016.

[6] Z. Yang, W. W. Cohen, and R. Salakhutdinov. "Revisiting semi-supervised learning with graph embeddings". In: *ICML*. 2016.