**Natural Language Processing  
Multiple Choice Automatic Alternate Choice Selection  
Andrew Ferlitsch**

**Abstract:** Hand generation of alternate answers (wrong choices) for multiple choice questions can be problematic. Given an existing corpus of questions and correct answers (e.g., 1000 Q/A), one needs to hand generate three alternate choices per question (e.g., 3000 alternate choices). Beyond the tediousness, there is also the question of quality. If the alternate choices are too far away, they will be obvious for elimination and if too close, they will be confusing. Therefore, the human generator must be highly skilled and experienced within the domain of the corpus.

In this paper, a method is discussed on using NLP techniques to auto-generate the alternate choices from the correct answers to other questions in the corpus that are similar.

***Experiment 1:***

In the first method tried, each question was processed into a bag of words. For bag of words, each question is processed to remove stop words (e.g., the, is, a, and, …), lowercased and punctuation removed, followed by stemming (i.e., find root form of word: reads, reading => read), and duplicate words eliminated. For example:

How many bytes are in a short => *byte, short*

For each question in the corpus, its bag of words were compared to the bag of words of all the remaining questions for the number of matches between the bag of words of the target question and each of the remaining questions. The results were then matched by count of matches. Each question had a unique id, so if question 4 had 3 matches, question 17 had 2 matches and question 6 had 1 match, the following result was produced, in JSON representation:

[ { id: 4, count: 3 }, { id: 17, count: 2 }, { id: 6, count: 1 } ]

At this point, the Q&A database consisted of 24 categories (corpuses) with a total of 1100 questions. Categories average 40+ questions were hand tested to subjectively observe the quality of the selection of alternate wrong choices (where three alternates were selected).

Each multiple choice generation was evaluated whether an alternate choice was out of context or not (‘the label’). Each choice was looked at subjectively as whether it was out of context (i.e., obvious it did not belong). No interpretation was made of the inverse (i.e., was subjectively within).

Using this subjective measurement, the quality of the choices was deemed marginal. In general, each question had one alternate choice that was consider not out of context (“true positive”). Occasionally it was observed a question with two alternate choices were considered not out of context. Rarely it was observed where all three alternate choices were not out of context. The result was subjectively deemed as 30/70 (30% acceptable choices, 70% not acceptable).

***Experiment 2:***

A spot analysis of the out of context alternate choices (“false positive”) showed a pattern of question pairs with some count matches, but the nature of the questions had answers that were very divergent from each other.

The method was modified to process each answer into a bag of words. For each question/answer, the bag of words of the answer was compared to the bag of words of the answer of each remaining question/answer in the category (“corpus”). The count matches of both the question and answer were then combined and sorted.

The results of the selected alternate choices of the multiple choice generation were again subjectively evaluated with the same criteria. It was deemed the quality of the alternate choices was better than marginal but not sufficient. In general, each question had two alternate choices that were consider not out of context (“true positive”). It was still rarely observed where all three alternate choices were not out of context. The result was subjectively deemed as 50/50 (50% acceptable choices, 50% not acceptable).

***Improvement to Experiment 2:***

It was occasionally observed that an alternate choice was identical to the true answer. For example, there were two different C# questions with the same answer:

{id: 1, ‘question’: ‘What is the byte length of a short?’, ‘answer’: ’16 bits’ }

{ id 2:, ‘question’: ‘What is the byte length of a char?’, ‘answer’: ’16 bits’ }

The bag of words reduction produced the following:

{ id: 1, ‘qbow’: ‘byte,length,short’, abow: ’16, bit’ }

{ id: 2, ‘qbow’: ‘byte,length,char’, abow: ’16, bit’ }

A similarity match between the two questions/answers produced:

{ id: 2, count: 4 }

A further rule was added to eliminate (“prune” similarity matches if the bag of word match of the answer is identical.

***Experiment 3:***

In this method, an attempt was made to find out how big a corpus needed to be to achieve a high quality of alternate choice matches. Two categories were chosen (C++ and C#). The question count in both corpuses was raised to 100.

The results of the selected alternate choices of the multiple choice generation were again subjectively evaluated with the same criteria. It was deemed the quality of the alternate choices was good, but not yet sufficient. About two-thirds of the questions had three alternate choices that were consider not out of context (“true positive”). The result was subjectively deemed as 75/25 (75% acceptable choices, 25% not acceptable).

It was guessimated that using this method would require a category size of 200 words to achieve a rate of 90/10.

***Experiment 4:***

In the prior methods, the questions/answers in a category were added ad-hoc. There was no prior plan or strategy. Questions/answers were added as they occurred to the author. The relationship between the questions, while not entirely independent, were slightly better than random.

It was postulated that if a category was written with a prior plan and strategy, that the quality rate of the alternate choices could be achieved at smaller corpus sizes.

In this method, two new categories, machine learning and artificial intelligence, were added with a plan and strategy. The plan was to predetermine the classes of questions/answers to add, and the strategy was for each class, to add four related but different answered questions.

The categories were evaluated first at 40 questions and then 100 questions. It was found, subjectively, that at 40 questions the quality was deemed slightly better than the quality of the categories without a plan/strategy (C++/C#) of 100 questions. At 100 questions, it was subjectively deemed that we achieved the goal quality of 90/10 (90% acceptable, 10% not acceptable).

***Conclusion:***

Using this method of bag of words and similarity matching will produce good alternate choices in multiple choice questions at substantially smaller category (“corpus”) sizes if a good plan and strategy is chosen prior to the generation of the questions/answers.

***Postulation:***

It would desired that the system could then learn to detect the presence of the remaining out of context (“false positive” alternate choices. It was initially postulated combining time metrics with A/B testing. A hypothesis is made that with a sufficient sampling base of users, that the time distribution of multiple choice questions with high quality alternate choices would fall into a normal equation distribution. The system could then monitor for multiple choice questions whose mean average time is too far to the left or right limits of the normal equation. Too far to the left, too many are out of context, and too far to the right, too many are too close in context.

For these, the system could use A/B testing techniques to randomly replace alternate choices with other unselected alternate choices until the mean average time falls into an acceptable range within the normal equation distribution.

It is perceived that this technique would require a substantially large sampling rate to find the normal equation time distribution per category, and then an equal or greater sampling rate of the A/B testing technique to learn a combination that brings the outlying multiple choice pairings within an acceptable range.

***Alternate Postulation:***

It is deemed that the previous postulation would not be desirable in time complexity, and another method of equal results in substantially lessor time complexity would be desired.

It is postulated that when a user selects the correct answer in a multiple choice question that it is preceded by two mental passes over the alternate choices. On the first pass, the user decides which choice is the most correct. The user then scans the alternate choices a second pass and uses the process of elimination of the choices they feel are incorrect. If both the first and second pass result in the same choice, they select it.

In an alternate representation of the multiple choice, instead of having radio boxes to select a single correct answer, each alternate choice would have a X box (elimination). The user would select alternate choices to eliminate until there was only one remaining choice – as the correct answer.

Under this representation, it is postulated that the user would eliminate not in chronological order by in most likelihood not correct order. The first choice eliminated would be the choice must obvious not correct, and so forth.

If this postulation is behavioral correct over most of the population, then a substantially smaller sample size (sooner) would detect which choices are too far out of context. In this case, the determination of A in A/B testing is deterministically determined instead of randomly determined. Likewise, the determination of B would be deterministic and not randomly determined.