# Question Answering

**Team Members**
Rakesh Chada (110012490)
Pranneetha Bellamkonda (110166357)

## 1. Introduction

The task of "Question Answering" lies at the heart of the Natural Language Processing (NLP) challenges. This also takes NLP closer to one of its major goals of using Artificial Intelligence to solve problems. Despite many exciting advances in this field, the central goal of building a knowledgeable machine that can answer questions through reasoning from text still seems distant. In this project, we study such approaches for the task of Question Answering and present our techniques that describe our attempts to solve it. The specific problem we are studying is the task of answering several "**Multiple Choice** Questions" of the **4th grade science exam** level.

Our work is based on several existing state-of-the-art techniques used in the field. A straight-forward approach is to just do a plain Bag-of-Words (BoW) matching between the question and the candidate answers and return the answer that matches the most based on a score measure. Other sophisticated approaches involve the use of the syntactic and semantic features of the text. A solution proposed by Shen et al. (2007) [1] describes the use of semantic role labels in Question Answering. They had used **Shalmaneser** (Erk and Pado, 2006)[2] for the generation of semantic role labels. After the generation of labels, they used a novel "Graph-Based" matching approach to match the labels of the question and the candidate answer. Although there is an improvement in the accuracy over the existing approaches at that time, the effectiveness of this solution was severely limited by the FrameNet coverage issues at that time and poor label generation. A few years later, Dipanjan Das and his team had come up with a probabilistic framework [3] for the semantic role labeling. This tool, named **SEMAFOR**[3], seemed more successful in producing the semantic role labels than the existing approaches and hence this got the researchers interested again in using semantic roles for Question Answering.

Our idea is to combine the power of a better semantic role labeler and the techniques that rely on such labels to answer questions. We also use a novel multi-level matching based on the semantic frame names and the inherent semantic frame relations. Furthermore, we had tried several new interesting approaches such as combining Semantic Frames and the Bag-of-Words approaches based on a "**Confidence** Measure" and using "**Focus** Word" based matching on candidate sentences and Semantic Frames. We were able to achieve a reasonable accuracy using our best model. We found that our accuracy was majorly affected by the limitations in semantic role labeler, SEMAFOR. We believe that we can observe a significant improvement in the accuracy given better semantic role labels. Our conclusion is that semantic role labels does help for the task of Question Answering and new interesting approaches that leverage their power should definitely be encouraged.

## 2. Problem Statement

The specific problem we are attempting to solve is the task of answering several "Multiple Choice Questions" of the 4th grade science exam level. These questions are taken from the **NYSE 4th grade science exam**.

### Input

An input to our system is the set of Multiple Choice Questions. Each question has four options. Our dataset consists of a total of 120 questions and corresponding decomposed answers. Another input to our system is a set of several sentences from the Barron's corpus. There are a total of 1200 sentences. We manually annotated the correct answers for each of the questions.

**Output**
The output of our system is an answer that is scored the maximum among the given four options.

An example input describing the Multiple Choice Question would be as below:

*If an object is attracted to a magnet, it is most likely made of: A) wood B) plastic C) cardboard D) metal*
Correct input sentences for the above question from the Barron's corpus are : "A magnet is an object that attracts metals." , "The distance between two magnets or a magnet and another metal object has an effect on the force of attraction.".
The output would be the highest scored answer which would be "**metal**" in this case.

**Tools**
We used the tools nltk (python), WordNet, FrameNet (comparison of frames), SEMAFOR (for extracting semantic frames) and Porter's stemmer.

**Evaluation measure**
The percentage of questions with the correct predicted answers has been used as the evaluation measure.

**Overall Framework Architecture**

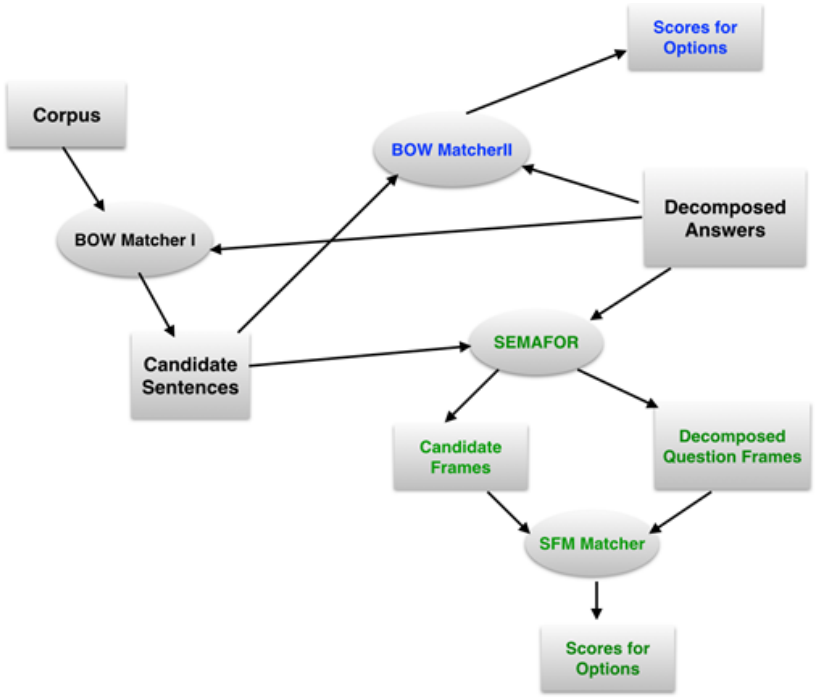



**Figure 1: Architecture of our Question Answering System**

In Figure 1, "Scores for Options" in blue indicate results obtained by using the baseline method and "Scores for Options" in green indicate results after using the extended approach of Semantic Frame Matching (SFM).

## 3. Baseline Model and Evaluation

For our baseline solution, we used Bag of Words matching for both candidate extraction and answer scoring. In **Figure 1** above, **BoW matcher I** takes as input, the decomposed answers (or question text) for the question and the sentences from the corpus and outputs the candidate sentences. BoW Matcher II takes decomposed answers and candidate sentences (extracted with BoW Matcher I) as inputs and gives the options with their respective scores as the output. Both the BoW matchers use the algorithm given below.

***BoW matcher***(words_sentence1, words_sentence2):

```
Update words_sentence1 and words_sentence2 after removing stop words
Stem and lemmatize words in  words_sentence1 and words_sentence2 and update
score = 0
For word1 in words_sentence1
        For word2 in words_sentence2
                score = score +  similarity(word1,word2)
return score
```

For the baseline, two methods have been used for candidate sentence extraction.

1. Inputs given to BoW MatcherI were sentences from the corpus and question text.
2. Inputs given to BoW Matcher II were sentences from the corpus and each of the decomposed answers individually. The final candidate sentences are the union of candidates extracted for each of the decomposed answers.

The example shown below gives the nature of candidate sentences extracted for both the methods.

*When plants and animals die, which organisms help return nutrients to the food chain ?*

*a. decomposers b. predators c. prey d. producers*

**Table 1 : Candidate Sentences for Method 1 and Method 2**

| Method 1 | Method 2 |
|---|---|
| Living things need to take in **nutrients** in the form of **food** which help **them** grow and create energy. | **Decomposers** are living things that break down dead organisms and recycle their **nutrients** into the soil. |
| ***Plants*** *and* ***animals*** *require air, water, light, and* ***nutrients*** *in order to live and survive.* | ***Plants*** *are called* ***producers*** *because they provide the* ***food*** *supply for themselves and animals.* |

For the question given above, the sentences in red that correspond to Method 1 do not go with the meaning of the question. After using Method 2 we get candidate sentences that are relevant to the question. When we use Method 1 we may miss out on the correct candidate sentences as incorrect sentences are ranked highly. Hence the accuracy is better if Method 2 (accuracy of 45.3% as opposed to 43.7% for Method 1) is used for candidate sentence extraction.

The example below shows a question which gives wrong candidate sentences when Method 2 alone was used, but gave correct candidates when the option words were weighted heavily.

*Which human activity most often has the most harmful effect on the environment ?*

*a. breathing b. growing c. planting d. polluting*

Even after using Method 2 we found incorrect candidate sentences for a few questions. So we slightly modified the Bag of Words matching algorithm to give more focus to the words in the options. This resulted in the extraction of better candidate sentences for those questions and hence a slightly better accuracy (46.28%). The example below shows a question which gave the wrong candidate sentences when Method 2 alone was used, but gave correct candidates when the option words were weighed heavily.

*Which human activity most often has the most harmful effect on the environment ?*

*a. breathing b. growing c. planting d. polluting*

**Table 2 : Candidate Sentences for BoW without Focus and with Focus**

| **BoW without focus** | **BoW with focus on options** |
|---|---|
| **Harm** to the **environment** can be caused when animals are displaced and the landscape is changed. | Building shelters and the various forms of transportation are some of the ways **humans** have created **pollution**. |
| ***Human*** *decisions and activities have had a major impact on the physical and living* ***environments****.* | *Plants and animals depend on each other and the nonliving* ***environment*** *for survival.* |

From Table 2 we observe that BoW with focus on options gives correct candidates. **For all the subsequent approaches Bag of Words with focus has been used for candidate extraction.** We also observe that extraction of right candidate sentences improves the overall accuracy.

The shortcoming of our baseline method is that it considers each sentence as a bag of independent words and BoW matcher does not take the meaning of words into account. For example, for the question below, though the right candidate sentence is extracted, the correct answer does not have the highest score because BoW does not score the word pair <length, size> highly.

*Which unit of measurement can be used to describe the length of a desk ?*

*a. degrees celsius b. grams c. litres d. centimeter*

The correct candidate sentence for the above question is “We can describe the paper in terms of its color (white) and its size (centimeter) ”. To capture the meaning of the sentences being compared we used an improved approach, “Semantic Frame Matching” described in Section 4.

Another shortcoming in the implementation of Bag of Words is that we used similarity between two words based on their synsets in WordNet. The synsets in WordNet do not accurately give the synonyms of all words. This also affects the performance of our baseline method.

## 4. Our Extensions and Evaluation

### 4.1 Extension Methodologies

The main theme behind all our extensions is to use the semantic role labels for matching the candidate sentences with the answers. This would take care of the issue of not taking the semantic information of the sentences in our baseline solution. Before we explain the three different extensions we had implemented, it is essential to understand what Semantic Frame Matching means. Any existing work based on Semantic Frame Matching used just the frame names. However, we perform a 2-level matching based on semantic frames and their inherent relations. We extract the semantic frame names from “SEMAFOR” [4]. Then, we obtain the corresponding “FrameNet” [5] frame elements using the frame names and extract all the frame relations. The second level of matching happens on this frame relations (such as “inheritance”, “using” etc). This leads to a more robust matching of the semantic structures between the answers and the candidate sentences. The

following sections explain in detail about the three different extensions, how each of them addresses a particular problem and their limitations.

**4.1.1 Semantic Frames Matching**

Once we have a good set of candidate sentences extracted using the BoW + focus approached described in the previous section, we match the semantic frames (generated using SEMAFOR) of each of the candidate sentence with the decomposed answers. This addition of the new semantic information would now result in a better match. This method gave us an accuracy of **53.67%** on our dataset. An example is illustrated below:
**Eg:-** *Which unit of measurement can be used to describe the length of a desk ?* would rightly predict "centimeters" as the answer.

However, there are still certain scenarios where this kind of matching isn't sufficient to generate the right answer. For instance, consider the below scenario:
**Eg:-** *Growing thicker fur in winter helps some animals to a)hide from danger b)attract a mate c)find food d)keep warm*
**Candidates:**
1)Animals grow thicker fur during winter season
2)A polar bear's fur keeps it warm in extreme temperatures.

Here, the number of frame matches with candidate 1 is more than that with candidate 2. So, the candidate 2 gets a low score despite having the right information for the question. However, a match on the frame representing "warm" is more important than any other match in this scenario. To address this issue, we have added a heuristic where we give an additional score whenever there is a semantic frame match on the focus word ("warm" here) between the candidate sentence and the answer. So, our next extension is to incorporate this new heuristic while matching.

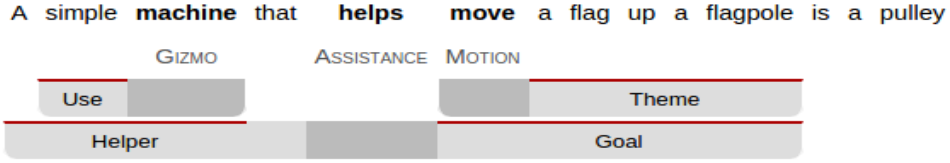
**4.1.2 Semantic Frames Matching with Focus**

The new heuristic to give an additional score to a Semantic Frame match on focus word helped solve the above scenarios. For instance, the answer is rightly predicted as "warm" in the above question. This method gave us an accuracy of **59.12%** on our dataset. However, we had noticed that in many scenarios, SEMAFOR produced either incorrect semantic labels or no semantic labels at all. For instance, consider the example below:

**Eg:-** *A simple machine that helps move a flag up a flagpole is a*
*(a) lever (b) pulley (c) inclined plane (d) bar magnet*

**Candidate** - A pulley is used to raise the flag up the flagpole at your school each day.

The semantic frames generated by SEMAFOR are as below:



Notice that there are no semantic frames generated for the important focus word "pulley". So, despite having good candidate sentences and a good semantic role matcher, we end up being limited by the lack of the semantic label. However, an interesting thing to note here is that there are a lot of "word matches" between the candidate sentence and the answer. So, a plain BoW matcher would predict the right answer in this case. This gave us the idea to incorporate some sort of heuristic that helps us switch between the BoW and semantic frame matching. So, our next extension is to add such a heuristic.

#### 4.1.3 Semantic Frame Matching + Confidence Heuristic

The heuristic that we have incorporated is an indication of a "**confidence**" measure for the top scored answer. For instance, consider we have scores of <4,4,2,2> for the four options using Semantic Frame Matching and scores of <1,4,2,0> for the four options using BoW matching. It is evident from the set of scores that BoW matcher is more "confident" than the Semantic matcher that second option is the right answer (as semantic matcher gave equal scores of 4 for both first and second options). So, it is more probable that the second option is the right answer. We use this confidence measure to switch between the BoW and semantic frame predictions picking the one with higher confidence. The confidence is calculated by dividing each score with the maximum value and then calculating the **variance** of the resulting numbers. Although helpful in certain scenarios, this heuristic didn't prove to be much effective in improving the results.

### 4.2 Other Error Scenarios

#### 4.2.1 Joint Inference problem

One of the issues that seemed to affect our accuracy is the problem of the inference from multiple sentences. This is illustrated with an example below:

**Eg:-** In New York State , the longest period of daylight occurs during (a) September (b) March (c ) December (d) June

Evidence 1 - The longest period of daylight hours occurs at the beginning of Summer.
Evidence 2 - Summer lasts from June 20 to September 21.

Here, none of the two sentences **individually** provide a direct evidence for the required answer. But a combination of the both certainly does. These kind of scenarios aren't addressed in our solution. We could probably generate the candidate sentences recursively to address this but the branching factor would be an important issue to consider in that scenario.

#### 4.2.2 Sentiment Association Problem

Another case that we have not considered is the sentiment of the decomposed answers and the corpus sentences for both candidate extraction and answer scoring. Since we are only comparing individual words in the first approach and frames corresponding to words in the second, questions like the one given below fail with our approaches.

*Which characteristic can a human offspring inherit?*

*a. facial scar b. blue eyes c. long hair d. broken leg*

**Candidate:**

Facial scar is **not** an inherited trait.

Hence it might be a useful idea to also match the sentiment of the question to improve the accuracy.
Another source of error that we observed was the lack of appropriate candidate sentences in the corpus.

## 5. Results

Table 3 shows the summary of the "**Accuracy**" results of each of our methods.

**Table 3 : Comparison of accuracies for different methods**

| **Candidate Sentence Extraction Method** | **Answer Scoring Method** | **Accuracy** |
|---|---|---|
| BoW without focus (**Baseline**) | BoW without focus | **45.3%** |
| BoW with focus | BoW without focus | 46.28% |
| BoW with focus | Semantic Frames without focus | 53.67% |
| BoW with focus (**Best Model**) | Semantic Frames with focus + other heuristics | **59.12%** |

In our best model, we had improved on the accuracy from our Baseline Model by **13.82%.**

## 6. Conclusions

The use of semantic role labels for Question Answering has not been extensively studied by the researchers. We conclude, from our observations and analysis, that the semantic role information is certainly valuable for the task of question answering. We observed a spike in the accuracy of about 13% when we incorporated the semantic role labels. This should definitely encourage researchers to consider semantic role labels more seriously. Especially, the task of accurate generation of semantic roles would be of prime importance given their benefits for the Question Answering task.

We also learnt valuable insights from this project. We realized that a limitation in one component would affect the functionality of the entire framework even though the other components are robust. For instance, a weak candidate sentence generator would result in poor results even though there is a robust semantic labeling. Similarly, an inaccurate Semantic Role Labeling would affect the accuracy even though there is a robust matching. The joint inference and the sentiment association problems made us realize that the accuracy could still get affected despite having robust candidate sentences, semantic frames and matching. Through all these experiences and analyses, we realized that Question Answering is indeed a hard and challenging task that requires an interplay between several components. Nevertheless, this challenge element is what makes Question Answering an exciting research area to pursue.

## 7. References

[1] Shen, Dan, and Mirella Lapata. "Using Semantic Roles to Improve Question Answering." EMNLP-CoNLL. 2007.

[2] K. Erk, S. Pado. 2006. Shalmaneser - a flexible toolbox ´for semantic role assignment. In Proceedings of the LREC, 527–532, Genoa, Italy.

[3] Das, Dipanjan, et al. "Probabilistic frame-semantic parsing." Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics. Association for Computational Linguistics, 2010.

[4] http://ark.cs.cmu.edu/SEMAFOR/

[5] C. J. Fillmore, C. R. Johnson, M. R. Petruck. 2003. Background to FrameNet. International Journal of Lexicography, 16:235–250