**Semantic Information Availability (SIA) Task Guidelines**

Contact: Man Luo ( [mluo26@asu.edu](mailto:mluo26@asu.edu) )

**Table of Contents**

[**Background**](#_ppnm5ib01nwc) **2**

[**Goal**](#_tobcl647y4qd) **2**

[The Difference between SIA and STS](#_3x3vh8t5ky8y) 2

[**SIA Scale**](#_aqinjhh4eg4j) **3**

[Guideline of Creating SIA dataset](#_4c1tivwol497) 3

[**Dataset Creation Task**](#_73akh8kxsxmi) **5**

# 

# Background

**Information Retrieval (IR)** is a fundamental component of search engines: With a database, given a query, the system is able to retrieve relevant sentences(documents) to answer the question.

*For example:*

*Q: “Is there any role of interleukin-11 in cardiovascular fibrosis?”*

*A: “These data reveal a central role of IL11 in fibrosis and we propose inhibition of IL11 as a new therapeutic strategy to treat fibrotic diseases.”*

# Goal

The goal is to devise a general approach to create **Semantic Information Availability (SIA)** dataset to do better IR. We will use the SIA dataset to train a baseline and test the baseline on a public dataset, like MSMARCO, bioASQ, to show improvement on baseline. SIA dataset is that given query Q and sentence S, the SIA score reflects how much information does S have to answer Q.

### The Difference between SIA and STS

1. SIA dataset is Question, Answer pair <Q,S> and the SIA score of <Q, S> depends on how well S answers Q. STS dataset is two sentences <S1, S2>, the STS score of <S1, S2> depends on how close the semantic meanings of two sentences are.
2. E.g.

Q: Is the protein Papilin secreted?

S: Apilins are homologous, secreted extracellular matrix proteins which share a common order of protein domains.

From the perspective of SIA, <Q,S> has the highest score 4 since S contains extract information to answer Q. From the perspective of STS, <Q, S> does not have the highest score since they are not semantically closely related.

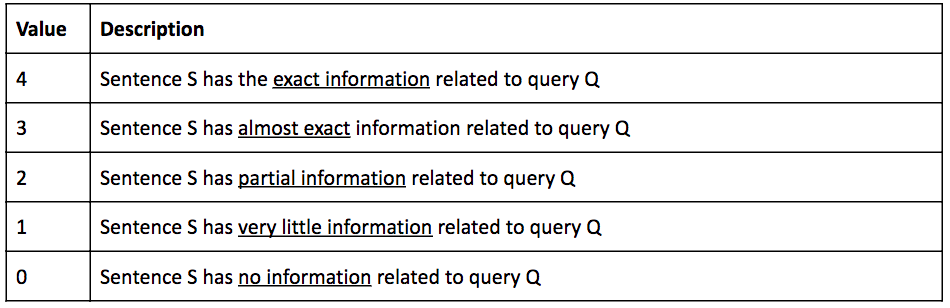
# 

# 

# SIA Scale

**General SIA Scale:**

*Given a Query Q and a Sentence S:*



# Guideline of Creating SIA dataset

### Annotation steps

1. Given a question and sentence pair <Q,S>, extract key phrases in Q.
2. Align each keyword in Q with context in S: how well the context in S aligns with the key phrase.
3. For each alignment, label the type of the alignment. For each labeling, assign a score[0,4].
4. Average the score of all alignment, which is the SIA score of <Q, S>. If one alignment is OPPO type(see below), then the score of <Q,S> is 0.

### Label of alignment

1. **EQUI**: the key phrase in Q can be matched with some content in S, they are semantically equivalent in this context. This label always has score 4.
   1. E.g. Q: What causes a [blocked nose]? [The blocked, runny or stuffy nose] is purely a symptom caused by a multitude of complaints.
2. **SIMI**: the key phrase and some content in S have close meaning. Score should be between [1-4).
   1. the key phrase and some content in S could be synonyms, or one is the abbreviation to the other. E.g. Q: Which [acetylcholinesterase inhibitors] are used for treatment of myasthenia gravis? S:Novel [AChE inhibitors] with oral antisense oligonucleotides have been developed and preliminary results appear to be promising.
   2. the key phrase in Q is more specific than some content in S. E.g Q: how do you know when you have [congenital night blindness]? S:There are several common signs of [night blindness]: Eyes adjust slowly to light changes, such as when entering a dark room from the sunny outdoors.
   3. the key phrase in Q is more general than some content in S. E.g. Q: what causes [blocked nose]? S: [Unilateral blocking of the nose] is usually caused by a congenital or a traumatic deviation of the nasal septum.
3. **ANS**:​ Some content in S is corresponding to the answer to Q. Score should be between [1-4].
   1. Q: [what causes] a blocked nose ? S: The 3 most frequent causes of snoring in children are enlarged adenoids, enlarged tonsils and blocked nose [due to allergic rhinitis].
4. **OPPO**: ​the meanings of the key phrase and some context in S are in opposition to each other, lying in an inherently incompatible binary relationship. If OPPO appears, then the final SIA score should be 0.
   1. E.g. Q: I have a cough but I [don’t have a fever], what medicine should I take? S:In the case of [a high fever, or a low fever that's causing discomfort], your doctor may recommend an over-the-counter medication, such as acetaminophen. be
5. **NOALI**: the key phrase in Q can not be aligned with any content in the S.Therefore,it is left unaligned. Score should be 0.
   1. E.g Q: [what causes] [blocked nose]? S:What Not To Do: See IIIb & IIIc a. Superficial and Saphenous Nerve Block.

# 

# Dataset Creation Task

**Phase 1:** you will manually annotate 60 question and answer pairs followed above instruction.[expected hour: 1.5 hour]. Specifically, you need to do:

1. Extract the key phrases in Q.
2. Identify the content in S aligning with each key phrase.
3. Label each alignment and give a score.

You can find the file named by you in this folder: [https://drive.google.com/dri ve/folders/1oIxmlSIBhSJBUizr7cXcyI64LWJfVStk?usp=sharing](https://drive.google.com/drive/folders/1oIxmlSIBhSJBUizr7cXcyI64LWJfVStk?usp=sharing)

**Phase 2:** you will devise an algorithm to automatically (or semi-automatic) create SIA dataset based on **one existing IR relevant dataset**. Consider the following approaches:

Approach 1:

1. Extract a list of key phrases from questions.
2. Generate a sentence which has SIA score 4, called this sentence as the *ideal answer*.
3. For each key phrase, generate a sentence by replacing the content in the *ideal answer* with other content such that the label is any type of SIMI, SEP1, SEP2, OPPO, etc….
4. Take each key phrase as S, such that <Q, S> has a low SIA score.
5. For each question Q, generate a similar question Q’ as a negative sentence (<Q, Q’> SIA score is 0).
6. For each sentence, increase the language diversity.

Approach 2:

1. Train(rely on) a STS model.
2. Generate the SIA dataset by approximate SIA of <Q, S> to STS of <Q+exact answer, S>.

Note:

1. The NLTK Python library may be helpful to easily perform many sentence preprocessing.
2. The above step is just some reference, you are encouraged to devise a novel approach to automatically create SIA dataset.

# 

# 

# First Report Requirement

The report includes three parts:

1. Your thought of the manual annotation(60 questions), for example, what do you think would be a good way to decide the score of SIMI and ANS; how do you decide the key phrases, how do you find the alignment in the sentence for each key phrase. Etc…
2. The algorithm you take to generate SIA for a particular dataset (could be a public dataset or you generate from scratch). If you take the Approach 1/2 suggested , be clear of each step, which tool/model you use in each step.
3. The analysis of how well the algorithm does, the evaluation could be based on some metric for the dataset, if there is no existing metric, you could sample some generated data and see how well it does and give some examples to support your statement.

# Reference

[1][QASC: A Dataset for Question Answering via Sentence Composition](https://arxiv.org/pdf/1910.11473.pdf)

[2][HOTPOTQA: A Dataset for Diverse, Explainable Multi-hop Question Answering](https://arxiv.org/pdf/1809.09600.pdf)

[3][MS MARCO: A Human Generated MAchine Reading COmprehension Dataset](https://arxiv.org/pdf/1611.09268.pdf)

[4][TREC Complex Answer Retrieval Overview](https://trec.nist.gov/pubs/trec26/papers/Overview-CAR.pdf)

[5][ANTIQUE: A Non-Factoid Question Answering Benchmark](https://arxiv.org/pdf/1905.08957.pdf)

[6][WikiPassageQA: A Benchmark Collection for Research on Non-factoid Answer Passage Retrieval](https://arxiv.org/pdf/1805.03797.pdf)

|  |  |
| --- | --- |
| Person Name | Dataset (the *No.* of above dataset or *scratch* if like to generate from scratch) |
| Jayasurya | 1 - QASC |
| Sahit | 3 - MS MARCO |
| Rajasree | 5 - ANTIQUE |
| Mihir | 6 - WikiPassageQA |
| Gouthami | 2 - HOTPOTQA |