The Ohio State University

Nationwide Children's
When your child needs a hospital, everything matters.

SurfCon: Synonym Discovery on Privacy-Aware Clinical Data

Zhen Wang

The Ohio State University

KDD 2019, August 6th

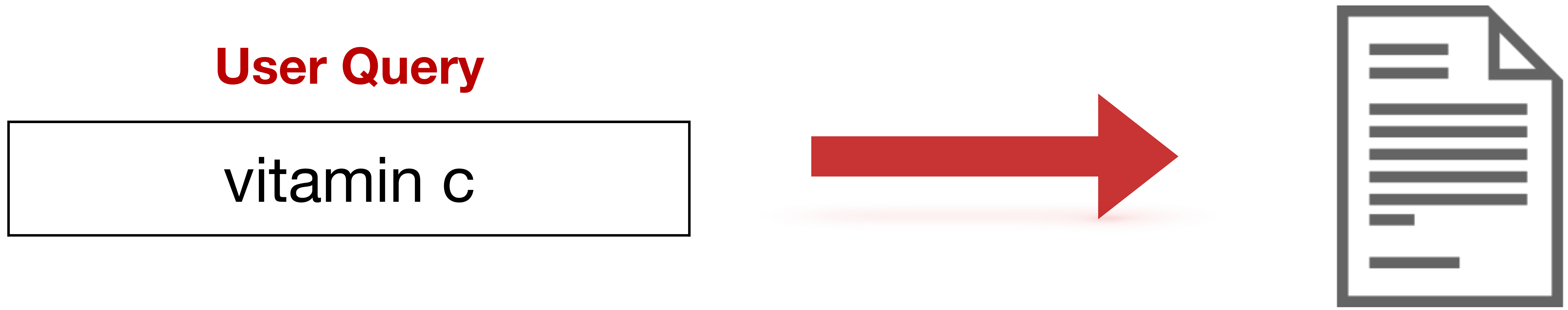
In collaboration with Xiang Yue (OSU), Soheil Moosavinasab (NCH), Yungui Huang (NCH), Simon Lin (NCH), Huan Sun (OSU)

Synonym Discovery in Clinical Data

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; …

Why Synonym Discovery?

One Application Scenario: Search



Why Synonym Discovery?

One Application Scenario: Search

User Query

vitamin c

User Query Expansion

vitamin c; ascorbic acid; vit c; c vitmin

More relevant documents

Existing Synonym Discovery Methods on Text Corpora

Text corpus with raw sentences

The **USA** is also known as **America**.
The **USA** (**America**) is a country of 50 states.
Illinois, which is also called **IL**, is a state in the US.
Michigan, also known as **MI**, consists of two peninsulas.

Examples from (Qu et al., KDD'17)

Concept Space [Wang et al., IJCAI'15]

DPE [Qu et al., KDD'17]

SynonymNet [Zhang et al., 2018]

…

Privacy Concerns in Clinical Data

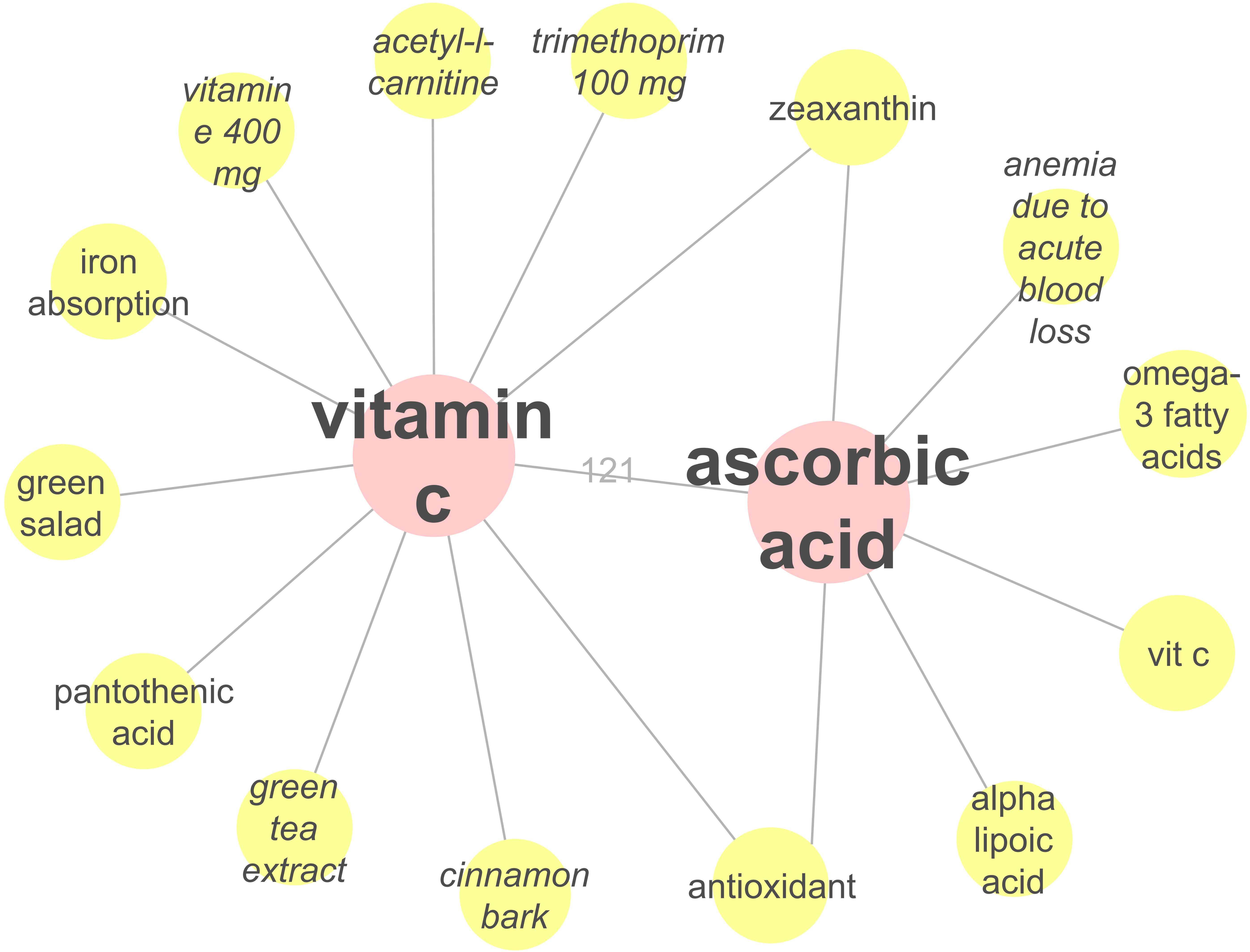
3. Echocardiogram on **DATE[Nov 6 2007] , showed ejection fraction of 55% , mild mitral insufficiency , and 1+ tricuspid insufficiency with mild pulmonary hypertension .
DERMOPLAST TOPICAL TP Q12H PRN Pain DOCUSATE SODIUM 100 MG PO BID PRN Constipation IBUPROFEN 400-600 MG PO Q6H PRN Pain
The patient is struggling to breathe at this time , and she is tachypneic , and she might have to be intubated right now but ; however , the patient 's family did not wish the patient to be intubated even after I explained to them that she could potentially die if she was not on a breathing machin but however , the patient 's family stressed to me again and wished that they do not want her mother to be on a breathing machine .
The patient had headache that was relieved only with oxycodone . A CT scan of the head showed microvascular ischemic changes . A followup MRI which also showed similar changes . This most likely due to her multiple myeloma with hyperviscosity .

Table 1: Examples of concepts (Problem, Treatment, and Test) from the i2b2 2010 corpus.

Examples from [Roberts, ClinicalNLP'16]

Raw clinical texts are rarely publicly available!

Privacy-Aware Clinical Data

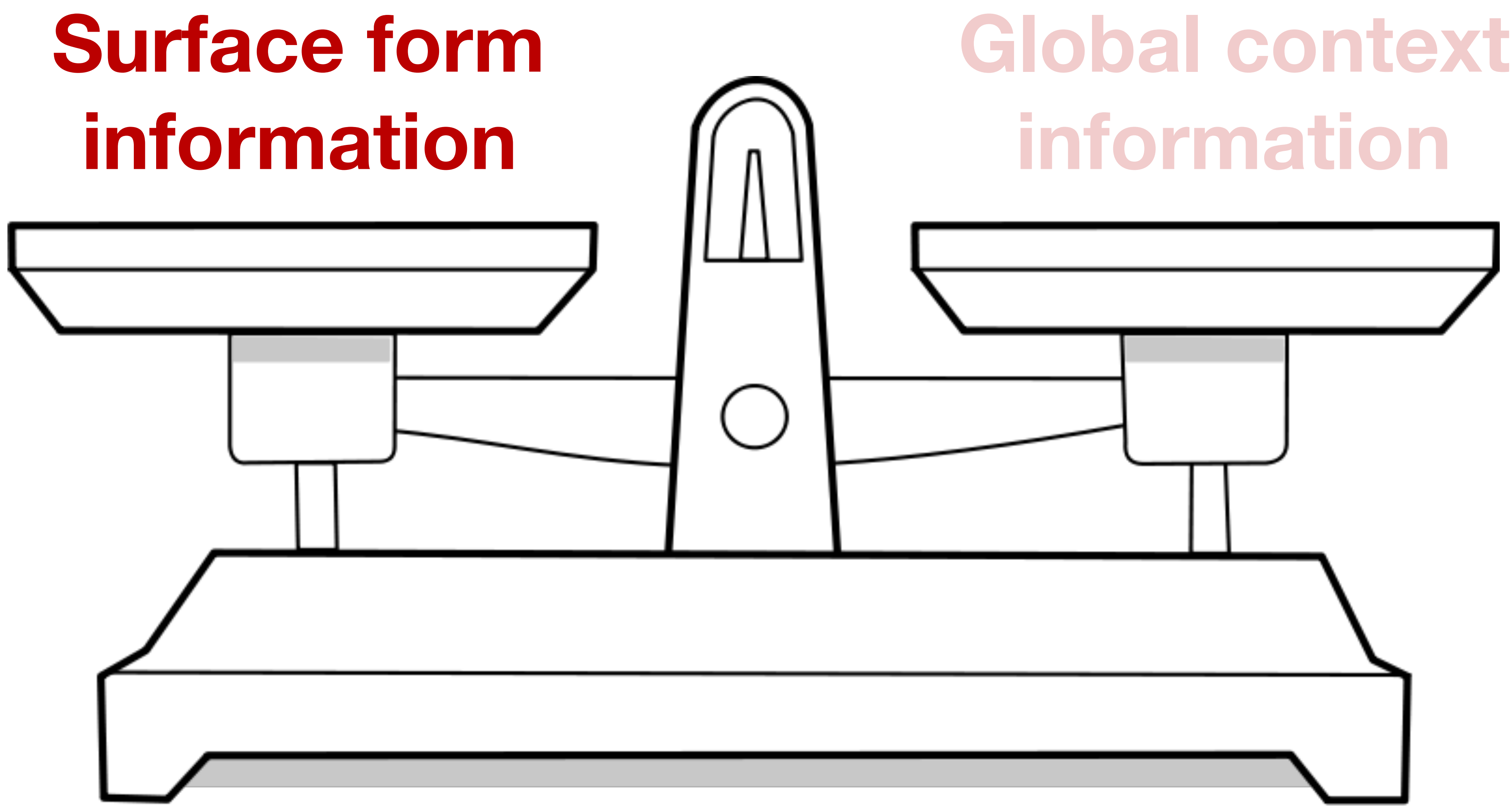


Medical term-term co-occurrence graph

Observations in Privacy-Aware Clinical Data

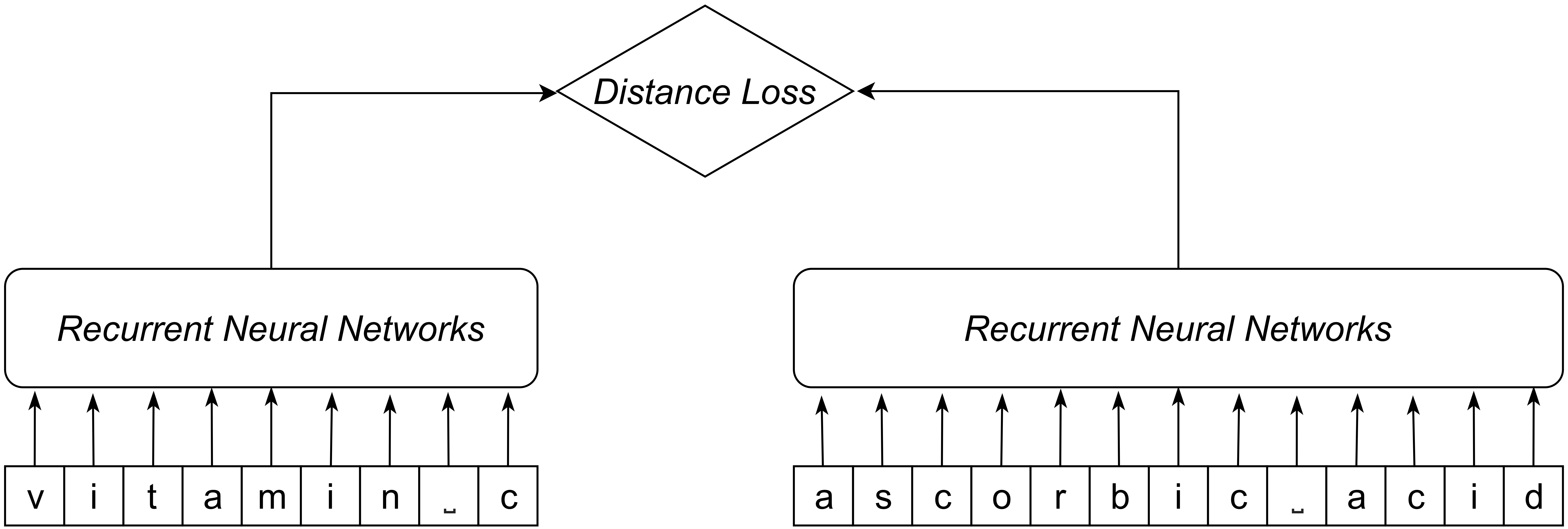
Observations in Privacy-Aware Clinical Data

Surface forms for "vitamin c"	
Character-level	*vit, ita, tam, min, …*
Word-level	*vitamin, c*



Modeling the surface form information

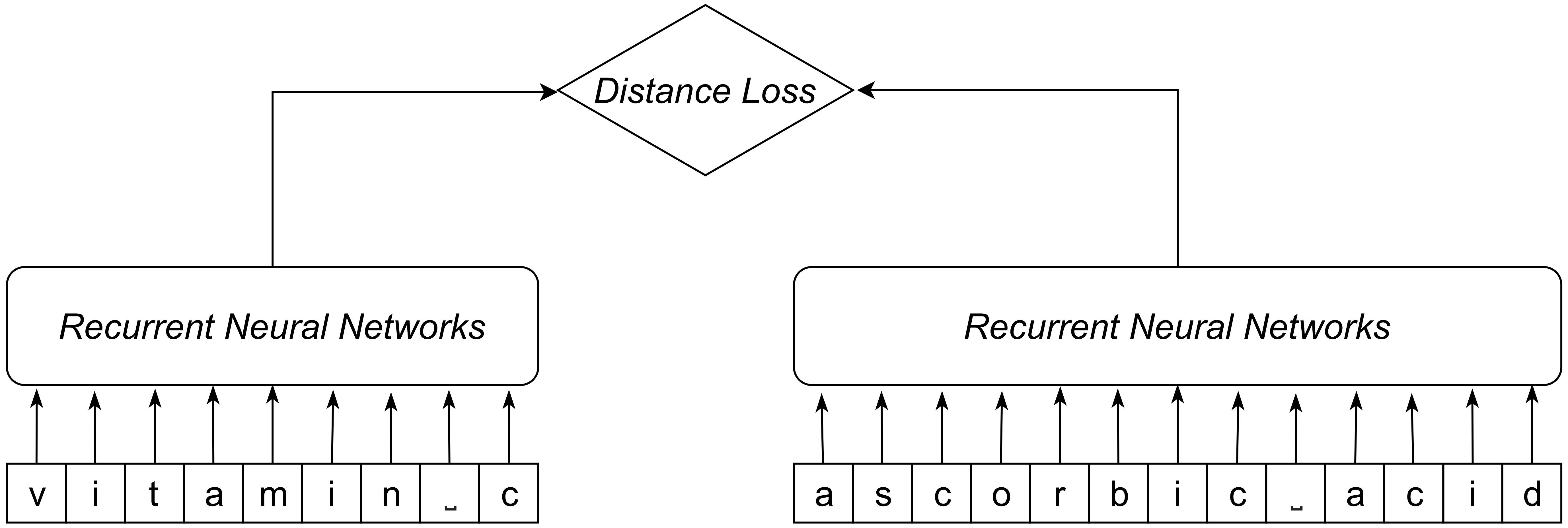
Siamese Recurrent Networks



[Mueller and Thyagarajan, AAAI’16]
[Neculoiu et al., 2016]

Good at capturing the string-level association

Modeling the surface form information



Challenge I

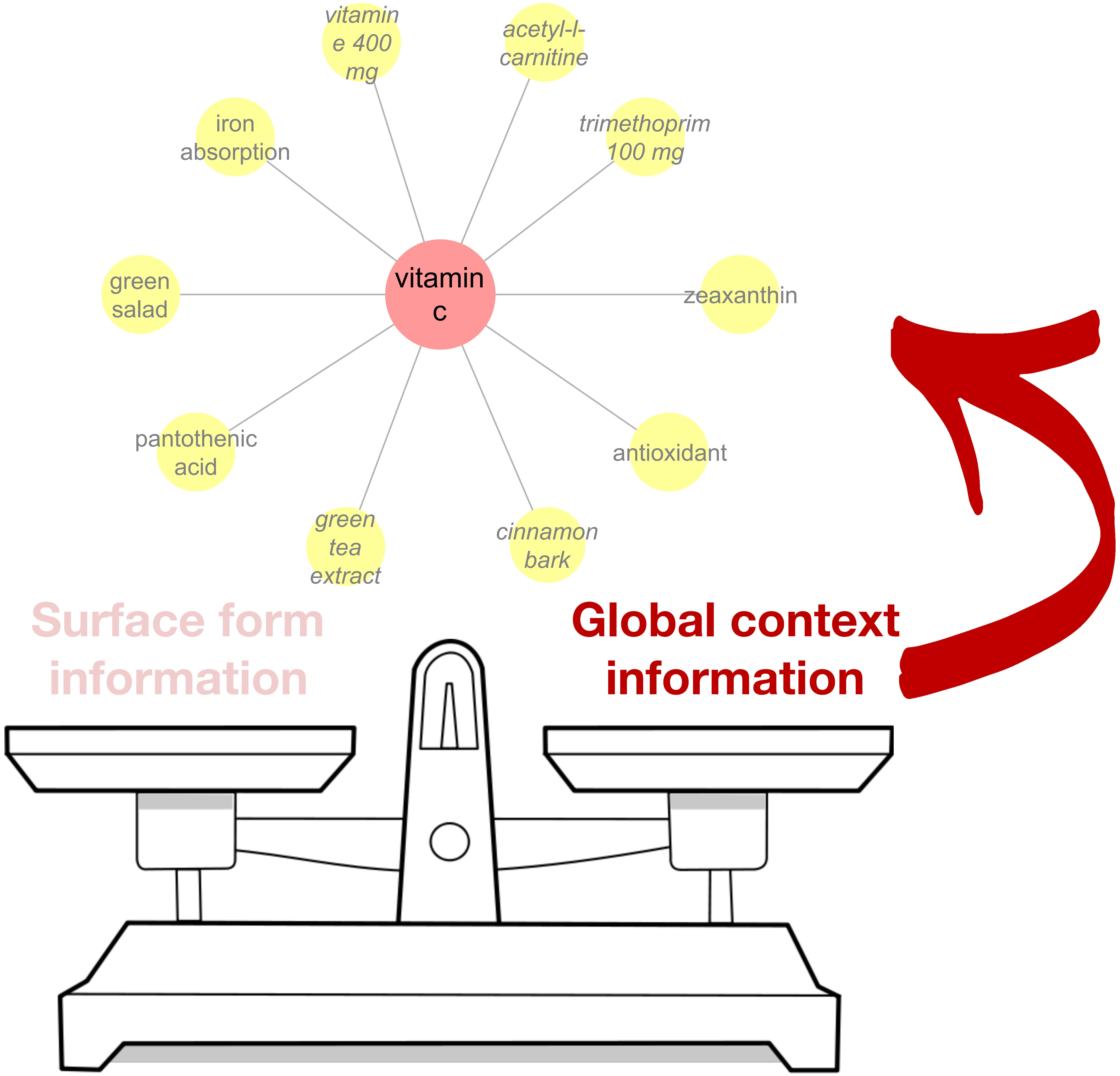
1. Similar in surface form with different meanings

- *hemostatic* (*stop bleeding*) *vs. homeostasis* (*stable inner environment*)

2. Similar in meaning with different surface forms

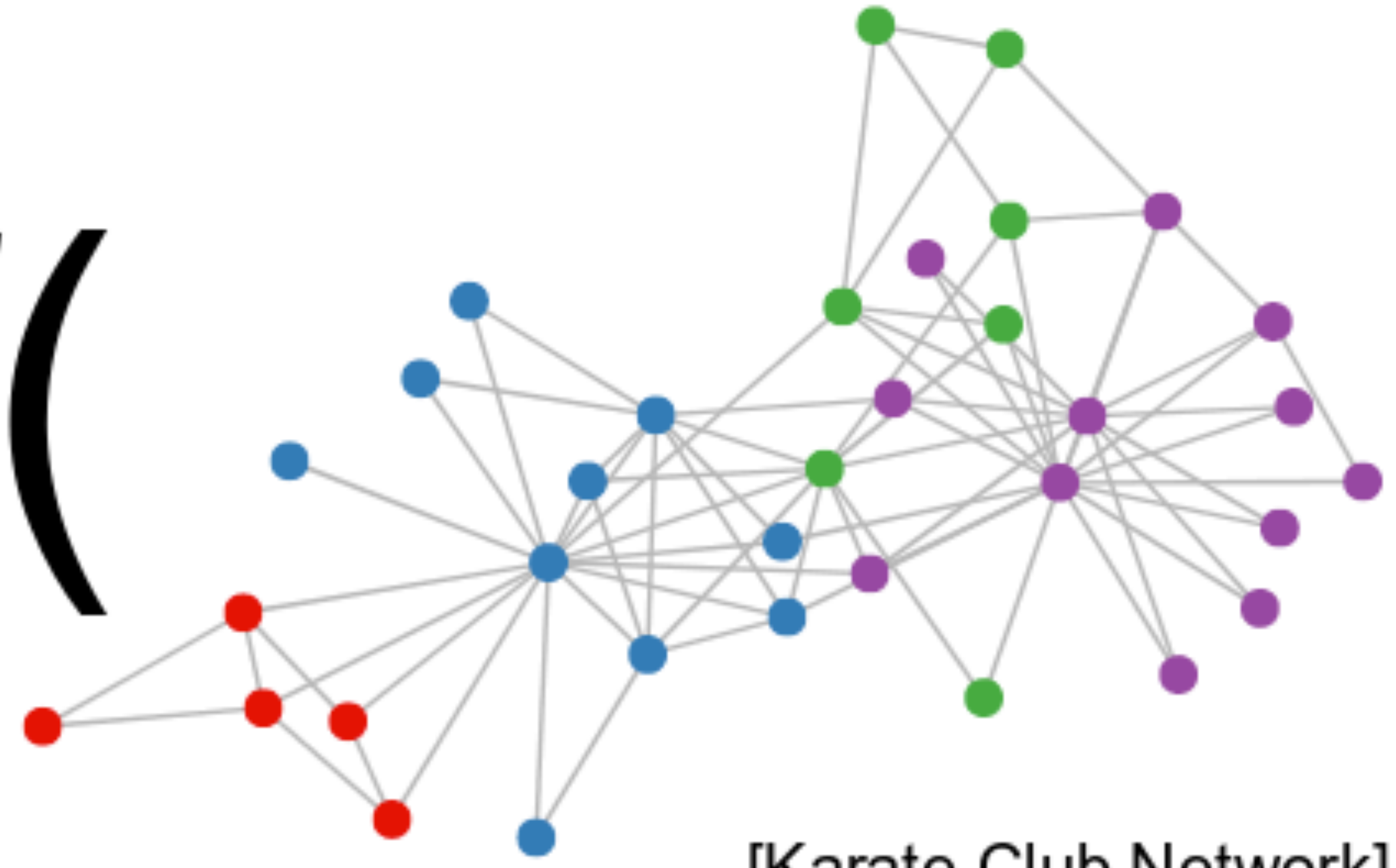
- *ascorbic acid vs. vitamin c*

Observations in Privacy-Aware Clinical Data



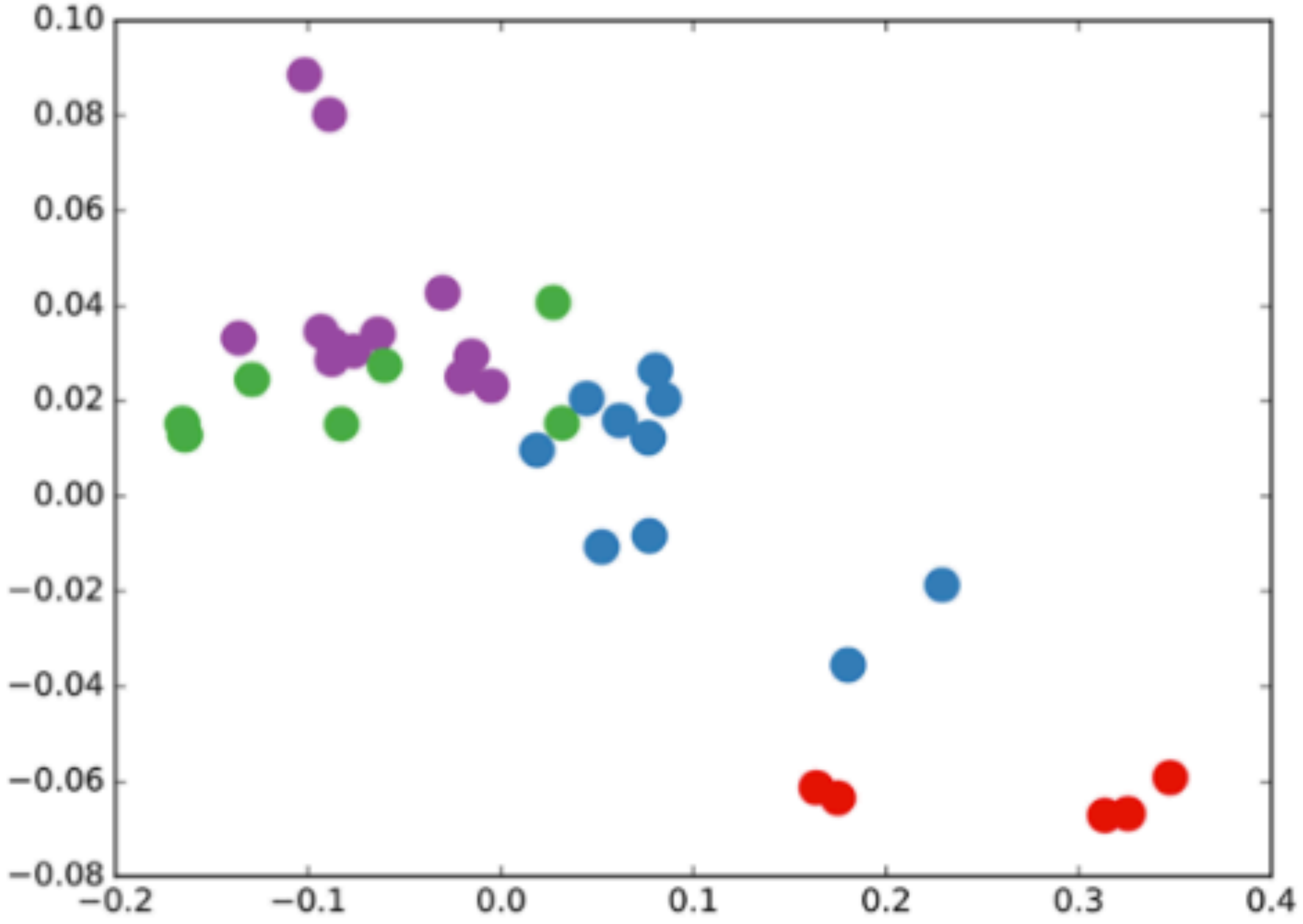
Modeling the global context information

Parameters initialized randomly

f() =

[Karate Club Network]

2-dim output per node

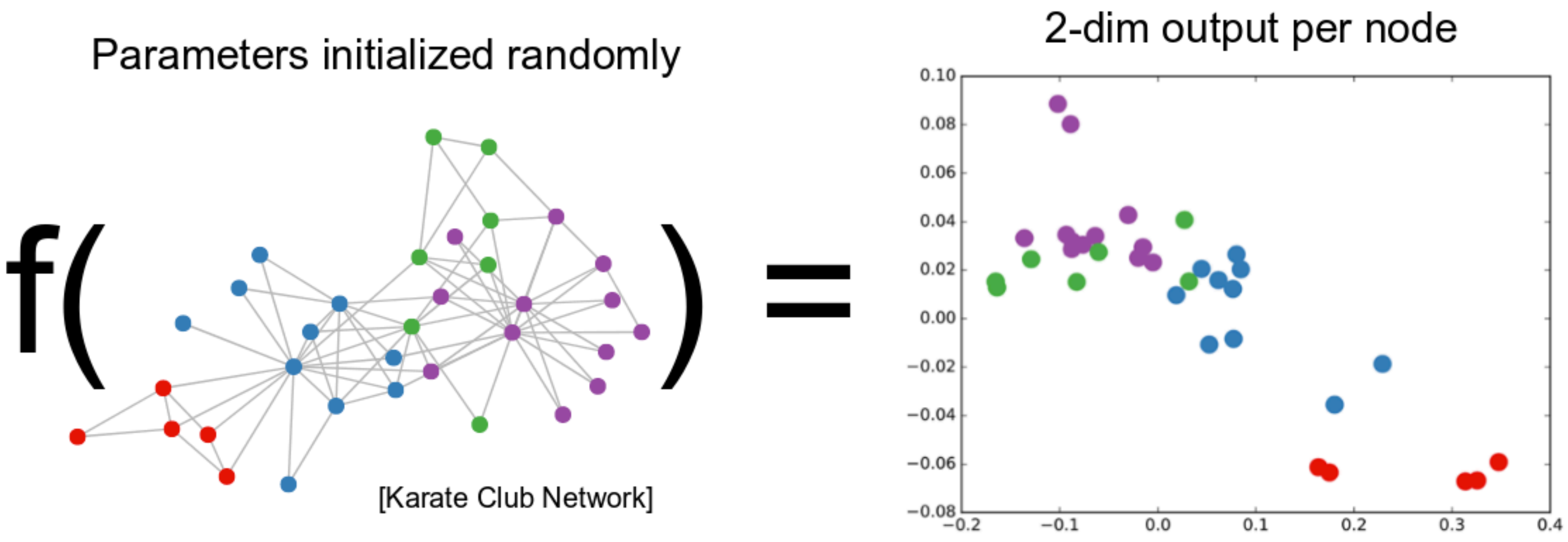


[Perozzi et al., KDD'14]
[Tang et al., WWW'15]
[Grover and Leskovec, KDD'16]

Learning semantic representations from graph structures

Figure from Siv's Blog
https://iamsiva11.github.io/graph-embeddings-2017-part1/

Modeling the global context information

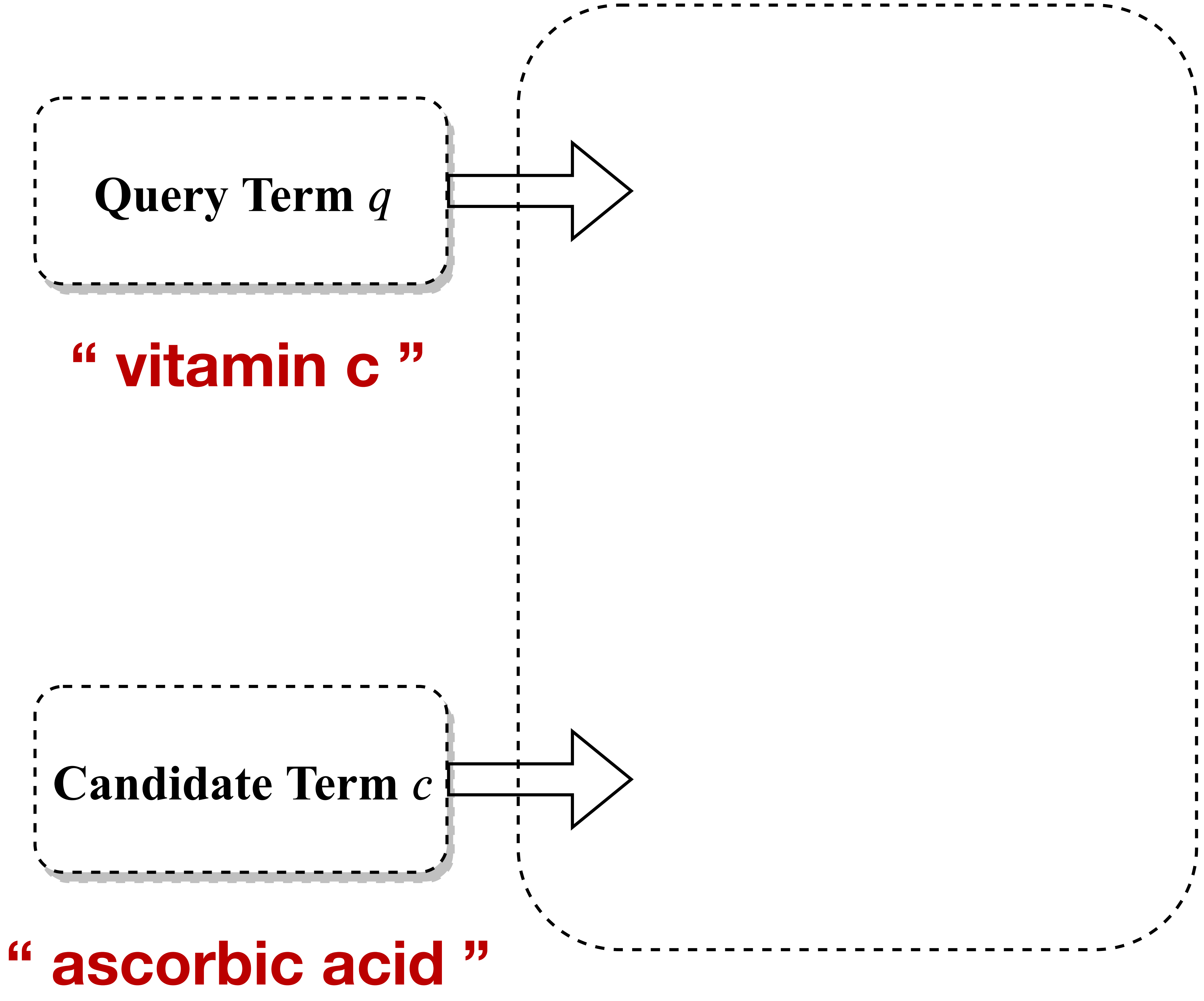


Challenge II

Traditional embeddings cannot deal with Out-Of-Vocabulary (OOV) Query terms

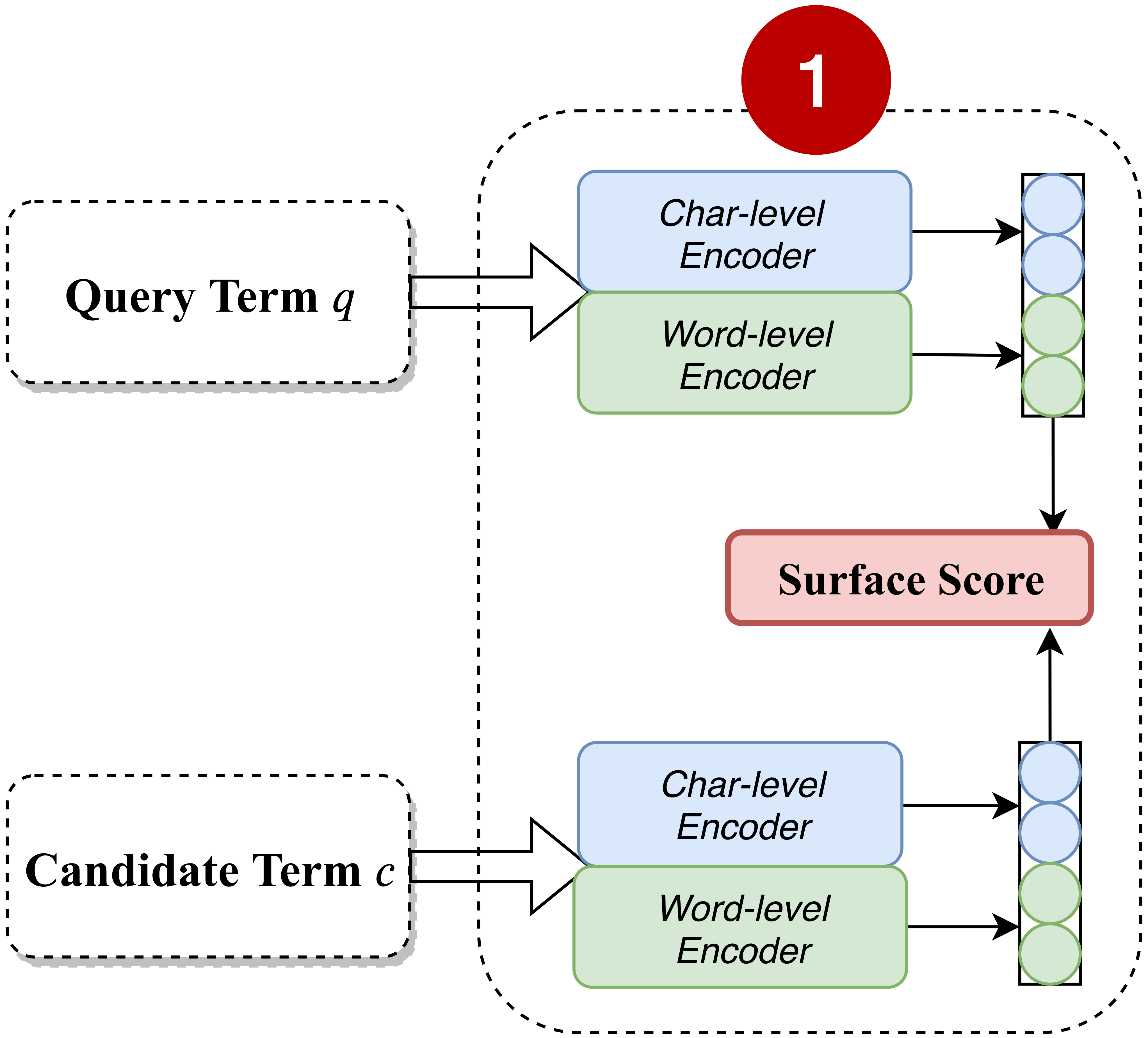
- *No global contexts available for OOVs in the graph*

SurfCon Framework: Surface form + Context

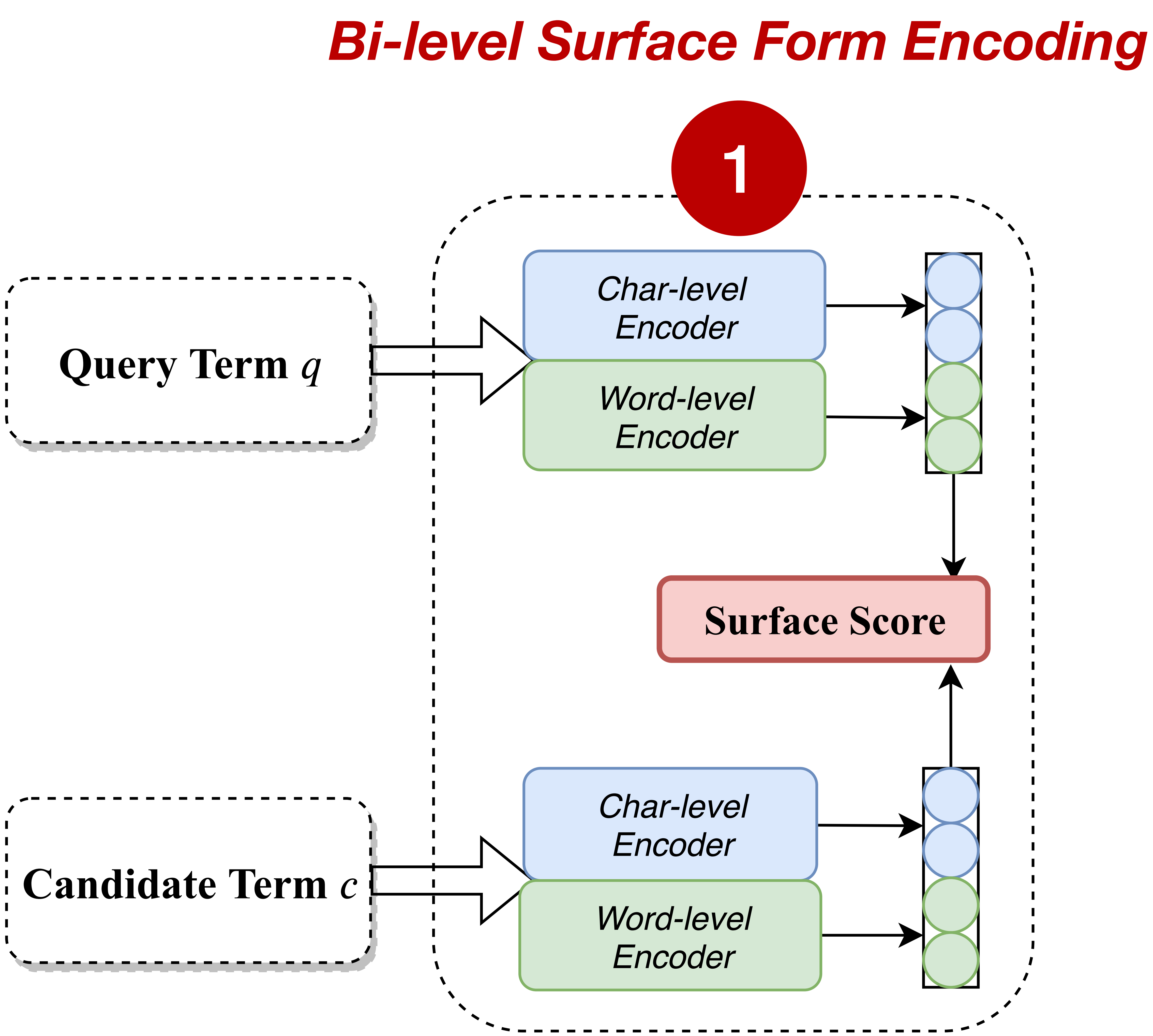


SurfCon Framework

Bi-level Surface Form Encoding



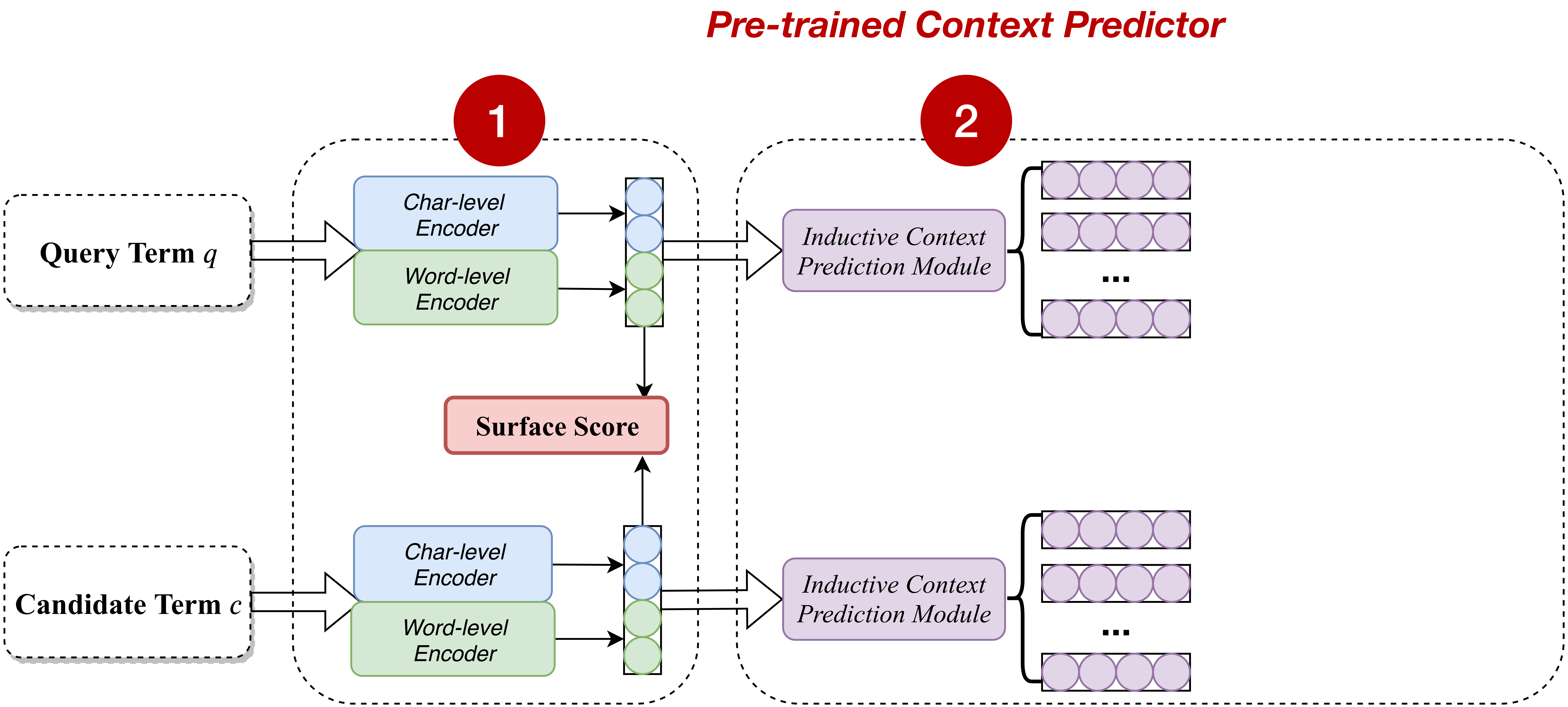
SurfCon Framework



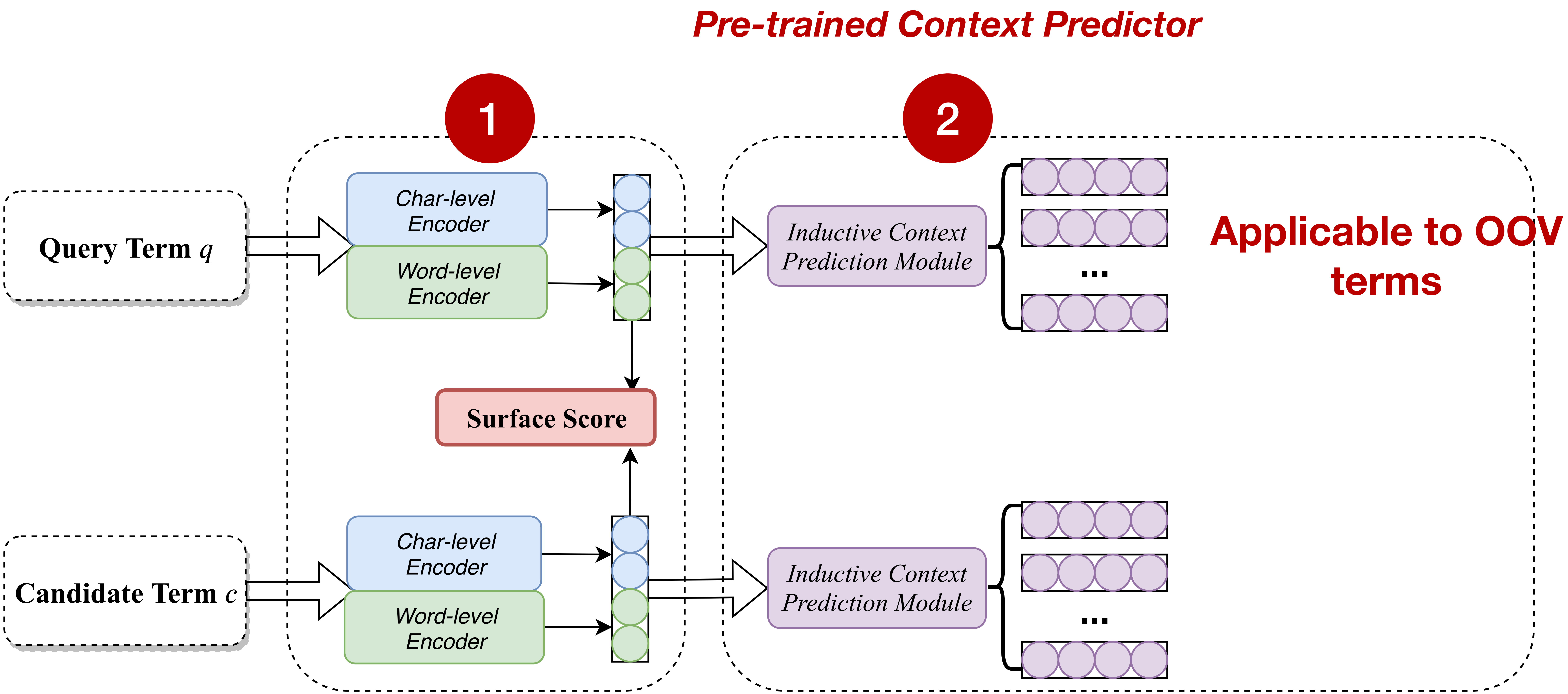
Capture character-level similarity by n-gram embeddings

Recover some semantic meanings by word embeddings

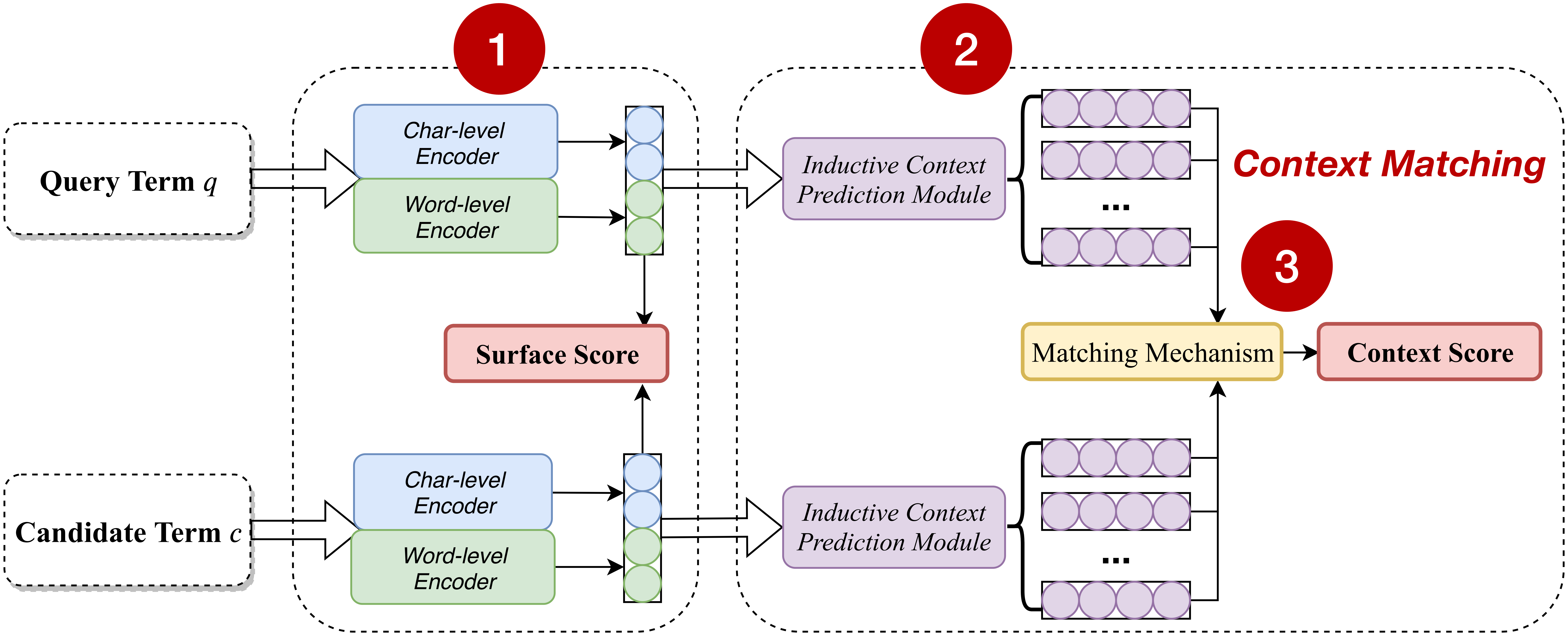
SurfCon Framework



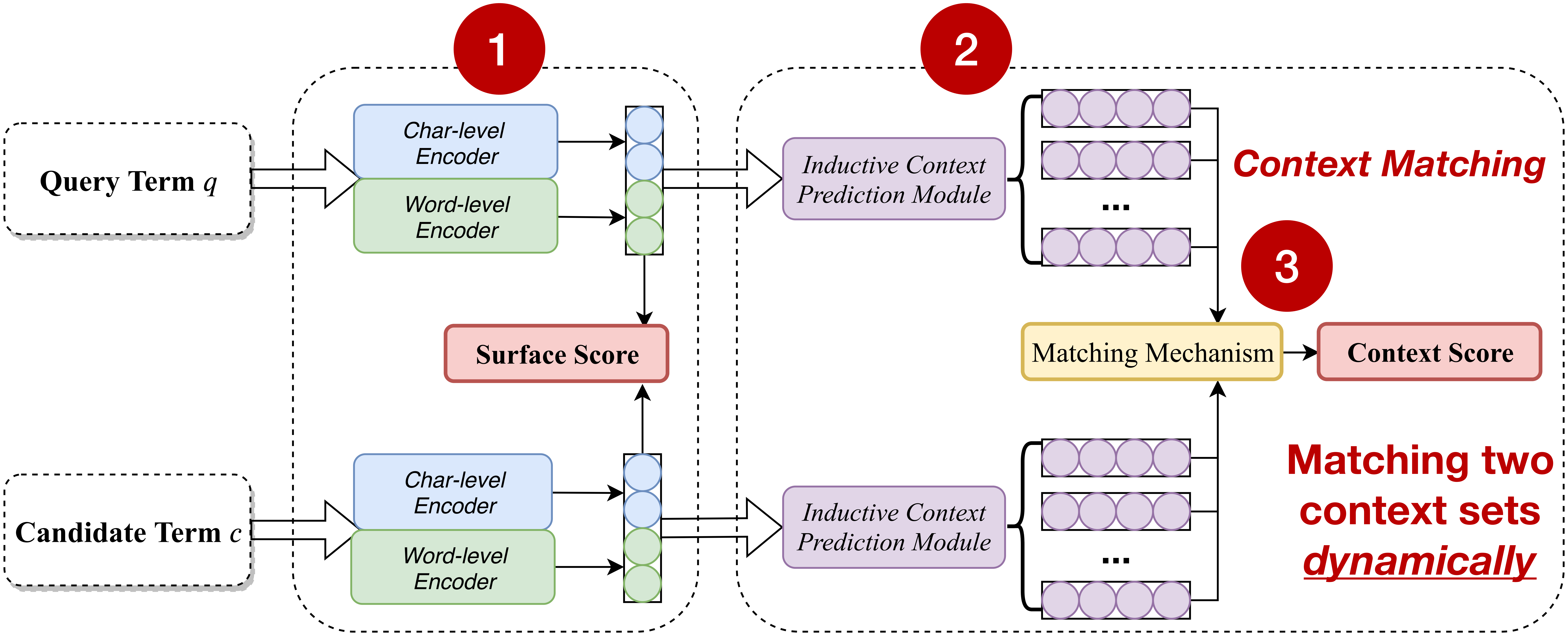
SurfCon Framework



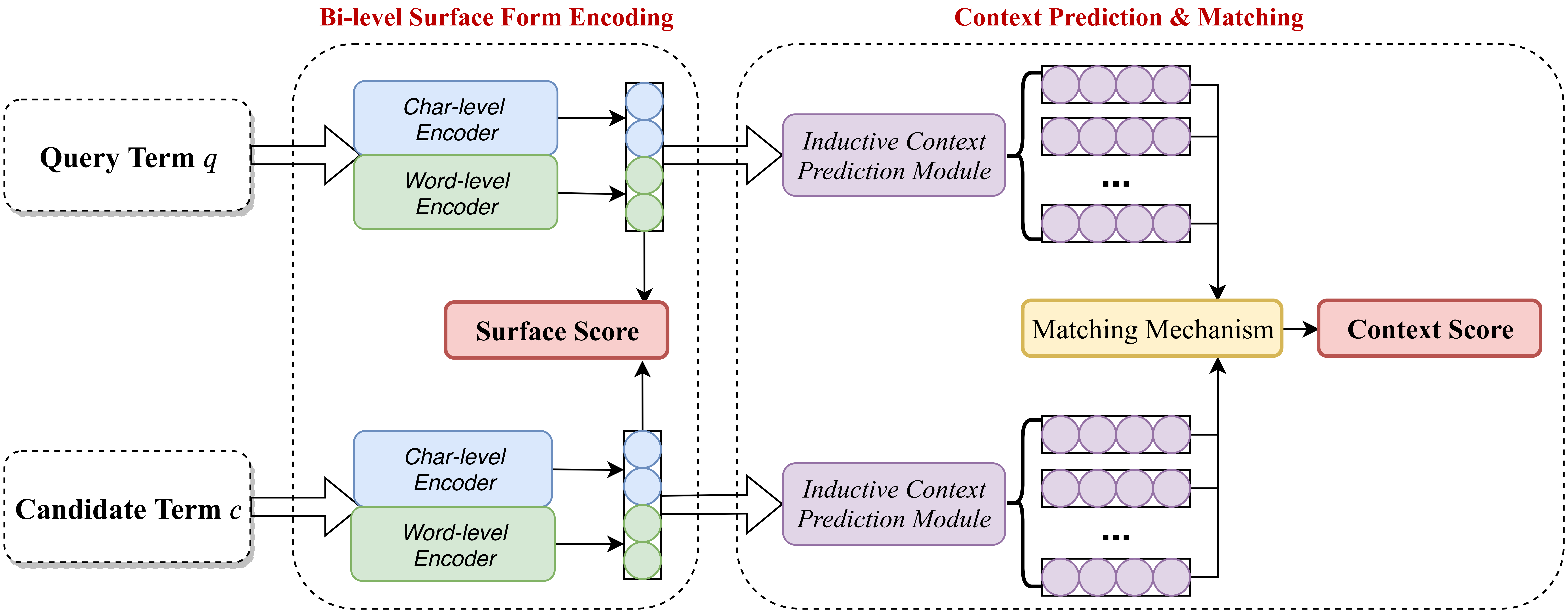
SurfCon Framework



SurfCon Framework

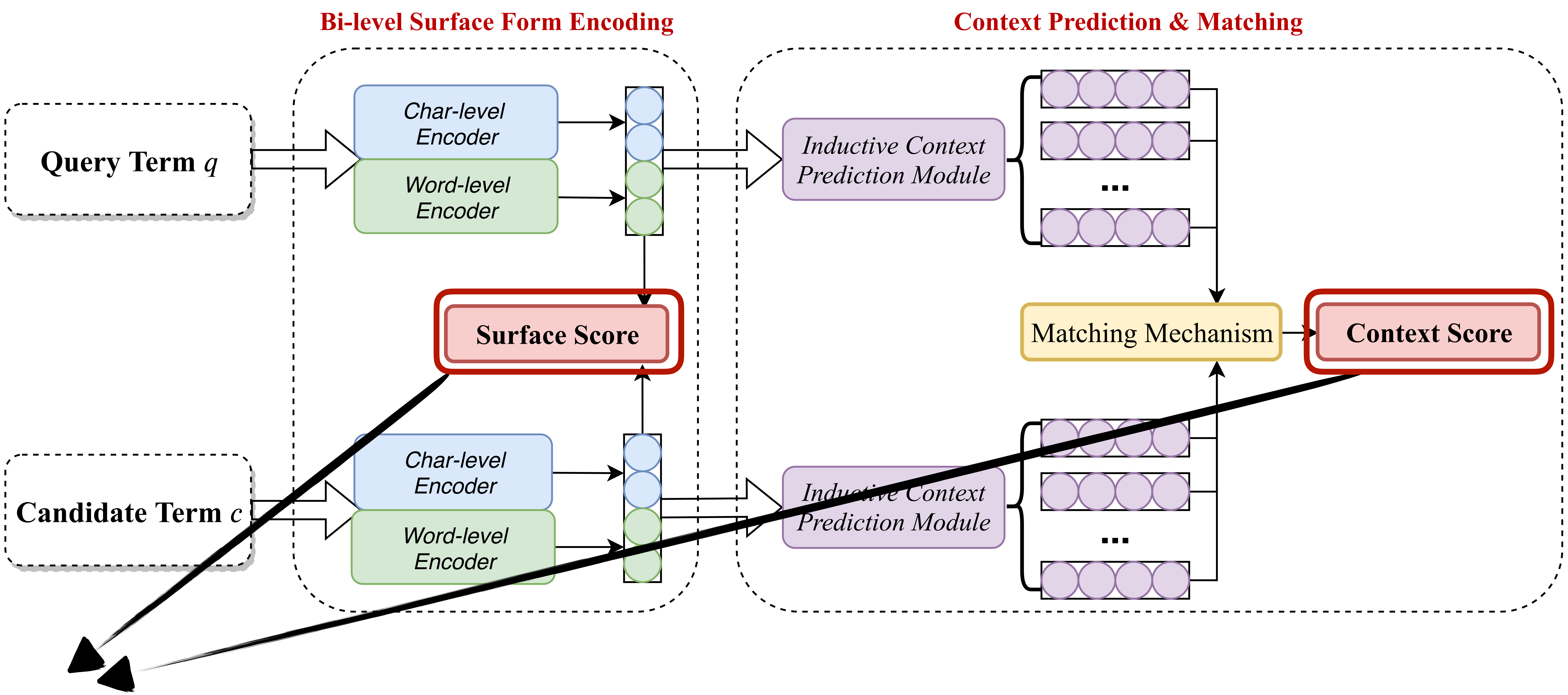


SurfCon Framework



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

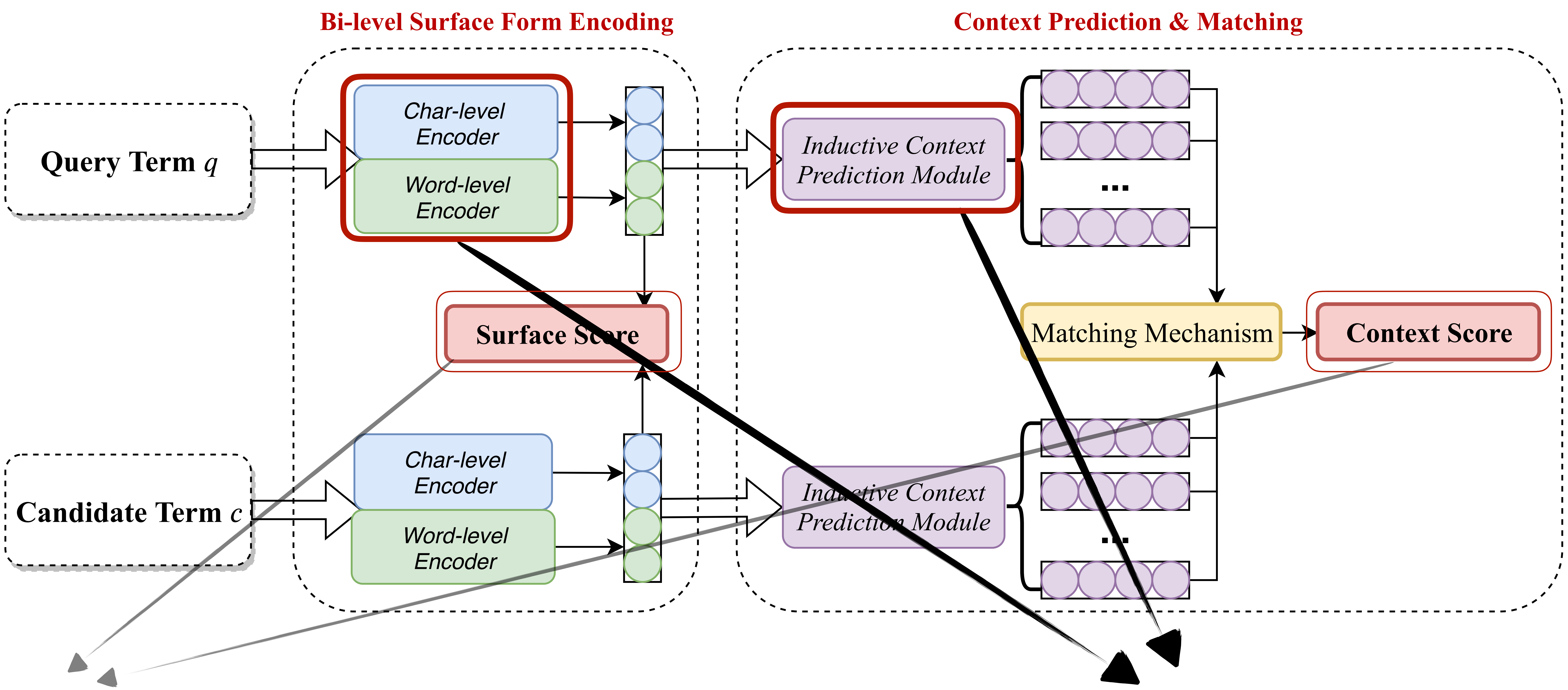
SurfCon Framework



Challenge I: balancing two information

1. Word-level encoder
2. Context score

SurfCon Framework



Challenge I: balancing two information

1. *Word-level encoder*
2. *Context score*

Challenge II: OOV query terms

1. ***Only require surface form as the input***
2. ***Infer global context based on surface form***

Experimental Setup

	1-day dataset	All-day dataset
# Nodes	52,804	43,406
# Edges	16,197,319	50,134,332
Average # Degrees	613.5	2310.0

- Two existing co-occurrence graphs from Finlayson et al. (2014) in *Scientific data*
  - 20 million clinical notes from Stanford Hospital and Clinics since 1995
  - 1-day dataset & all-day dataset
- Synonym labels from UMLS

Experimental Setup

		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

- ❖ Two testing scenarios
  - ❖ **InV** (query) testing and **OOV** (query) testing
  - ❖ A testing subset, **Dissim**, with string-dissimilar synonyms

Main Results

Category	Methods	InV - All	InV - Dissim
Surface form based methods	CharNgram (Hashimoto et al., EMNLP'17)	0.8473	0.4657
	CHARAGRAM (Wieting et al., EMNLP'16)	0.8507	0.5504
	SRN (Neculoiu et al., 2016)	0.8565	0.5102
Global context based methods	Word2vec (Mikolov et al., NeurIPS'13)	0.3748	0.3188
	LINE (2nd) (Tang et al., WWW'15)	0.4301	0.3494
	DEP-NoP (Qu et al., KDD'17)	0.6107	0.4855
Hybrid methods	Concept Space (Wang et al., IJCAI'15)	0.8109	0.4690
	Planetoid (Yang et al., ICML'16)	0.8514	0.5612
Our model and variants	SurfCon (Surf-Only)	0.9053	0.6145
	SurfCon(Static)	0.9151	0.6542
	SurfCon	**0.9176**	**0.6821**

Main Results

Category	Methods	InV - All	InV - Dissim
Surface form based methods	CharNgram (Hashimoto et al., EMNLP'17)	0.8473	0.4657
	CHARAGRAM (Wieting et al., EMNLP'16)	0.8507	0.5504
	SRN (Neculoiu et al., 2016)	0.8565	0.5102
Global context based methods	Word2vec (Mikolov et al., NeurIPS'13)	0.3748	0.3188
	LINE (2nd) (Tang et al., WWW'15)	0.4301	0.3494
	DEP-NoP (Qu et al., KDD'17)	0.6107	0.4855
Hybrid methods	Concept Space (Wang et al., IJCAI'15)	0.8109	0.4690
	Planetoid (Yang et al., ICML'16)	0.8514	0.5612
Our model and variants	SurfCon (Surf-Only)	0.9053	0.6145
	SurfCon(Static)	0.9151	0.6542
	SurfCon	**0.9176**	**0.6821**

Main Results

Category	Methods	InV - All	InV - Dissim
Surface form based methods	CharNgram (Hashimoto et al., EMNLP'17)	0.8473	0.4657
	CHARAGRAM (Wieting et al., EMNLP'16)	0.8507	0.5504
	SRN (Neculoiu et al., 2016)	$\boxed{\textbf{0.8565}}$	0.5102
Global context based methods	Word2vec (Mikolov et al., NeurIPS'13)	0.3748	0.3188
	LINE (2nd) (Tang et al., WWW'15)	0.4301	0.3494
	DEP-NoP (Qu et al., KDD'17)	0.6107	0.4855
Hybrid methods	Concept Space (Wang et al., IJCAI'15)	0.8109	0.4690
	Planetoid (Yang et al., ICML'16)	0.8514	$\boxed{\textbf{0.5612}}$
Our model and variants	SurfCon (Surf-Only)	0.9053	0.6145
	SurfCon(Static)	0.9151	0.6542
	SurfCon	**0.9176 (+7.1%)**	**0.6821 (+21.5%)**

Main Results

Category	Methods	OOV - All	OOV - Dissim
Surface form based methods	CharNgram (Hashimoto et al., EMNLP'17)	0.7427	0.4131
	CHARAGRAM (Wieting et al., EMNLP'16)	$\boxed{0.7609}$	$\boxed{0.5142}$
	SRN (Neculoiu et al., 2016)	0.7241	0.4341
Global context based methods	Word2vec (Mikolov et al., NeurIPS'13)	-	-
	LINE (2nd) (Tang et al., WWW'15)	-	-
	DEP-NoP (Qu et al., KDD'17)	-	-
Hybrid methods	Concept Space (Wang et al., IJCAI'15)	-	-
	Planetoid (Yang et al., ICML'16)	0.731	0.4714
Our model and variants	SurfCon (Surf-Only)	0.8228	0.5829
	SurfCon(Static)	0.8285	0.5933
	SurfCon	**0.8301 (+9.1%)**	**0.6009 (+16.9%)**

Case Studies

Query Term	**“unable to vocalize” (InV Query)**	**“marijuana” (OOV Query)**
SurfCon Top Ranked Candidates	<u>“does not vocalize”</u> <u>“aphonia”</u> <u>“loss of voice”</u> “vocalization” **“unable to phonate”**	**“marijuana abuse”** “**cannabis**” <u>“cannabis use”</u> <u>“marijuana smoking”</u> “narcotic”

- **Bold terms** are existing synonyms in the KBs
- <u>Underlined terms</u> are new synonyms we discover

Conclusions

- **SurfCon model for synonym discovery**
  - Leverages surface form and global context information
  - Handles InV and OOV query terms
  - Does not require raw clinical texts and works on co-occurrence graphs
- **Medical term co-occurrence graph as privacy-aware clinical data**
  - Better preserves privacy
  - Can be used for solving many data mining tasks

THE OHIO STATE UNIVERSITY

NATIONWIDE CHILDREN'S

When your child needs a hospital, everything matters.

Thanks! Any Questions?

Zhen Wang

The Ohio State University

SurfCon: Synonym Discovery on Privacy-Aware Clinical Data

Source Code and Datasets: **https://github.com/yzabc007/SurfCon**

References

- [Roberts, ClinicalNLP'16] Roberts, K., 2016, December. Assessing the corpus size vs. similarity trade-off for word embeddings in clinical NLP. In Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP) (pp. 54-63).

- [Mueller and Thyagarajan, AAAI'16] Mueller, J. and Thyagarajan, A., 2016, March. Siamese recurrent architectures for learning sentence similarity. In Thirtieth AAAI Conference on Artificial Intelligence.

- [Neculoiu et al., 2016] Neculoiu, P., Versteegh, M. and Rotaru, M., 2016, August. Learning text similarity with siamese recurrent networks. In Proceedings of the 1st Workshop on Representation Learning for NLP (pp. 148-157).

- [Perozzi et al., KDD'14] Perozzi, B., Al-Rfou, R. and Skiena, S., 2014, August. Deepwalk: Online learning of social representations. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 701-710). ACM.

- [Tang et al., WWW'15] Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J. and Mei, Q., 2015, May. Line: Large-scale information network embedding. In Proceedings of the 24th international conference on world wide web (pp. 1067-1077). International World Wide Web Conferences Steering Committee.

- [Grover and Leskovec, KDD'16] Grover, A. and Leskovec, J., 2016, August. node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 855-864). ACM.

- [Finlayson et al., 2014] Finlayson, S.G., LePendu, P. and Shah, N.H., 2014. Building the graph of medicine from millions of clinical narratives. Scientific data, 1, p.140032.

References

- [Yang et al., ICML'16] Z. Yang, W. W. Cohen, and R. Salakhutdinov. 2016. Revisiting semi-supervised learning with graph embeddings. In ICML.

- [Zhang et al., 2018] Zhang, C., Li, Y., Du, N., Fan, W. and Yu, P.S., 2018. SynonymNet: Multi-context Bilateral Matching for Entity Synonyms. arXiv preprint arXiv:1901.00056.

- [Wang et al., IJCAI'15] Wang, C., Cao, L. and Zhou, B., 2015, June. Medical synonym extraction with concept space models. In Twenty-Fourth International Joint Conference on Artificial Intelligence.

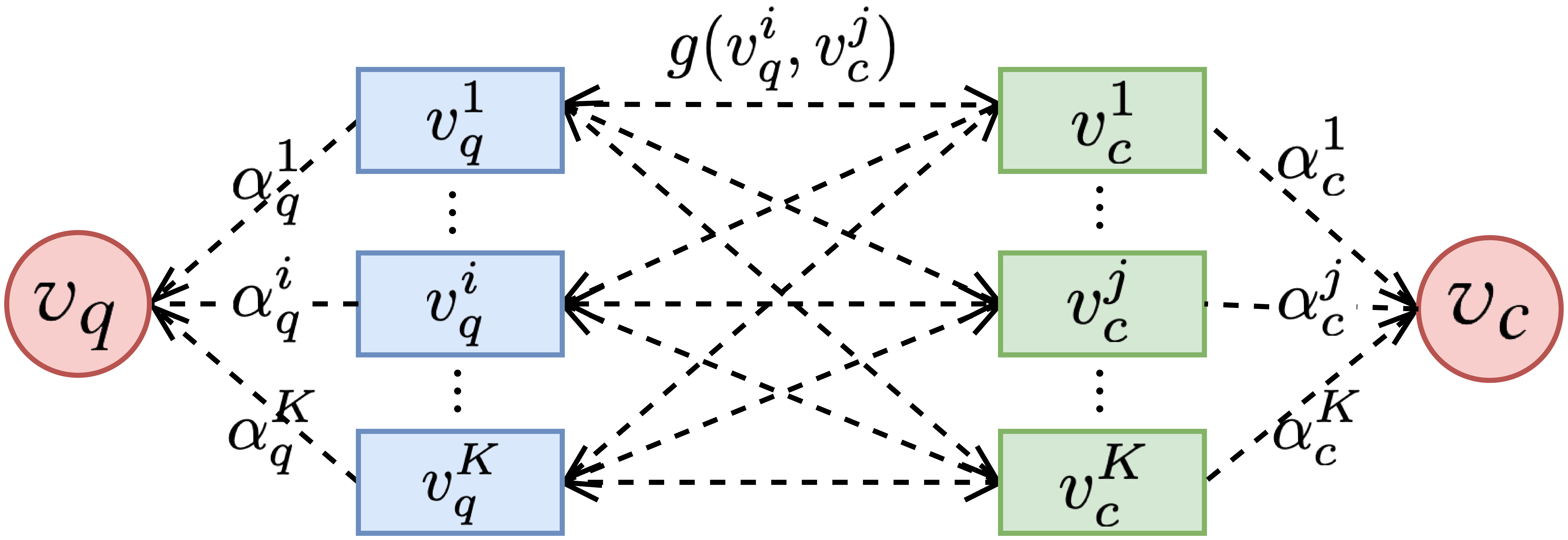
- [Qu et al., KDD'17] Qu, M., Ren, X. and Han, J., 2017, August. Automatic synonym discovery with knowledge bases. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 997-1005). ACM.

- [Hashimoto et al., EMNLP'17] Hashimoto, K., Xiong, C., Tsuruoka, Y. and Socher, R., 2017. A joint many-task model: Growing a neural network for multiple nlp tasks. In EMNLP.

- [Wieting et al., EMNLP'16] Wieting, J., Bansal, M., Gimpel, K. and Livescu, K., 2016. Charagram: Embedding words and sentences via character n-grams. In EMNLP.

- [Mikolov et al., NeurIPS'13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Dynamic Context Matching Mechanism



Weigh each context of query term based on its matching degree with contexts of the candidate term, and vice versa.