The 14th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics

ACM-BCB 2023 | Houston, TX, September 3-6, 2023

Text-to-ESQ: A Two-Stage Controllable Approach for Efficient Retrieval of Vaccine Adverse Events from NoSQL Database

Wenlong Zhang[1], Kangping Zeng[2], Xinming Yang[1], Tian Shi, Ping Wang[1]

[1]Dept. of Computer Science, Stevens Institute of Technology, Hoboken, NJ, USA

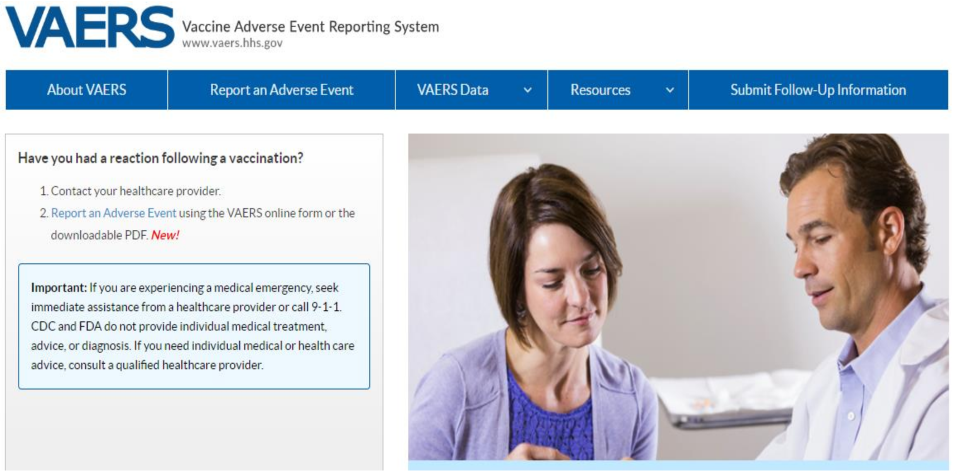
[2]School of Business, Stevens Institute of Technology, Hoboken, NJ, USA

STEVENS INSTITUTE of TECHNOLOGY

Outline

- Introduction
- Challenges
- Contribution
- Experiments
- Conclusion

Vaccine Adverse Events Report System (VAERS)

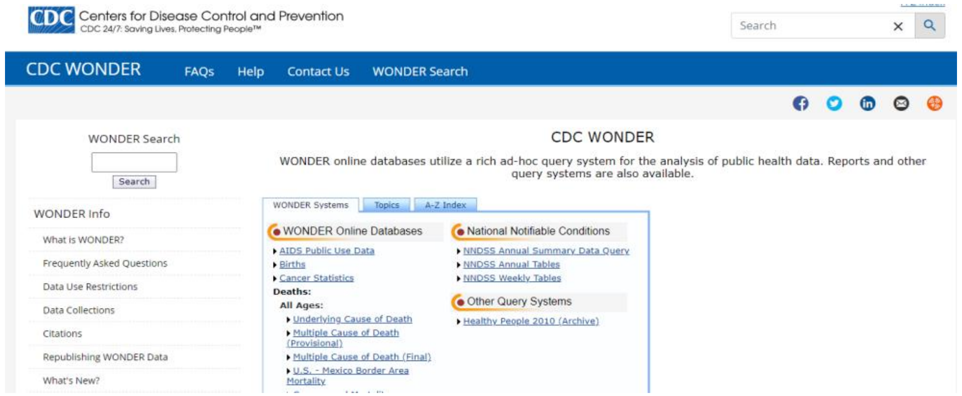


Have you had a reaction following a vaccination?

1. Contact your healthcare provider.
2. Report an Adverse Event using the VAERS online form or the downloadable PDF. *New!*

Important: If you are experiencing a medical emergency, seek immediate assistance from a healthcare provider or call 9-1-1. CDC and FDA do not provide individual medical treatment, advice, or diagnosis. If you need individual medical or health care advice, consult a qualified healthcare provider.

Vaccine Adverse Event Reporting System(VAERS1) co-managed by the U.S. FDA and CDC is an important platform for reporting and analyzing side effects after getting vaccines.



CDC WONDER

WONDER online databases utilize a rich ad-hoc query system for the analysis of public health data. Reports and other query systems are also available.

WONDER Online Databases

- AIDS Public Use Data
- Births
- Cancer Statistics

Deaths:

All Ages:

- Underlying Cause of Death
- Multiple Cause of Death (Provisional)
- Multiple Cause of Death (Final)
- U.S. - Mexico Border Area Mortality

National Notifiable Conditions

- NNDSS Annual Summary Data Query
- NNDSS Annual Tables
- NNDSS Weekly Tables

Other Query Systems

- Healthy People 2010 (Archive)

The VAERS data has been continuously updated since 1990,including structured information such as demographic information, vaccine details, and various coded symptoms, as well as narrative text descriptions. Currently, the VAERS data can be accessed via the CDC's WONDER system.

limitations

However, there are several limitations to such a system

(1) Complicated to use

(2) Inflexible to extend

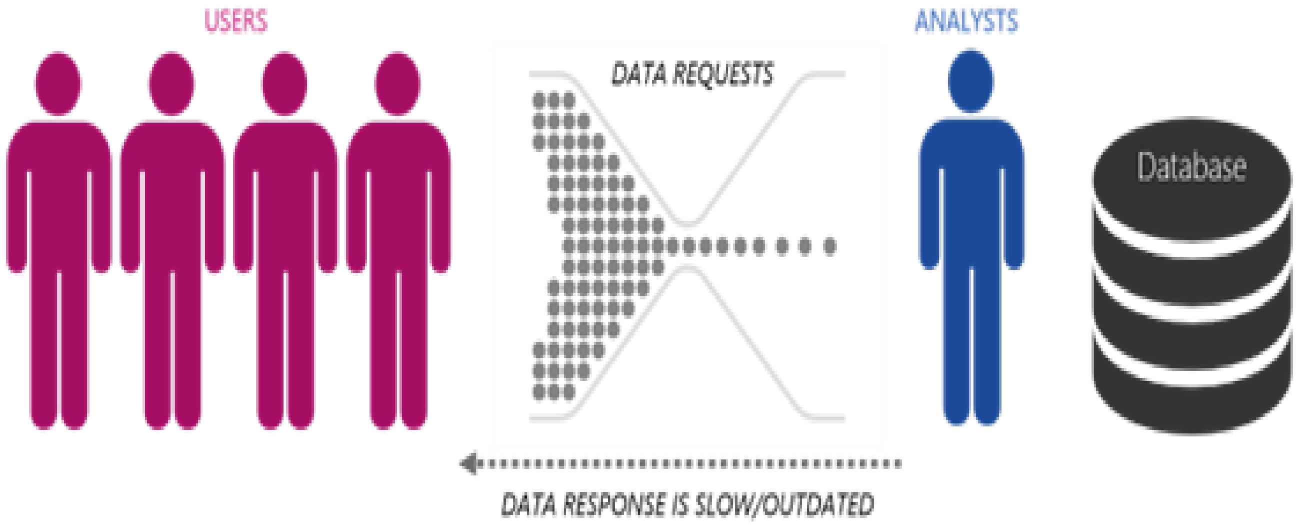
- These limitations can be potentially addressed by Text-to-SQL, which aims to automatically translate natural language questions to SQL queries with different NLP techniques. However, Text-to-SQL is primarily designed for retrieving information from SQL databases with relational tables.

Challenges with Text-to-SQL

- Traditional research focus on SQL database

Limitation

- Text-to-SQL capabilities are limited by the data structures and functionality of SQL databases
- It is difficult to incorporate external knowledge bases (KBs) into relational tables



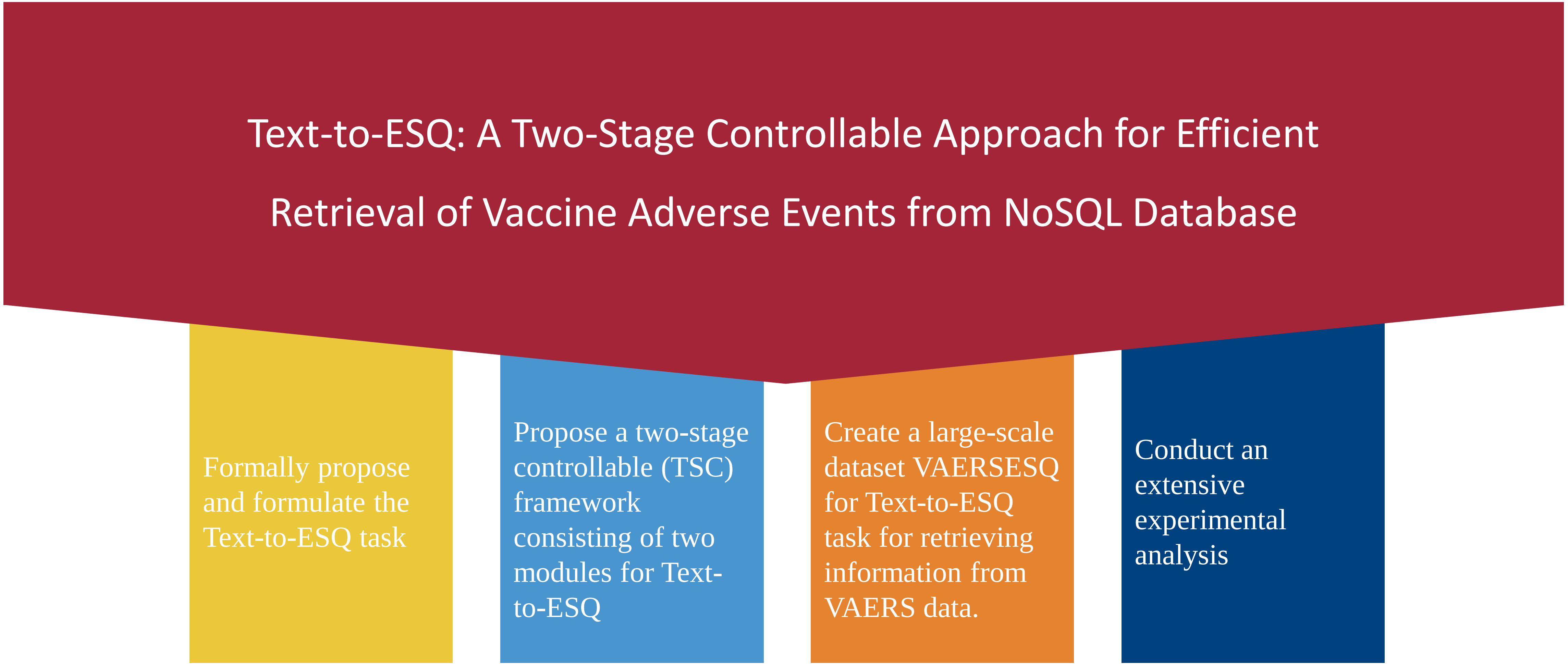
elasticsearch

Solution

No-SQL database

- Handle large volumes of data at high speed with a scale-out architecture
- Store unstructured, semi-structured, or structured data

Contribution



Text-to-ESQ: A Two-Stage Controllable Approach for Efficient Retrieval of Vaccine Adverse Events from NoSQL Database

Formally propose and formulate the Text-to-ESQ task

Propose a two-stage controllable (TSC) framework consisting of two modules for Text-to-ESQ

Create a large-scale dataset VAERSESQ for Text-to-ESQ task for retrieving information from VAERS data.

Conduct an extensive experimental analysis

Our Contributions

Task

- Formally propose and formulate the Text-to-ESQ task

Module

- Propose a two-stage controllable (TSC) framework consisting of two modules for Text-to-ESQ

Dataset

- Create a large-scale dataset VAERSESQ for Text-to-ESQ task for retrieving information from VAERS data.

Experiment

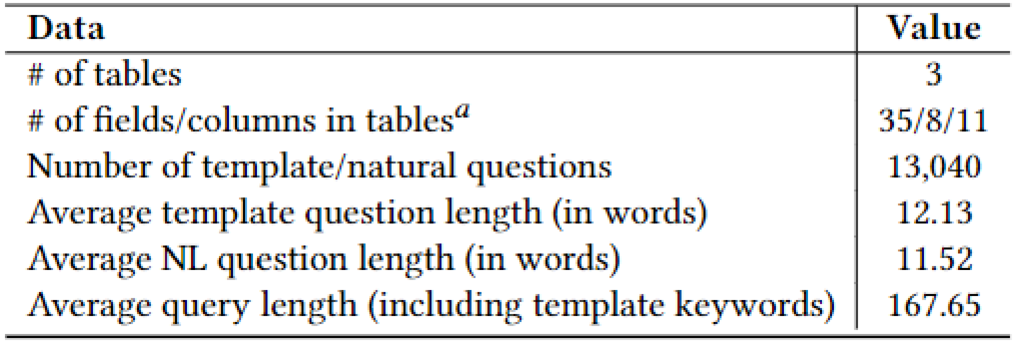
- Conduct an extensive experimental analysis

VAERSESQ Data Generation

- Question Template Collection and Population.

  - How many people have [SYMPTOM] after vaccination?
  - Give me all the patients who got [VAX_NAME_1] vaccine and [VAX_NAME_2] vaccine.
  - Search all the patients who are diagnosed of [HISTORY].

- Natural Language Question Generation with Back-Translation.



Data	Value
# of tables	3
# of fields/columns in tables[a]	35/8/11
Number of template/natural questions	13,040
Average template question length (in words)	12.13
Average NL question length (in words)	11.52
Average query length (including template keywords)	167.65

- Elasticsearch Query Generation.

  when generating the template questions, the corresponding Elasticsearch queries are generated at the same time by populating the placeholders in the query templates.

VAERSESQ Data Generation

Question Template Collection and Population.

- How many people have [SYMPTOM] after vaccination?
- Give me all the patients who got [VAX_NAME_1] vaccine and [VAX_NAME_2] vaccine.
- Search all the patients who are diagnosed of [HISTORY].

Natural Language Question Generation with Back-Translation.

- Number of template/natural questions:13040
- Average template question length (in words):12.13
- Average NL question length (in words) :11.52

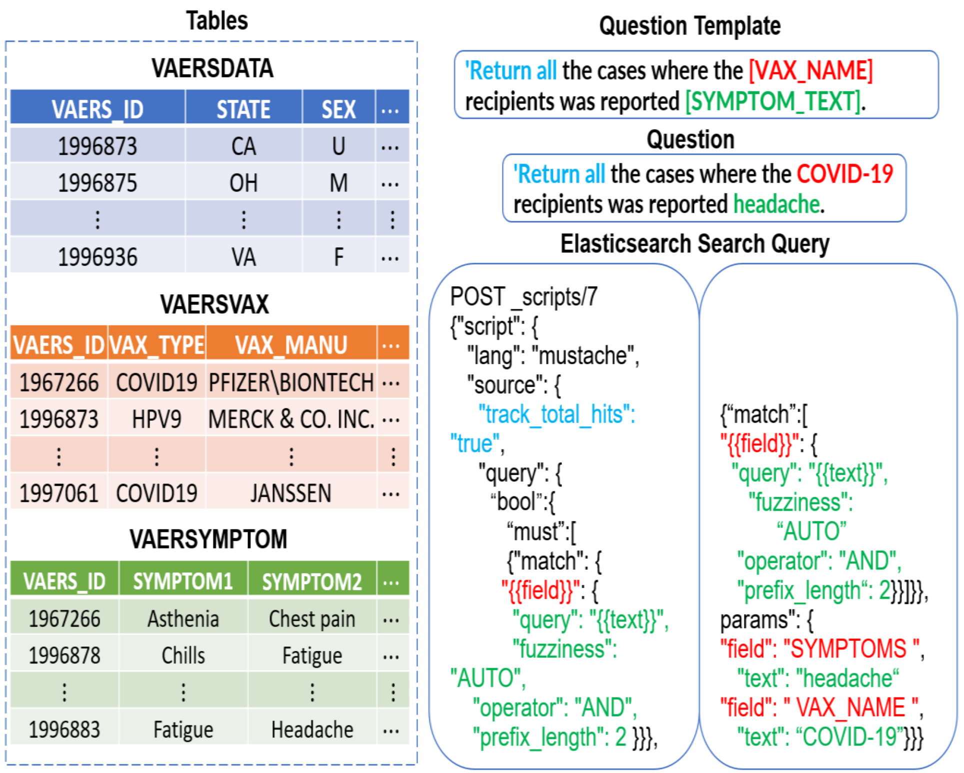
Elasticsearch Query Generation.

- when generating the template questions, the corresponding Elasticsearch queries are generated at the same time by populating the placeholders in the query templates.

Dataset

- The VAERSESQ dataset is publicly available at https://github.com/LEAF-Lab-Stevens/Text2ESQ.

An example from VAERSESQ data.



Tables

VAERSDATA

VAERS_ID	STATE	SEX	⋯
1996873	CA	U	⋯
1996875	OH	M	⋯
⋮	⋮	⋮	⋮
1996936	VA	F	⋯

VAERSVAX

VAERS_ID	VAX_TYPE	VAX_MANU	⋯
1967266	COVID19	PFIZER\BIONTECH	⋯
1996873	HPV9	MERCK & CO. INC.	⋯
⋮	⋮	⋮	⋮
1997061	COVID19	JANSSEN	⋯

VAERSYMPTOM

VAERS_ID	SYMPTOM1	SYMPTOM2	⋯
1967266	Asthenia	Chest pain	⋯
1996878	Chills	Fatigue	⋯
⋮	⋮	⋮	⋮
1996883	Fatigue	Headache	⋯

Question Template

'Return all the cases where the [VAX_NAME] recipients was reported [SYMPTOM_TEXT].

Question

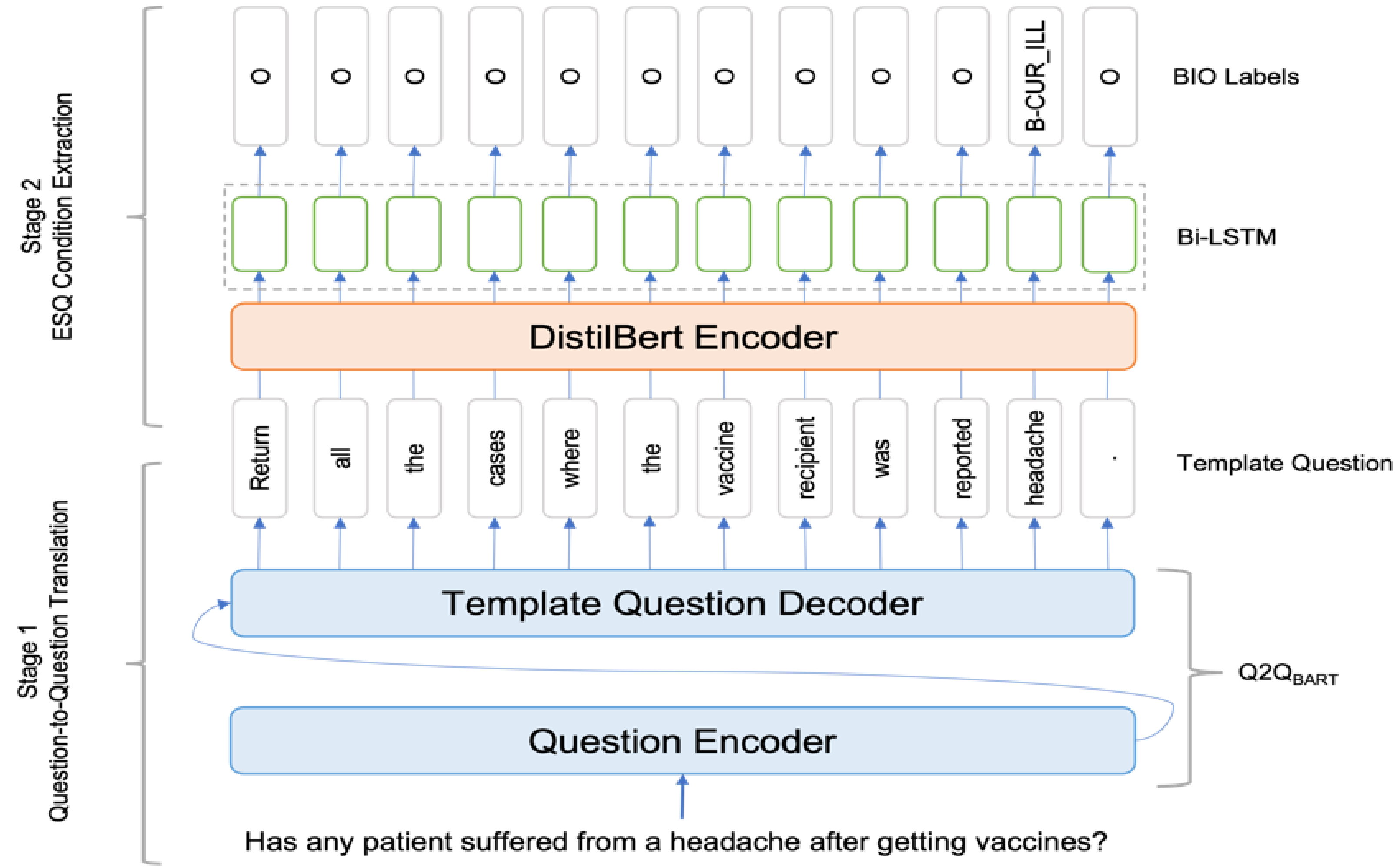
'Return all the cases where the COVID-19 recipients was reported headache.

Elasticsearch Search Query

```
POST _scripts/7
{"script": {
  "lang": "mustache",
  "source": {
    "track_total_hits":
"true",
    "query": {
      “bool”:{
        “must”:[
        {"match": {
       "{{field}}": {
         "query": "{{text}}",
         "fuzziness":
"AUTO",
     "operator": "AND",
     "prefix_length": 2 }}},
```

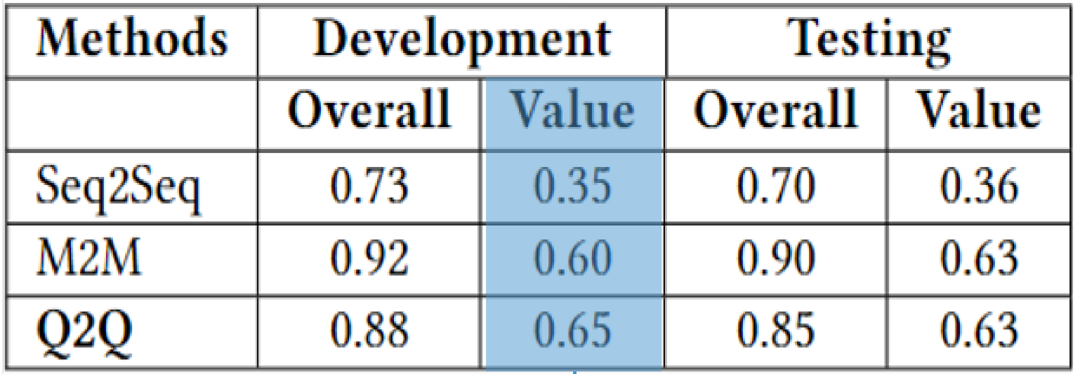
```
{“match”:[
"{{field}}": {
  "query": "{{text}}",
    "fuzziness":
        “AUTO”
   "operator": "AND",
   "prefix_length“: 2}}]}},
params": {
"field": "SYMPTOMS ",
   "text": "headache“
"field": " VAX_NAME ",
   "text": “COVID-19”}}}
```

TSC Model


Stage 2
ESQ Condition Extraction
O
O
O
O
O
O
O
O
O
O
B-CUR_ILL
O
BIO Labels
Bi-LSTM
DistilBert Encoder
Return
all
the
cases
where
the
vaccine
recipient
was
reported
headache
.
Template Question
Stage 1
Question-to-Question Translation
Template Question Decoder
Question Encoder
$Q2Q_{BART}$
Has any patient suffered from a headache after getting vaccines?

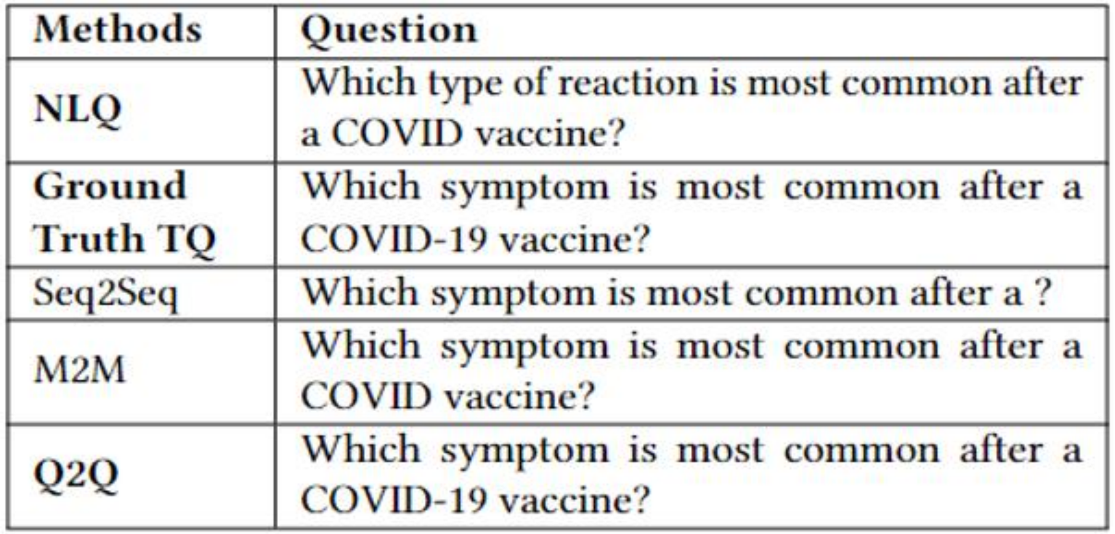
Experiments

Results of Stage1: Question to Question translation



Methods	Development		Testing	
	Overall	Value	Overall	Value
Seq2Seq	0.73	0.35	0.70	0.36
M2M	0.92	0.60	0.90	0.63
Q2Q	0.88	0.65	0.85	0.63

An example of translating the natural language questions (NLQ) into template questions (TQ) with the Q2Q module

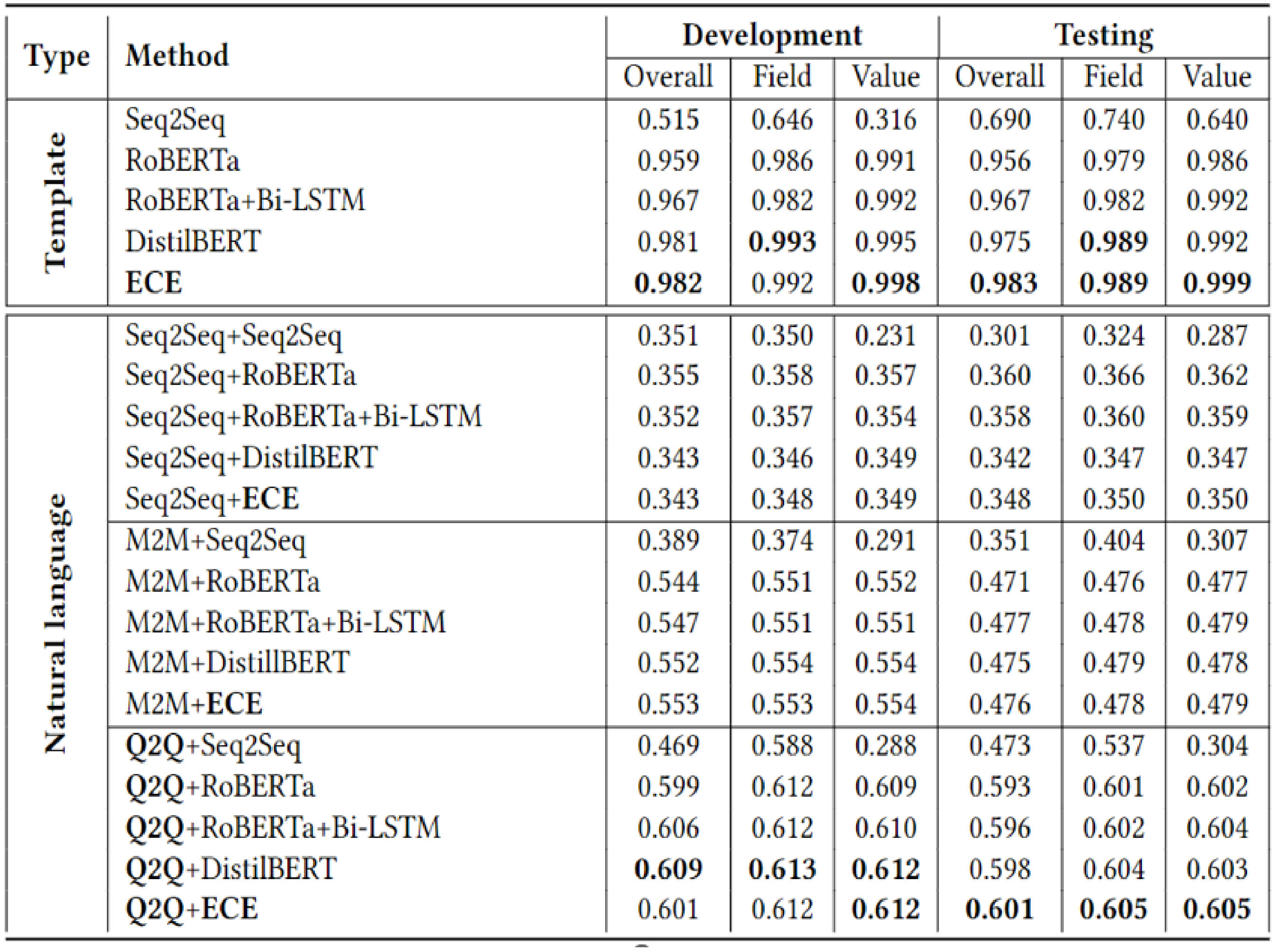


Methods	Question
NLQ	Which type of reaction is most common after a COVID vaccine?
Ground Truth TQ	Which symptom is most common after a COVID-19 vaccine?
Seq2Seq	Which symptom is most common after a ?
M2M	Which symptom is most common after a COVID vaccine?
Q2Q	Which symptom is most common after a COVID-19 vaccine?

Most challenge part: we employ strict rules, including thorough punctuation and spacing checks based on the ground truth, to evaluate the results.

Experiments

Results of Stage2: ESQ Condition Extraction



Type	Method	Development			Testing		
		Overall	Field	Value	Overall	Field	Value
Template	Seq2Seq	0.515	0.646	0.316	0.690	0.740	0.640
	RoBERTa	0.959	0.986	0.991	0.956	0.979	0.986
	RoBERTa+Bi-LSTM	0.967	0.982	0.992	0.967	0.982	0.992
	DistilBERT	0.981	**0.993**	0.995	0.975	**0.989**	0.992
	ECE	**0.982**	0.992	**0.998**	**0.983**	**0.989**	**0.999**
Natural language	Seq2Seq+Seq2Seq	0.351	0.350	0.231	0.301	0.324	0.287
	Seq2Seq+RoBERTa	0.355	0.358	0.357	0.360	0.366	0.362
	Seq2Seq+RoBERTa+Bi-LSTM	0.352	0.357	0.354	0.358	0.360	0.359
	Seq2Seq+DistilBERT	0.343	0.346	0.349	0.342	0.347	0.347
	Seq2Seq+**ECE**	0.343	0.348	0.349	0.348	0.350	0.350
	M2M+Seq2Seq	0.389	0.374	0.291	0.351	0.404	0.307
	M2M+RoBERTa	0.544	0.551	0.552	0.471	0.476	0.477
	M2M+RoBERTa+Bi-LSTM	0.547	0.551	0.551	0.477	0.478	0.479
	M2M+DistillBERT	0.552	0.554	0.554	0.475	0.479	0.478
	M2M+**ECE**	0.553	0.553	0.554	0.476	0.478	0.479
	Q2Q+Seq2Seq	0.469	0.588	0.288	0.473	0.537	0.304
	Q2Q+RoBERTa	0.599	0.612	0.609	0.593	0.601	0.602
	Q2Q+RoBERTa+Bi-LSTM	0.606	0.612	0.610	0.596	0.602	0.604
	Q2Q+DistilBERT	**0.609**	**0.613**	**0.612**	0.598	0.604	0.603
	Q2Q+**ECE**	0.601	0.612	**0.612**	**0.601**	**0.605**	**0.605**

The TSC model outperforms the baseline in terms of performance, and there remains untapped potential for further exploration.

Conclusion

- Introduce the Text-to-ESQ task, facilitating NLQ on NoSQL databases.
- Introduces the novel Two-Stage Controllable (TSC) framework
- Contributes a substantial VAERSESQ dataset for Text-to-ESQ
- A comprehensive experimental analysis

THANK YOU

Link to VAERSESQ dataset and codes:

https://github.com/LEAF-Lab-Stevens/Text2ESQ

Please feel free to send questions and suggestions to:

wzhang71@stevens.edu

pwang44@stevens.edu