CS598 DLH Project, Summer 2023

Project Presentation by Mitch & Sathish

{sbrama2, mm109}@illinois.edu

Group ID: 195
Paper ID: 175

.

Paper :

SurfCon: Synonym Discovery on Privacy-Aware Clinical Data,

Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery

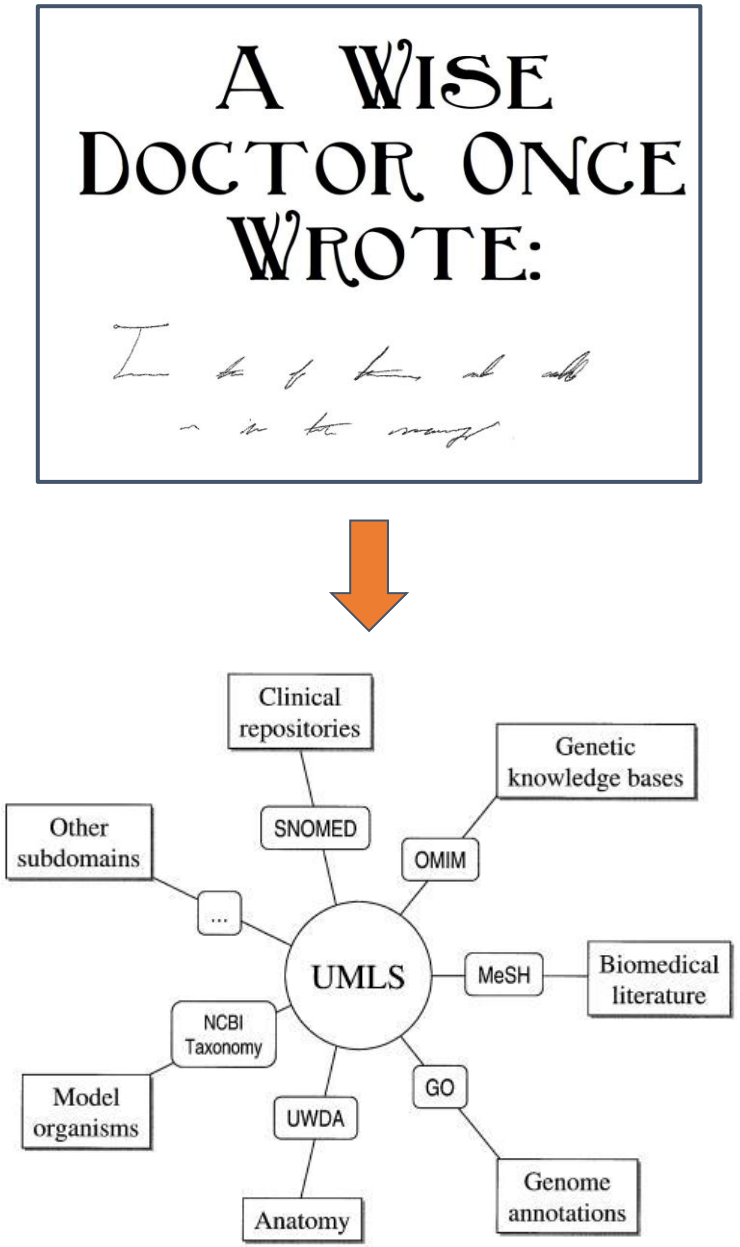
Link to Paper : https://dl.acm.org/doi/pdf/10.1145/3292500.3330894

Problem

- Automatically discovering synonyms (e.g., "c vitamin", "vit c", "ascorbic acid") or misspelled variations (e.g. "viatmin c") can help find valuable information such as patient-clinical interactions and disease treatment outcomes.
- Documentation of medical terms in unstructured fields is error prone and often captured in shorthand
- Current taxonomies (e.g. UMLS) help to group medical concepts, but do not bridge the gap to this noisy documentation

Solution

- Utilize co-occurrence data of medical terms in clinical notes to build a robust synonym generation model
- Consider visual similarity of terms and semantic similarity



Approach

SurfCon for Synonym Generation

Build Surface Form and Global Context representation of medical terms using Character and word embeddings and co-frequency data from Clinical notes

Two information categories in privacy-aware clinical data:

- Surface form information of a medical term
- Global contexts from the given co-occurrence graph.

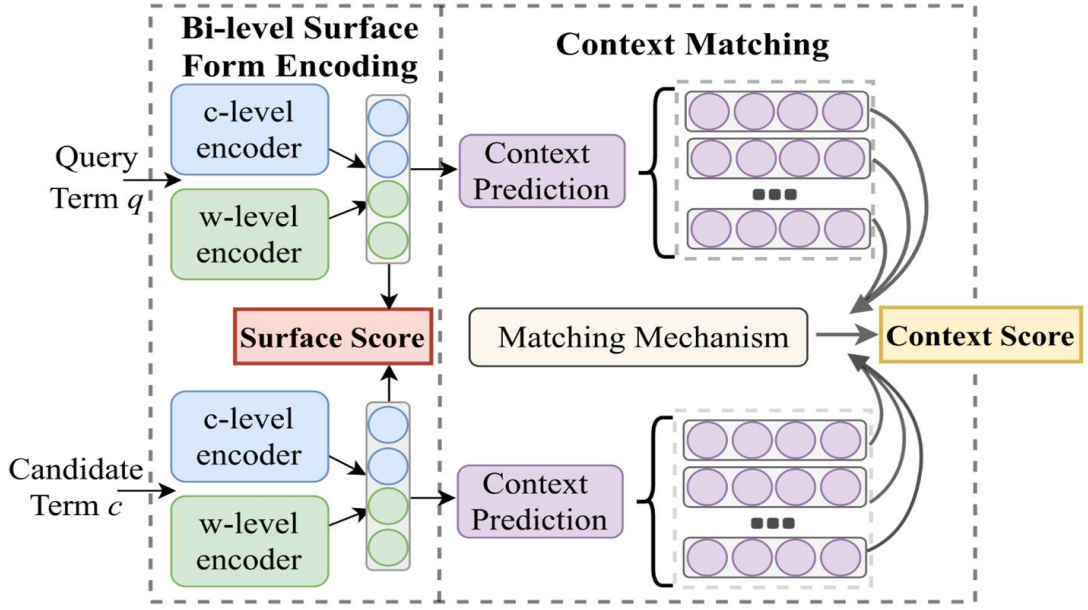


Figure 2 : SurfCon Architecture

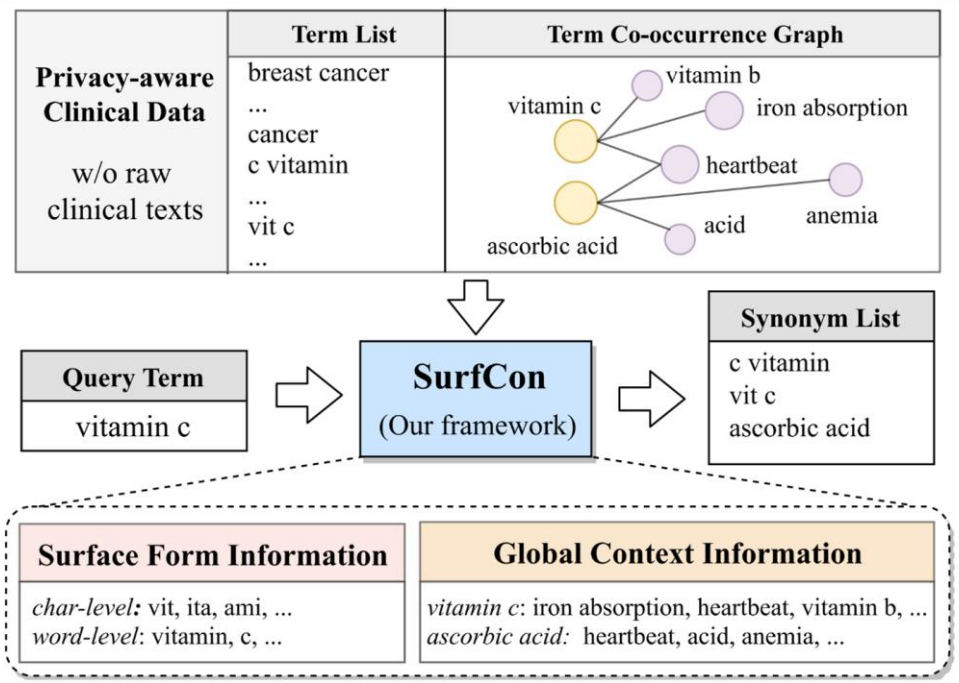


Figure 1 : SurfCon Approach

Bi-level surface form encoding component:

- Model the similarity between two terms at the surface form level
- Exploits both character and word-level information to encode a medical term into a vector.
- Computes a surface score of two terms based on similarity
- Works well for detecting synonyms similar on surface form only.

Context matching component:

- Aims to discover synonyms that are not similar in surface form
- Uses term's context to represent semantic meaning
- Utilizes the co-occurrence graph as a representation of a term's global context
- Generates context semantic vector for the candidate term wrt query term
- Using training set of term's contexts, seeks to predict the global context of a novel term

Results

Synonym Generator:
Model successfully handles shorthand and misspellings to produce synonymous terms

Handles Out of Vocabulary Terms:
Model performs well on terms that were not present in the frequency graph

Improvement over existing methods:
Model shows marked improvement from surface form, global context, and existing hybrid models.
SurfCon is also robust to dissimilar terms, showing significant improvement against other methods

Method Category	Methods	1-day Dataset				
		Dev	InV Test		OOV Test	
			All	Dissim	All	Dissim
Surface form based methods	CharNgram [13]	0.8755	0.8473	0.4657	0.7427	0.4131
	CHARAGRAM [40]	0.8705	0.8507	0.5504	0.7609	0.5142
	SRN [25]	0.8886	0.8565	0.5102	0.7241	0.4341
Global context based methods	Word2vec [23]	0.3838	0.3748	0.3188	-	-
	LINE(2nd) [34]	0.4279	0.4301	0.3494	-	-
	DPE-NoP [30]	0.6222	0.6107	0.4855	-	-
Hybrid methods (surface+context)	Concept Space [37]	0.8094	0.8109	0.4690	-	-
	Planetoid [41]	0.8813	0.8514	0.5612	0.731	0.4714
Our model and variants	SurfCon (Surf-Only)	0.9160	0.9053	0.6145	0.8228	0.5829
	SurfCon (Static)	0.9242	0.9151	0.6542	0.8285	0.5933
	SurfCon	**0.9348**	**0.9176**	**0.6821**	**0.8301**	**0.6009**

```
Input your query (Press 'exit' to exit): epi tissue
Searching 52804 candidate terms by 18.916 seconds

Top ranking Terms:
1       epithelial tissue
2       osseous tissue
3       fibrofatty tissue
4       skin tissue
5       adventitial tissue
6       deep tissue
7       splenic tissue
8       periurethral tissue
9       lung tissue
10      brain tissue
```

```
Input your query (Press 'exit' to exit): cncer
Searching 52804 candidate terms by 22.216 seconds

Top ranking Terms:
1       cancer, liver
2       liver cancer
3       primary liver cancer
4       pulsatile liver
5       appearance of liver
6       colonic cancer
7       bile duct cancer
8       cecal cancer
9       cancer other
10      ca - cancer
```

Reproduction Study

Replaced CharNGram with FastText embeddings:

- Utilized pre-trained subword embeddings from FastText as a replacement for the 2-4 Gram embeddings with CharNGram

Implemented Data Loading and Preprocessing:

- Build data loading and transformation code for building input datasets and labels
- Implemented code to replicate data preprocessing measures including PPMI, subsampling, and data splitting

Performed Ablation Experiment:

- Executed SurfCon training without Context Matching to re-evaluate significance in Synonym generation

Performed reproduction of full SurfCon Architecture:

- Trained Surfcon Surface-form and Global context data, utilizing inductive global context prediction with dynamic context matching

Study Outcomes

SurfCon Innovative Approach Verified:

- Even with utilizing a different pre-trained subword embedding, we were able to verify the claim that SurfCon improved the Synonym generation task over existing methods

Context Matching Claim Questioned:

- Our model saw better results without the use of Context Matching for Test sets (including Dissimilar terms), calling into question the significance of this component on the task

Table 1 : Model evaluation in MAP between Paper Claims vs Our reproduction

Methods	Paper Claims			Our reproduction results		
	Dev	InV Test(All)	InV Test (Dissim)	Dev	InV Test(All)	InV Test(Dissim)
SurfCon (Surf-Only)	.9160	.9053	.6145	0.9465	0.9097	0.6757
SurfCon	.9348	.9176	.6821	0.921	.8766	.6176

Using Paper Model :

```
Input your query (Press 'exit' to exit): vitamin c
Searching 52804 candidate terms by 22.11 seconds

Top ranking Terms:
1       vitamin b
2       vitamin d3
3       vitamin b-12
4       vitamin b complex
5       cyanocobalamin
6       vitamin a
7       vitamin e
8       vitamin d2
9       vitamin b-6
10      vitamin b12
```

Using our model

```
Begin querying:

Input your query (Press 'exit' to exit): vitamin c
Searching 52804 candidate terms by 21.737 seconds

Top ranking Terms:
1       vitamin b
2       b vitamins
3       vitamin b12
4       vitamin b 12
5       c vitamin
6       vitamin b deficiency
7       vitamin b-12
8       b vitamin
9       vitamin b 1
10      vitamin b complex
```