Automatic Synonym Generator Tool

Go through the document before testing it for understanding how it is designed

As we know, the wordnet(online dictionary from princeton university) db is organized in "synsets", so each word has several synonyms, and those synonyms in turn have their own synonyms, and so on. This arrangement can be used to dynamically generate a multiple choice question for each word. Here this will pick a random word from a attached file and generate a multiple choice question having five answer choices, of which one choice alone is correct (ie it is one of the valid synonyms of the given word) and the other four choices are randomly picked from entries in the db.

So far, so good. But there is a problem: What if, among the four "wrong" choices, there are words close to the main word? Maybe another direct synonym, or the synonym of a synonym? Thus we should define a metric, **"Minimal Path Length" or MPL** that tells us "how far apart two words are, in their meanings".

For Eg, the word "be" has a synonym "live", and "live" in turn has a synonym "endure", which in turn has a synonym, "suffer". So the words **<"be", "suffer"> have an MPL of 3** as you have to hop 3 synonyms to get to one from the other.

Of course, this is assuming that there are no common synonyms across this set (eg. "suffer" is not a synonym of "live") - that's what "Minimum" means, ie it's the shortest path.

So, use the above logic to ensure that your four wrong answer choices are at an MPL >= 4 from the main word.

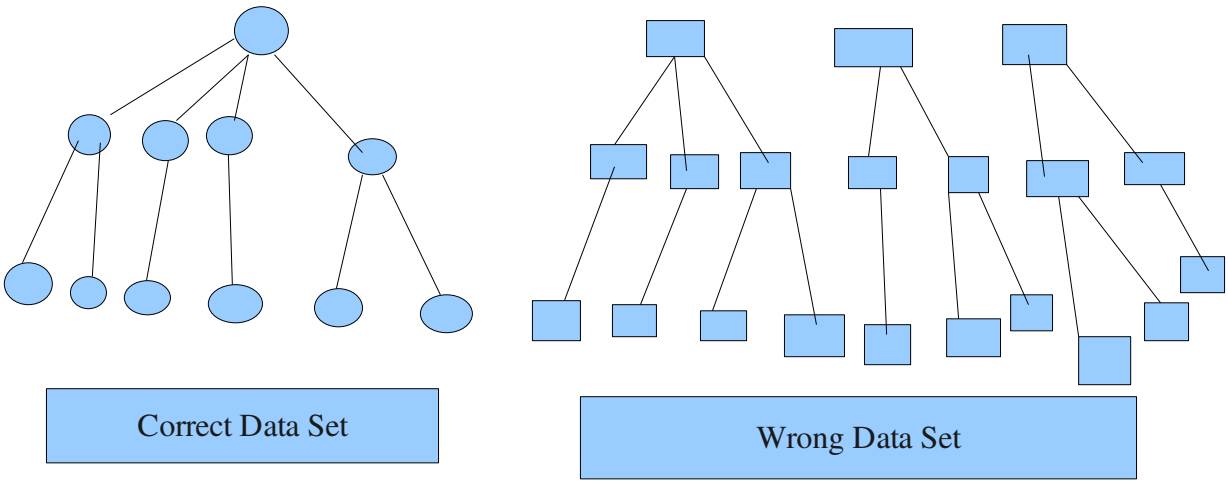
Now, for each question, the generated answer choices have 1 right choice and 4 very wrong choices. But an intelligent question will also have a "distracter" answer choice, ie one that's close to the right answer but is not the right answer. Ie 1 right choice, 1 distracter choice and 3 very wrong choices.

I think you have understood the scenario now. Okay. For achieving these things,

- First I will get the User input keyword which is selected randomly from the file.
- In wordnet db synsets are related with synset_id, synonyms are taken by mathcing the synset_id alone. So all the manipulation here are done using synset_id.
- Then I will find the child(synsets)nodes of that word. The word should not repeat on the child node also. It is manipulated using SQL query.
- Then for each synset in level 1(MPL==1) I will find its children and also it is checked with its parents for avoiding repitition.
- So, when we go for the second level there will be around 200 synsets in average. Now from the first level children we can get one correct answer for the question which is a direct synonym from the correct answer set. All those values in three levels will be considered as correct data set.
- Now our target is to find a three wrong answers with MPL>=4. To achieve this we have to go upto four levels in the correct answer set. Since it is a web based application we can't go by this way, because if we go upto 4 levels there will be thousands of synsets which is not possible to hold in memory.
- So this is achieved like this. Here we are selecting three random synsets using its ID from the table. After selecting each value it is compared with correct data set and also with

previous random values for avoiding repitition.

- By the way we are generating upto two levels of data in the wrong data set so that no value is repeated from the correct data set.



- So by doing this the parent nodes in wrong data set is with MPL of 4 with the user input keyword.
- So now we have obtained four answer choices. For the distractor answer choice we will select it from the second level of correct data set which will not be a direct synonym for the user keyword. It will be a synonym with MPL = 2. It won't be a direct synonym for the question since it is compared with prvious values . Otherwise we can have another method for distractor: We can get a synset using "LIKE" clause. For example for a keyword "advertise" we can give one of the option as "advert" using Like. But it doesn't work well for all the words. If there is no distractor means it is generated from random values like before.
- So we have got the solution for selecting five answer choices for the question.
- Ajax is used for selecting question.
- This is how this Automatic Synonym Generator Tool works.

Problems With this application:

- All the manipulations here are done with arrays. While running this script it takes almost 40MB - 50 MB in memory. It takes more memory for some words. Any way all the memory will be freed after execution.
- Query optimizations are done but still we have to optimize for reducing time complexity.
- It won't work for some words which doesnot contain any synonym. Eg: xenophobia
- Answer is checked with javascript .So answer can be viewed if you go through page source in your browser. By default answer is 4^{th} option always here.
- There is no session or cookie is applied here.

Future enhancements:

- Session/cookie must be applied.
- Correct answer should not be like the same user input word. It must be checked. Many type of algorithms should be used in selecting distractor depending on the question.
- Query must be optimized more and results should be arrived in a faster manner using ajax or

jquery.

- Memory consumption must be reduced for a faster browser interaction. Data structure can be changed to something other than arrays.
- Ajax auto suggestion must be organized in to a drop down list like Google suggestions.
- This should be conducted as a test with many questions. Answers should not be displayed in page source. It must be handled through database.

For any queries,

Karthikeyan NG
intrepidkarthi@gmail.com
www.intrepidkarthi.com
Cell: 9786836407