TOWARD AUTOMATIC GENERATION OF MULTIPLE-CHOICE QUESTIONS IN AN EDUCATIONAL CONTEXT

BY

ZHENG KANG

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Advisor:

Professor ChengXiang Zhai

Abstract

Traditionally, textbooks often include thought-provoking questions at the beginning or at the end of a piece of text. These questions help readers to self-evaluate their comprehension of the text, and to reinforce their knowledge.

The Internet has given people access to much more information than ever available before. With it, we also witness the blossom of educational websites and articles online. While the emergence of these websites and content themselves are great for child and adult learners alike, most of them are not accompanied by such reviewing questions traditionally available in textbooks. Without these questions, readers cannot easily assess their understanding, and forget what they learn in a shorter time span. This makes for a less satisfactory learning experience.

It is impractical to generate such questions manually due to the sheer amount of articles available online. Therefore, we propose a method of generating these questions automatically without human intervention. We first select the most important sentences in an article. For each sentence, we automatically choose the most important phrase in it, and blank out that phrase. Finally, we attempt to create distracting answers for the phrase blanked out, making a multiple-choice question out of the sentence.

We believe this is a step forward to creating a more enjoyable learning environment for life-long learners and all Internet users.

To My Parents, My Friends, and Rifadel Schelder.

Acknowledgments

First order of gratitude goes to my academic advisor, Professor ChengXiang Zhai, who patiently helped me in many regards in and out of class, and without whose advising this thesis wouldn't have been possible. Thank you, Professor Zhai.

I would also like to thank my parents for their unconditional support throughout my study, who were always there for me during my most difficult times. Love you, mom. Love you, dad.

Table of Contents

Chapter 1

Introduction

1.1 Motivation

The Age of the Internet witnessed the exponential increase of the amount of information available to regular people, and with it, an increase in educational materials, often presented in revolutionary manners. From crowd-sourced encyclopedia, most notably Wikipedia, to manually compiled articles written by invited authors, such as HowStuffWorks.com, the Internet has truly become a perfect place for people to further their knowledge regardless of their age, job, or location.

What has not increased in quantity at as high a speed, are the questions that accompany a piece of text one often finds in old-style textbooks. These questions serve to emphasize on key information in the text and to test a reader's comprehension of the text, providing a convenient measure of assessment for self-learners, as well as for instructors in school. Traditionally, these questions are manually written by field experts who have read and reflected on the text in question. It is a time-consuming process and requires of the question-setters a level of competence that is not commonplace, which is why the number of questions hasn't been able to keep up with the amount of text available on the Internet.

In this thesis we propose a method of automated generation of such questions and a method of evaluating thees automatically generated questions, in hopes of improving people's learning experience and contributing to the developing trend of life-long learning.

1.2 Problem

To generate multiple-choice questions from an educational article, we break down the process into three steps:

1. Identifying the most important sentences in the article.
2. Identifying the most important phrases in each sentence identified in step 1.
3. Creating distracting answers for each phrase blanked out in step 2.

Many automatic summarization algorithms and tools exist to pick out important sentences in a piece of texts.[1][2][3]. Therefore, this problem is not extensively discussed in this thesis. Instead, we focus on the second and the third step of the process.

Since the questions are intended to accompany the text (either appearing before the text as "preview questions" or after the text as "review question"), we assume the reader has the ability of read the text immediately before or after reading the questions. This allows us to generate more questions, including those that need to be read in the context of the article.

An example. Given this article:

`We've all seen movies in which a character has a retinal scan to prove his or her identity before walking into a top-secret installation. That's an example of a biometric system. In general, biometrics is a collection of measures of human physiology and behavior. A biometric system could scan a person's fingerprint or analyze the way he or she types on a keyboard. The purpose of most biometric systems is to authenticate a person's claimed identity.`

We might identify this sentence as important:

`In general, biometrics is a collection of measures of human physiology and behavior.`

Within that sentence, we carve out a crucial piece of information:

`In general, biometrics is a collection of measures of ____.`

Finally, we might generate these phrases as potential options:

`a. human physiology and behavior (correct answer)`

`b. human activities and thinking`

`c. human conceptions and ideas`

1.3 Summary of Contribution

In order to solve the problem stated above, we address several challenges in natural language processing in an educational context. We need to capture the main idea of a text, and evaluate which sentences and phrases best explain those ideas. We propose a numerical measurement of relevancy in order to pick out the most useful phrase for a question. We then create a test set to evaluate the accuracy of our automatic generation. After defining an index to measure our work, we improve upon ourselves by tuning the parameters introduced in our formulae. We discover several challenges beyond our initial hypothesis, and we discuss how to address them.

With the methods discussed in this thesis, we can build a crude system that generates questions from

arbitrary articles found on the Internet (though it works best on educational articles). We are one step closer to producing high-quality multiple-choice questions that Internet readers could use to improve their learning experience.

Chapter 2

Literature Review

The broader topic of automatic question generation has been extensively studied. Since natural language can be intricate, most such studies sacrifice either the variety of questions they could ask, or the the variety of information source with which they can work with. Existing work on this topic therefore falls into two categories.

In the first category, people try to construct a structured knowledge base, in which they could understand the relation of all concepts contained within, thereby obtaining some flexibility of how they can produce the questions. One study[11] creates a concept map with relations such as 'has-a' and 'is-a'. In another[12], they tried to understand whether a word refers to a person, a place or an organization, etc., using various language clues.

In the second category, the authors try to build their questions on free-formed text, but restrict the generation of their questions to be based on patterns they know how. For example, one looks for a particular pattern in the syntactic tree[4], or named entities[5]. They provide great insights into our problem, but they address a different aspect of it and do not provide a solution to our problem.

Text summarization through identifying key sentences has been extensively studied[1][2][3] and provide the prerequisite tools we need for the experiments we carry out in this thesis.

Key phrase extraction has been studied on various corpora by multiple authors[13] and the methods proposed usually depend on a large body of text which we don't have available. That said, these methods provide an insight into the general problem, and we adapt some of these methods for our purposes.

To the best of our ability, we could not find existing work on generating distracting answers for a given blank in a given sentence.

Chapter 3

Problem Definition

In this thesis:

Text refers to one or more paragraphs on the same topic. **Title** refers to the title of the text. A piece of text does not include its title.

Question refers to a sentence with a phrase blanked out. On its own, it serves as a fill-in-the-blank question. **Distracting answers** refer to similar phrases that could fill in the blank of a question but provide the incorrect information. A question, the phrase carved out, and the distracting answers together make a **multiple-choice question**.

As stated in chapter 1, in this thesis we focus on (1) generating a question from a given sentence (by identifying the most important phrase in it), and (2) generating a multiple-choice question from a question (by discovering good distracting answers).

The first task takes as input a piece of text, its title, and an important sentence. It produces as output a question based on the sentence given.

The second task takes as input a question and the sentence it's based on. It produces as output several distracting answers.

Chapter 4

Technology Employed

In the pursuit of a solution to our problem, we made use of following tools and technology:

4.1 NLTK (Natural Language Toolkit)

NLTK[6] is a python module with many features in natural language processing. It has a straightforward API and is a great help in computational linguistics. We used NLTK for many tasks, including tokenization, stemming, and computing semantic similarity.

4.2 Stanford CoreNLP

Stanford CoreNLP[7] is the leading toolset in natural language analysis and is still actively developed by the Stanford NLP Group. It is written in Java, but many wrappers exist in other languages, including a few in Python. We used Stanford CoreNLP for part-of-speech (POS) tagging, syntactic parsing and more.

4.3 Word2Vec

Word2Vec[8] is an open source project released by Google that produces a vector representation of words learned given a corpus as input. It can be used to compute semantic distances between words and phrases, and answer questions like 'king' + 'woman' - 'man' = ? (answer: 'queen'). We used Word2Vec to discover similar words and phrases in the attempt to generate distracting answers.

4.4 PyTeaser

PyTeaser[9] is a python module that summarizes an article by selecting the most important sentences in it. It uses TextRank as its principal algorithm. PyTeaser could be used if a human being is not available to pick out sentences from a piece of text.

Chapter 5

Making a Blank

Given a piece of text, its title, and the key sentences extracted from the text, we're ready to identify the most important phrase in the sentence.

5.1 What's an Important Phrase?

An important phrase is one that carries the most information in a given sentence, or one that is the most relevant to the theme of the text. For example, in the sentence

`In general, biometrics is a collection of measures of human physiology and behavior.`

"human physiology" is more relevant to "in general", because "human physiology" gives to the reader new information about biometrics (theme of the article), while "in general" is simply a grammatical construct, a metadiscourse that facilitates the understanding of the article.

Thus, we aim to find such phrases that are most relevant to the theme of the text.

5.2 Identifying the Theme of the Text

Given a piece of text and its title, we wish to identify the theme of the text.

The title naturally gives us a big clue. After all, the title is usually the author's idea of what the text is about. Therefore, we propose consider all words in the title important and part of the theme. However, stop words like "the" and "of" don't contribute to this information, therefore is excluded. The rest of the words are identified as **theme words from the title**.

For example, in the title "What Is Boredom", the only theme word from it would be "boredom". In the title "How Will Biometrics Affect Our Privacy", the theme words from it are "biometrics", "affect", and "privacy".

The body of the text often provides additional clues. Important concepts tend to be repeated over and over, so words that keep appearing tend to be important. Again, since stop words don't contribute to this information, they're excluded from consideration.

Therefore, we first tokenize the text into words (with NLTK), then remove the stop words and count the frequency of appearance of remaining words. The five most frequently used words are identified as **theme words from the text**.

5.3 Parsing Candidate Sentences

Before identifying *important* phrases, we need to identify *valid* phrases as candidates. We do this because we want the questions to read normally (even with a blank), and if the blank spills across the border of a phrase, it won't read as nice and make as good a question.

For example, in the sentence

`In general, biometrics is a collection of measures of human physiology and behavior.`

This would not make a good question:

`In general, biometrics is a collection of measures of human ____ behavior.`

because "physiology and" is not a phrase that can stand alone, despite "physiology" being an important word in the sentence.

To do so, we parse the sentence into its syntactic tree. This is done with the help of Stanford CoreNLP package, which nicely produces a hierarchy of phrase boundaries and helpfully label their part of speech for us. We collect all phrases identified and consider them our candidate blanks, among which we will attempt to pick the best one.

Upon experimenting, we find that noun phrases (labeled "NP") and verb phrases (labeled "VP") work the best as potential blanks. Therefore we will restrict our candidate pool to NPs and VPs only.

5.4 Calculating Relevancy Score

We now have obtained the theme of the text and a list of candidate phrases to blank out. We wish to rank the candidates in a way that reflects their usefulness as a blank. As discussed above, this mostly equates to finding the phrase most relevant to the theme of the text. Since words are easier to process than phrases, we tokenize the phrases into words and calculate the relevancy score for each word independently.

5.4.1 Calculating the Similarity Between Two Words

We wish to measure the similarity between any two given words, in such a way that, "chess" and "game" would return a higher score than, say, "chess" and "book". To do so, we need to understand the semantic differences (or similarities) between two words, which we can do with the help of NLTK's WordNet corpus.

WordNet maintains a network of synonyms, hypernyms and meronyms for every word. The closer two words are on this graph, the more semantically related they are. Using NLTK's WordNet corpus, mostly its collection of cognitive synonyms ("synsets"), we could calculate the semantic similarity of two words as follows:

```
synsets1 = wordnet.synsets(word1)
synsets2 = wordnet.synsets(word2)
if not synsets1 or not synsets2:
    return 0
score =  max(x.wup_similarity(y) for (x, y) in itertools.product(synsets1, synsets2))
return score
```

...where "wup_similarity"[10] and is a value between 0 and 1, and measures how close the two words are on the graph.

5.4.2 Calculating the Relevancy Score of a Word

Given the theme words from the title, the theme words from the text, and a word, we wish to find out how relevant the word is to the theme words.

With the method discussed above, we could obtain the semantic similarity of the word in question to each of the theme words we have. Since the theme words could potentially cover a range of concepts, and relatedness to any such concept makes the word more relevant as a blank, we take the sum of these similarity measurements as a starting point.

Considering the title is the author's summarization of the text, and most likely the intended topic of the text, if a word is related to a word in the title, we are more confident that it is more relevant to the theme of the text. For this reason, all similarity measurements to words in the title are multiplied by a boost factor of a before the summation. a is a parameter that can be tuned, which we set to 3 and obtained good results.

However, if a word appears exactly as in the title, or is a derived form of a word in the title, the answer would be too obvious and such a blank will not make a good question. For example, in the article titled "How will biometrics affect our privacy", this question:

`In general, ____ is a collection of measures of human physiology and behavior.`

is painfully useless as the answer is obviously "biometrics". We wish to test the reader on concepts related to the theme, not on what the theme is, therefore if a word shares a stem with a word in the title, we reduce its relevancy score to zero.

In summary, we calculate the relevancy score of a word as follows:

```
stemmer = SnowballStemmer('english')
title_key_words_stemmed = [stemmer.stem(w) for w in title_key_words]
if stemmer.stem(word) in title_key_words_stemmed:
    return 0

score = 0
for key_word in text_key_words:
    score += similarity(word, key_word)
for key_word in  title_key_words:
    score += a * similarity(word, key_word)
return score
```

5.4.3 Calculating the Relevancy Score of a Phrase

For each valid (candidate) phrase we obtained from the parse, we now wish to calculate its relevancy score for ranking purposes.

A naive solution would be to sum up the relevancy scores of each individual word in the phrase. However, long phrases will have an advantage simply by having more words in them and thus a better chance to score higher. Moreover, the highest scoring phrase would always be the entire sentence, which does not help our cause. We need to penalize long phrases to balance out this unfair advantage.

However, some phrases are good candidates precisely because they contain more keywords. Again, in the sentence,

`In general, biometrics is a collection of measures of human physiology and behavior.`

"human physiology and behavior" would make a better blank than "behavior" or "physiology" because the longer phrase covers two important concepts instead of one. If we simply use "relevancy score per word" as a measurement, the longer phrase would lose because the less useful words would drag down the score.

We therefore propose this formula for calculating the relevancy score of a phrase, given the relevancy score of each of its component words:

$$\text{relevancy score} = \frac{\sum \text{relevancy score of each word}}{1 + \frac{\text{length of phrase - 1}}{b}} \tag{5.1}$$

...where "length of phrase" is the number of tokenized words in the phrase, and b is a parameter. When b is set to 1, it's equivalent to the per-word average score. When b is infinity, it's equivalent to summing up

the score of each word. Between 1 and infinity, the higher b is, the less penalty we give long phrases. We set b to 5 and obtained good results.

5.5 Blanking Out a Phrase

We calculate the relevancy score of each candidate phrase and rank them. We blank out the top choice to generate a question.

5.6 Limitations

This method works generally well, but there are certain limitations. Two of them stand out:

1. Sometimes the most important word in a sentence is either a rare word or a proper noun, which appears exactly once and is not related the theme words as measured by WordNet (because its usage is rare). For example, in this sentence:

```
When we experience joy and excitement in a new situation, a chemical messenger,
or neurotransmitter, called dopamine, triggers that response in our brains.
```

"dopamine" would be a great blank to make, yet because it's neither a theme word (appearing only once), nor related to any theme word (WordNet limitation), it scores a zero in relevancy.

2. Natural language is intricate. Sometimes authors use metaphors and references which can be difficult to capture. In one article, the author mentioned two ways of carrying out a certain task, calling one the "high road" and the other the "low road". He then kept mentioning the "high road" and the "low road" throughout the article, making our algorithm think both phrases are important. While it's not incorrect, the questions would be better if we could resolve the reference and substitute the words "high road" with the actual concept referred to.

Chapter 6

Evaluating a Blank

To evaluate the accuracy of the blanks generated, we've compiled a test corpus of 6 articles and prepared questions manually.

6.1 Criteria for a Blank

When carving out the blanks manually, two central criteria are borne in mind:

1. The phrase blanked out should be central to the sentence's meaning;
2. The phrase blanked out should give the reader a piece of information they most likely did not know before reading the article.

We've selected 5 sentences from each article and blanked out the most important phrases in each sentence according to the principles above stated, making a total of 30 questions. When multiple phrases work equally well in a sentence, all are included as best choices of blanks. These manually created questions are considered the golden standard.

We've included the test set in the appendices for reference.

6.2 Calculating Overlap Score

We feed both the articles and the selected sentences into our system, which then produces 30 questions. We now have 30 automatically generated questions, and 30 corresponding manually generated questions, and we wish to evaluate how close our generated questions are to the manually created ones, quantitively measuring the accuracy of our generated questions.

For each sentence, we calculate the overlap score with this formula:

$$\text{overlap score} = \frac{\text{number of tokenized words carved out in both questions}}{\text{number of tokenized words carved out in either questions}}$$

We believe this is a fair evaluation because, if an automatically generated question produces a blank that

overlaps exactly with a manually created question, the overlap score is 1. If the automatically generated blank completely misses the manually generated blank, the overlap score is 0. If the automatically generated blank overlaps partially with the manually generated blank, the overlap score is between 0 and 1, depending on how much of a match there is between the two blanks.

When multiple versions of best answers are provided, the highest score is considered to be the overlap score.

The "overall overlap score" for the entire test set is the average overlap score of all questions, i.e. the sum of each individual overlap score divided by the number of questions in the test set, in our case 30.

6.3 Evaluation

Using the overall overlap score as defined above as a measurement of accuracy, we ran our algorithm on the test set.

Naively, with parameters $a = 1$ and $b = 1$, a base score of .327 was achieved.

After tuning the parameters, we found that $a = 3$ and $b = 5$ work best on our test set, with an overall overlap score of .506 achieved. That is, our automatically generated questions match the human-created ones over half the time.

The interaction of the two parameters and their effect on the overlap score are shown in this table:

	a = 1	a = 2	a = 3	a = 4
b = 1	.327	.332	.321	.321
b = 2	.416	.424	.446	.446
b = 3	.472	.472	.461	.477
b = 4	.463	.476	.475	.442
b = 5	.494	.497	**.506**	**.506**
b = 6	.472	.487	.470	.478

6.4 Sample Results

Good results:

```
A check card is a quick and easy alternative to handwritten checks.
A check card is a quick and easy alternative to ____.
```

(100% match)

```
In general, biometrics is a collection of measures of human physiology and behavior.
In general, biometrics is a collection of ____. (generated question)
In general, biometrics is a collection of measures of ____. (closest match)
```

(66.7% match. The word 'measures' scored just enough to overcome the long phrase penalty.)

Bad results:

```
As mentioned earlier, boredom can arise from both external and internal stimuli, muddying
    the answer to that question.
As mentioned earlier , boredom can arise from both external and internal stimuli, ____.
```

(0% match. Bad question from a structural point of view too — reader doesn't know what kind of an answer is expected.)

```
By the later 1700s, the scientific community was beginning to get a clearer picture of
    how electricity worked.
By the later 1700s , the scientific community was beginning to get ____ of how
    electricity worked .
```

(0% match. A generic phrase is identified as important.)

In future work, grammatical constructs should be taken into account when analyzing the importance, since certain structures are usually used to introduce new knowledge, while others are often used to clarify existing knowledge. In addition, a list of common phrases that's typically used in educational texts should be compiled, so that phrases such as 'a clearer picture' could be excluded from the candidates.

Chapter 7

On Generating Distracting Answers

One of the ideas we played with, now that we have a method of generating fill-in-the-blank style questions, was to automatically generate phrases that fit in the blank but provide the incorrect information, i.e. to generate distracting (wrong) answers. We were unable to produce useful results for this task, and the experiments we carried out are discussed below.

7.1 Proposed Method

We're given a sentence, and the phrase carved out for the question. Our goal is to produce phrases that are similar to the phrase carved out, in such a way that, when mixed in random order, it is not obvious which phrase belongs to the original sentence. A reader who is not familiar with the topics discussed in the text should believe all choices are plausible.

To generate such phrases, we look for phrases that occur in a similar context in the English language, that is, they often precede and succeed the same words. Phrases that occur in a similar context could usually substitute one another without breaking the grammatical construct, and they're often semantically related.

However, many of our blanks are phrases, and many of them probably put together in that order for the first time in history. Seeing it's difficult to analyze phrases as a whole, we break them down into words and consider each word individually.

Google's open source tool Word2Vec allows us to obtain a list of words and phrases used in similar contexts in real life. The model we experimented with was pre-trained on Google News articles, containing billions of words. With each word, we choose the best candidate, and we reassemble the best candidate from each word to form a new phrase.

7.2 Hypothesis

We had thought that words that appear in a similar context should provide a good list of substitutes we could use to generate distracting answers. For example, "red" should appear in similar contexts with

"blue", "green", etc., and "table" should appear in similar contexts with "chair", "desk", etc., allowing us to substitute "red table" with "blue chair" or "green desk".

7.3 Actual Result and Problems

While our anticipation was not incorrect, it turned out that "words that appear in a similar context" provides a very weak signal as to whether those words are a good candidate for a distracting answer.

One of the problems was in sense disambiguation, which is practically non existent in Work2Vec. While "red" provided "orange" as a similar word, "green" is most associated with concepts such as "eco friendly", which is definitely not what we're looking for.

Another problem was the noise in the data. Trained on real-world texts, the model sometimes behaves in a bizarre banner and producing, as top answers, words that don't make much sense. For example, the "most similar" phrase to "green" was "wearin o", and a candidate produced for the word "president" was "daunting platoni", which are borderline gibberish and unfit for a distracting answer.

A third and final problem, and probably the most important one, was that all kinds of words could appear in similar contexts. While words we're looking for — words with similar but *different* meanings — do appear in similar contexts, so do many other words. For example, similar words to "president" included synonyms such as "chairman" and "chief executive", which in many cases would fit the original blank and provide the *correct* information. For another example, similar words to "engineering" included meronyms (words on a lower level) such as "mechanical engineering" and "bioprocess engineering", while the best choices for a distracting answer are often words on the same level such as "design" or "art".

Much work still needs to be done in the future to further address this problem. In particular, it's worth exploring how to computationally capture the idea of "just different enough", i.e. phrases that refer to different concepts, but fall into the same category. Even a method of evaluating two given phrases for their just-different-enough-ness would be helpful. Another area worth looking into is noise canceling. It would be helpful to exclude gibberish from the list of candidates. Finally, a good method of telling whether the phrases are on the "same level" is needed. Natural language is tricky, and words don't fall neatly into buckets, but if we could tell whether two phrases are parallel to each other, or one encompasses the other, we'd have a good method of excluding many unreasonable answers.

Chapter 8

Conclusions and Future Work

In the age of the Internet, self-learning has become more possible and less costly than ever before. While the amount of information soared, we hardly see accompanying questions with these texts. It is in our belief that if such questions existed, they would greatly improve readers' learning experiences.

In this thesis we proposed a method for automatically generating questions for educational articles. We introduced a three step process: (1) identifying important sentences, (2) identifying important phrases, and (3) creating distracting answers. We focused on the last two tasks in this thesis.

For task (2), we discussed how to capture the theme of a piece of text from itself and its title, and how to calculate the relevancy of a phrase to such themes, in preparation for choosing the best phrase for a fill-in-the-blank question. We then created a test set with manually created questions, and proposed a measurement — "overlap score" — that measures how well we've generated these questions automatically.

The overall overlap score we achieved in this thesis was about 0.5, which, while significant for a first step, is not ready for real-life usage. Further improvement on the accuracy of automatically generated questions will be useful. We discussed some of the limitations of the current method, and what we can do to make it better.

For task (3), we discussed our sub-optimal results, and identified several challenges we need to address before we can get decent results. Words and phrases that work as a distracting answer must fit an extensive checklist of requirements, which, as we've discussed, is difficult to capture computationally. Other tools, and deeper analysis of the text is needed to solve this problem.

One problem that remains to be tackled is the selection of sentences, which in this thesis we relied on either humans or existing tools. Natural language articles make use of pronouns often, and the most important information could come from sentences such as "and that's why the mission failed", where the word "that" refers to an unknown concept or object when standing alone. It is important to either recognize which sentences, when combined, form an independent paragraph that's meaningful on its own, or resolve the reference and substitute it when generating the question.

With the methods discussed in this thesis, we take a first step toward generating high-quality multiple-

choice and fill-in-the-blank questions automatically from the educational materials freely available on the Internet today. We believe these questions would help life-long learners reinforce their knowledge, and create a better learning experience for children and adults alike.

Appendix A: Texts Used for Evaluation

All these texts are taken and adapted from HowStuffWorks.com.

1. Do I really need a bank account?

You may be fed up with overdraft charges, ATM fees and inconvenient bank hours. Sometimes keeping money in the bank seems like more trouble than it's worth. Nevertheless, most people assume bank accounts are essential. But why is this? Maybe the only thing preventing you from taking your money out is the fear of looking like that paranoid old granny who still keeps all her money under her mattress. Or maybe you pride yourself on being an independent thinker who never gives in to social customs that don't make sense to you. Let's see if having a bank account is all it's cracked up to be.

Why should you hand over your hard-earned money to a bank? Your money is protected there. The Federal Deposit Insurance Corporation (FDIC) insures your money up to $100,000. Savings accounts and some checking accounts offer another perk – you can earn interest on the money you deposit. Your mattress can't make such promises.

Some people find managing finances easier with a bank account. Looking at your bank statement makes creating a budget easier. Bank accounts also make getting paid simpler. You can arrange for your employer to direct deposit your paycheck automatically into your bank account.

As technology advances, the advantages of having a bank account grow. A check card is a quick and easy alternative to handwritten checks. With online banking, you can manage your money by keeping a real-time watch on your funds versus a monthly bank statement.

More and more people use their bank's online bill pay to pay anyone from credit card and utility companies to their local church. An automatic bill pay tool sends weekly or monthly payments so you never have to worry about forgetting. However, some people recommend using this tool only for payments that consistently stay the same amount like mortgage or car loan payments – not credit card bills.

Nevertheless, bank accounts aren't mandatory, and they're not the only smart place to put your money.

2. How will biometrics affect our privacy?

We've all seen movies in which a character has a retinal scan to prove his or her identity before walking into a top-secret installation. That's an example of a biometric system. In general, biometrics is a collection of measures of human physiology and behavior. A biometric system could scan a person's fingerprint or analyze the way he or she types on a keyboard. The purpose of most biometric systems is to authenticate a person's claimed identity.

Biometrics tend to be more convenient than other methods of identity authentication. You might forget your ID at home when you head out the door, but you'll still be able to use biometric devices. Imagine verifying your identity while at the store by swiping your finger across a sensor.

But along with convenience and security comes a concern for privacy. For biometrics to work, there needs to be a database containing the relevant information for each individual authorized by the system. For example, at that top-secret installation, every employee's biometric signature would have to be recorded so that the scanners could verify each person's identity.

This might not present much of a problem on its own. If the only data the system stores relates to the actual biometric measurements, privacy violations are at a minimum. But by their very nature, biometric systems collect more information than just the users' fingerprints, retinal patterns or other biometric data. At a basic level, most systems will record when and where a person is at the time of a scan.

3. What is boredom?

Although references to the idea of boredom stretch back to the Greek philosophers, the word did not enter the written English language until 1766. Afterward, literature exploded with musing on it, including works by Kierkegaard, Dostoyevsky and Tolstoy, who called boredom "the desire for desires."

Everyone knows what boredom feels like, but even after hundreds of years of identifying boredom as a plague upon life, no scientific consensus exists of what exactly it is. One reason lies in rooting out the source of boredom, akin to the clichd "chicken or the egg" question. As mentioned earlier, boredom can arise from both external and internal stimuli, muddying the answer to that question.

Scientists do know something about brain activity in high-risk, boredom-prone people. When we experience joy and excitement in a new situation, a chemical messenger, or neurotransmitter, called dopamine, triggers that response in our brains. It appears that high-risk, boredom-prone people may have naturally lower levels of dopamine, meaning that they require a heightened sense of novelty to stimulate their brains. In this light, boredom may serve as the lackluster yin to our yang of excitement and pleasure.

Although the part of our brain controlling the boredom response remains unclear, patients with damage

to their frontal cortex experience greater risk-taking urges along with boredom proneness. Interestingly, the frontal cortex also controls our perception of time, which could be linked to the sensation of time passing more slowly when we're bored.

How can we combat this elusive pest? A study found that people who reported feelings of boredom more frequently tried to alleviate it with brief distractions including work breaks or doing laundry. But these boredom Band-Aids soon failed. On the other hand, people who meditated, engaged with other people or accepted the boredom were more successful.

Likewise, finding new interests or hobbies, physical exercise and mindfulness have all been shown to reduce boredom. One study of teenagers found that those with strong interests had significantly higher self-esteem and overall well-being than bored ones.

When searching for an activity, psychologists recommend finding an optimal amount of ease and challenge, called flow. In essence, flow means getting into a groove, like a runner's high or hitting a tennis ball back and forth. It demands more skill and agility than tedious tasks, but at a low enough intensity that you reap the mental reward of accomplishment.

4. How Electricity Works – Electrostatics and Coulomb's Law

Even though they didn't fully understand it, ancient people knew about electricity. Thales of Miletus, a Greek philosopher known as one of the legendary Seven Wise Men, may have been the first human to study electricity, circa 600 B.C. By rubbing amber – fossilized tree resin – with fur, he was able to attract dust, feathers and other lightweight objects. These were the first experiments with electrostatics, the study of stationary electric charges or static electricity. In fact, the word electricity comes from the Greek elektron, which means amber.

The experiments wouldn't continue until the 17th century. That's when William Gilbert, an English physician and amateur scientist, began to study magnetism and static electricity. He repeated the research of Thales of Miletus, rubbing objects together and charging them by friction. When one object attracted or repelled the other, he coined the term "electric" to describe the forces at work. He said these forces developed because the rubbing action removed a fluid, or "humour," from one of the objects, leaving an "effluvium," or atmosphere, around it.

This concept – that electricity existed as a fluid – persisted into the 1700s. In 1729, English scientist Stephen Gray observed that certain materials, such as silk, didn't conduct electricity. His explanation was that the mysterious fluid described by Gilbert could travel through objects or be hampered from traveling. Scientists even built jars to hold this fluid and study its effects. The Dutch instrument makers Ewald von

Kleist and Pieter van Musschenbroek created what is now known as a Leyden jar, a glass jar containing water and a nail that could store an electrical charge. The first time Musschenbroek used the jar, he received a massive shock.

By the later 1700s, the scientific community was beginning to get a clearer picture of how electricity worked. Benjamin Franklin ran his famous kite experiment in 1752, proving that lightning was electrical in nature. He also presented the idea that electricity had positive and negative elements and that the flow was from positive to negative. Approximately 30 years later, a French scientist by the name of Charles Augustin de Coulomb conducted several experiments to determine the variables affecting an electrical force. His work resulted in Coulomb's law, which states that like charges repel and opposite charges attract, with a force proportional to the product of the charges and inversely proportional to the square of the distance between them.

Coulomb's law made it possible to calculate the electrostatic force between any two charged objects, but it didn't reveal the fundamental nature of those charges.

5. What did Abraham Lincoln invent?

Well before becoming the 16th president of the United States, the young Abraham Lincoln was known for his interest in engineering and mechanics. A childhood centered on agriculture suited Lincoln's curiosity well; he loved the culture of designing and inventing new objects, especially anything that had the potential to improve or refine the efficiency of labor. Later in life, he expressed a belief that an inventor should have exclusive rights to his design for a period of time after completion. This, he said, might inspire more people to invent solutions to their problems.

It's fitting, then, that Abraham Lincoln became the first – and so far, only – U.S. president to gain official recognition as an inventor by being granted a patent. His invention, a device intended to help boats navigate shallows, was the result of an adolescence spent boating along the rivers of the Midwest.

As a teen, Lincoln used his river navigation skills to explore the Ohio and Mississippi rivers. As a young adult, he worked on the crew of several cargo ships, moving goods to New Orleans down the Mississippi River. His skills and intuition were essential on one such trip, when the boat was damaged after it ran aground on a shallow. Lincoln quickly led the effort to shift cargo, drain water and move the boat along without capsizing. This experience, and others, led to Lincoln's interest in improving the technologies and resources available to the shipping and boating industries.

Entering adulthood, Lincoln decided to pursue a career in politics, but even after years of political experience and achievements, he couldn't shake the adventures of his youth. Still motivated by rivers, boats

and all things mechanical, his future forays down United States waterways would prove fateful.

In 1848, then-Congressman Abraham Lincoln was inspired by a riverboat that ran aground on a sandbar. In an attempt to get back afloat, the captain ordered his crew to place supplies – anything that could float and support weight, like empty cargo containers – underneath the ship in an effort to lift it off its shallow spot. Accounts vary as to whether or not Lincoln was actually aboard this boat or merely a witness to the incident, but historians agree that it was a catalyst to Lincoln's imminent brainstorm. He spent about a year in between sessions of Congress developing a solution to this common river navigation scenario.

Patent 6469 was awarded to Abraham Lincoln on May 22, 1849. Called "Buoying Vessels Over Shoals," Lincoln envisioned a system of waterproof fabric bladders that could be inflated when necessary to help ease a stuck ship over such obstacles. When crew members knew their ship was stuck, or at risk of hitting a shallow, Lincoln's invention could be activated, which would inflate the air chambers along the bottom of the watercraft to lift it above the water's surface, providing enough clearance to avoid a disaster. As part of the research process, Lincoln designed a scale model of a ship outfitted with the device. This model (built and assembled with the assistance of a Springfield, Ill., mechanic named Walter Davis) is on display at the Smithsonian Institution.

Lincoln was convinced he'd made a great contribution to the boating and shipping industries. During his frequent travels as a politician and public speaker, he even made reference to his invention on a few occasions. To his disappointment, though, there is no record of "Buoying Vessels Over Shoals" being fitted to a watercraft (even for testing or development purposes), and the system was never manufactured. Though "Buoying Vessels Over Shoals" was never built, and no profit was ever made, Lincoln made history as the only president to hold a patent for an invention.

6. Do all major networks have to carry presidential addresses?

When it's time for the leader of the free world to speak to the nation, the best way to get him in front of the most people is via television. But are the networks required to air every presidential address? Well, it's more complicated than a simple "yes" or "no" answer.

While there's more than one way for a president to get television airtime, telling him to take a hike when the office has made a direct request to broadcasters is embarrassing for everyone involved. Networks look unpatriotic and the president can appear weak. On the other hand, saying yes and providing free airtime can cost millions in lost advertising revenue to networks.

In a typical East Room address, the president takes questions from the press corps on a particular topic the administration would like to address. In each case, networks and their news companies decide whether

the issue being addressed is newsworthy, and can broadcast or report on it at their discretion. However, in cases where the president would like to speak to the nation directly, the White House will specifically ask networks to set aside time for the address, usually in the evening when more people are watching television.

It's generally considered good form to agree to a president's official request, but this is where the issue gets murky for whether or not networks must consent. Part of the deal that local broadcast stations make with the Federal Communications Commission is they must prove they are performing some public service in order to continue operating. This includes showing a minimum amount of children's educational programming, as well as agreeing to disseminate information that contributes to public wellbeing, including presidential addresses. When broadcasters fail to uphold their end of the bargain, the FCC can subject them to fines or revoke their licenses.

However, this arrangement doesn't always work well for the networks. While a presidential address is always by definition newsworthy, it almost always airs during primetime when networks have their biggest audiences. Agreeing to the president's request could means a loss of advertising revenue during the most profitable time of day, and sometimes networks choose to protect their bottom line instead.

Networks rarely refuse presidential requests, but it does happen and may even be increasingly common. In 2009 for example, Fox declined to show one of Barack Obama's news conference during primetime, citing that the network would lose too much money in lost advertising revenue. In 2001, the network made the same decision for one of George W. Bush's speeches, despite the White House's request. In 2014, ABC, NBC and CBS all declined to air Obama's November 2014 speech on immigration reform, though none would comment on why.

Appendix B: Test Set for Evaluation

These are sentences selected from the texts in Appendix A, and the manually created blanks according the criteria discussed in the thesis. If multiple questions work equally well, multiple questions are listed.

1. Do I really need a bank account?

```
- The Federal Deposit Insurance Corporation (FDIC) insures your money up to $100,000.
The ____ insures your money up to $100,000.
The ____ (FDIC) insures your money up to $100,000.
The Federal Deposit Insurance Corporation (FDIC) insures your money up to ____.

- Savings accounts and some checking accounts offer another perk -- you can earn interest
    on the money you deposit.
Savings accounts and some checking accounts offer another perk -- you can earn ____ on
    the money you deposit.
Savings accounts and some checking accounts offer another perk -- you can earn interest
    on ____.

- A check card is a quick and easy alternative to handwritten checks.
A check card is a ____ to handwritten checks.
A ____ is a quick and easy alternative to handwritten checks.
____ is a quick and easy alternative to handwritten checks.
A check card is a quick and easy alternative to ____.

- More and more people use their bank's online bill pay to pay anyone from credit card
    and utility companies to their local church.
More and more people use their bank's ____ to pay anyone from credit card and utility
    companies to their local church.
```

More and more people use their bank's online bill pay to pay anyone from ____ and utility companies to their local church.
More and more people use their bank's online bill pay to pay anyone from credit card and ____ to their local church.

- Bank accounts aren't mandatory, and they're not the only smart place to put your money.
Bank accounts aren't ____, and they're not the only smart place to put your money.
Bank accounts aren't mandatory, and they're not the only ____ to put your money.

2. How will biometrics affect our privacy?

- In general, biometrics is a collection of measures of human physiology and behavior.
In general, biometrics is a collection of measures of ____.

- A biometric system could scan a person's fingerprint or analyze the way he or she types on a keyboard.
A biometric system could scan ____ or analyze the way he or she types on a keyboard.
A biometric system could scan a person's fingerprint or analyze the way ____.

- The purpose of most biometric systems is to authenticate a person's claimed identity.
The purpose of most biometric systems is to authenticate ____.

- For biometrics to work, there needs to be a database containing the relevant information for each individual authorized by the system.
For biometrics to work, there needs to be a database containing ____ for each individual authorized by the system.
For biometrics to work, there needs to be a database containing the relevant information for ____.

- At a basic level, most systems will record when and where a person is at the time of a scan.
At a basic level, most systems will record ____ at the time of a scan.

At a basic level, most systems will record when and where a person is at ____.

3. What is boredom?

- Although references to the idea of boredom stretch back to the Greek philosophers, the word did not enter the written English language until 1766.

Although references to the idea of boredom stretch back to the ____, the word did not enter the written English language until 1766.

Although references to the idea of boredom stretch back to the Greek philosophers, the word did not enter the written English language until ____.

- As mentioned earlier, boredom can arise from both external and internal stimuli, muddying the answer to that question.

As mentioned earlier, boredom can arise from ____, muddying the answer to that question.

- It appears that high-risk, boredom-prone people may have naturally lower levels of dopamine.

It appears that ____ may have naturally lower levels of dopamine.

It appears that high-risk, boredom-prone people may have ____.

- A study found that people who reported feelings of boredom more frequently tried to alleviate it with brief distractions including work breaks or doing laundry.

A study found that people who reported feelings of boredom more frequently tried to alleviate it with ____ including work breaks or doing laundry.

- Likewise, finding new interests or hobbies, physical exercise and mindfulness have all been shown to reduce boredom.

Likewise, ____, physical exercise and mindfulness have all been shown to reduce boredom.

Likewise, finding new interests or hobbies, ____ and mindfulness have all been shown to reduce boredom.

Likewise, finding new interests or hobbies, physical exercise and ____ have all been shown to reduce boredom.

4. How Electricity Works Electrostatics and Coulombs Law

- In fact, the word electricity comes from the Greek elektron, which means amber.

In fact, the word electricity comes from the Greek ____, which means amber.

In fact, the word electricity comes from the Greek elektron, which means ____.

- When one object attracted or repelled the other, he coined the term "electric" to describe the forces at work.

When one object ____, he coined the term "electric" to describe the forces at work.

When one object attracted or repelled the other, he coined the term ____ to describe the forces at work.

When one object attracted or repelled the other, he coined the term "electric" to describe ____.

- In 1729, English scientist Stephen Gray observed that certain materials, such as silk, didn't conduct electricity.

In 1729, English scientist Stephen Gray observed that certain materials, such as ____, didn't conduct electricity.

In 1729, English scientist Stephen Gray observed that certain materials, such as silk, didn't ____.

- By the later 1700s, the scientific community was beginning to get a clearer picture of how electricity worked.

By the later 1700s, the scientific community was beginning to get a clearer picture of ____.

- Coulomb's law made it possible to calculate the electrostatic force between any two charged objects.

Coulomb's law made it possible to calculate the ____ between any two charged objects.

Coulomb's law made it possible to calculate the electrostatic force between ____.

5. What did Abraham Lincoln invent?

- As a teen, Lincoln used his river navigation skills to explore the Ohio and Mississippi rivers.

As a teen, Lincoln used his ____ to explore the Ohio and Mississippi rivers.

As a teen, Lincoln used his river navigation skills to explore ____.

- Well before becoming the 16th president of the United States, the young Abraham Lincoln was known for his interest in engineering and mechanics.

Well before becoming ____, the young Abraham Lincoln was known for his interest in engineering and mechanics.

Well before becoming the 16th president of the United States, the young Abraham Lincoln was known for his interest in ____.

- Later in life, he expressed a belief that an inventor should have exclusive rights to his design for a period of time after completion.

Later in life, he expressed a belief that an inventor should have ____ to his design for a period of time after completion.

Later in life, he expressed a belief that an inventor should have exclusive rights to his design for ____ after completion.

Later in life, he expressed a belief that an inventor should have exclusive rights to his design for ____.

- Lincoln was convinced he'd made a great contribution to the boating and shipping industries.

Lincoln was convinced he'd made ____ to the boating and shipping industries.

Lincoln was convinced he'd made a great contribution to ____.

- Lincoln made history as the only president to hold a patent for an invention.

Lincoln made history as ____ to hold a patent for an invention.

Lincoln made history as the only president to hold ____.

6. Do all major networks have to carry presidential addresses?

- While there's more than one way for a president to get television airtime, telling him to take a hike when the office has made a direct request to broadcasters is embarrassing for everyone involved.

While there's ____ for a president to get television airtime, telling him to take a hike when the office has made a direct request to broadcasters is embarrassing for everyone involved.

While there's more than one way for a president to get television airtime, telling him to take a hike when the office has made a ____ to broadcasters is embarrassing for everyone involved.

While there's more than one way for a president to get television airtime, telling him to take a hike when the office has made a direct request to broadcasters is embarrassing for ____.

- In cases where the president would like to speak to the nation directly, the White House will specifically ask networks to set aside time for the address.

In cases where the president would like to ____, the White House will specifically ask networks to set aside time for the address.

In cases where the president would like to speak to the nation directly, the White House will specifically ask networks to ____.

- When broadcasters fail to uphold their end of the bargain, the FCC can subject them to fines or revoke their licenses.

When broadcasters fail to uphold ____, the FCC can subject them to fines or revoke their licenses.

When broadcasters fail to uphold their end of the bargain, the FCC can ____ or revoke their licenses.

When broadcasters fail to uphold their end of the bargain, the FCC can subject them to fines or ____.

- While a presidential address is always by definition newsworthy, it almost always airs during primetime when networks have their biggest audiences.

While a presidential address is always by definition newsworthy, it almost always airs

during ____ when networks have their biggest audiences.

While a presidential address is always by definition newsworthy, it almost always airs during primetime when networks have ____.

- Agreeing to the president's request could means a loss of advertising revenue during the most profitable time of day, and sometimes networks choose to protect their bottom line instead.

Agreeing to the president's request could means ____ during the most profitable time of day, and sometimes networks choose to protect their bottom line instead.

Agreeing to the president's request could means a loss of advertising revenue during ____ , and sometimes networks choose to protect their bottom line instead.

Agreeing to the president's request could means a loss of advertising revenue during the most profitable time of day, and sometimes networks choose to protect ____ instead.

References

[1] Luhn, H. P., "The Automatic Creation of Literature Abstracts," IBM Journal of Research and Development , vol.2, no.2, pp.159,165, Apr. 1958.

[2] Erkan, Gnes, and Dragomir R. Radev. "LexRank: Graph-based lexical centrality as salience in text summarization." Journal of Artificial Intelligence Research (2004): 457-479.

[3] Mihalcea, Rada, and Paul Tarau. "TextRank: Bringing order into texts." Association for Computational Linguistics, 2004.

[4] Kotov, Alexander, and ChengXiang Zhai. "Towards natural question guided search." Proceedings of the 19th international conference on World wide web. ACM, 2010.

[5] Kim, Han-joon, and Min-kyoung Kim. "Design of question answering system with automated question generation." Networked Computing and Advanced Information Management, 2008. NCM'08. Fourth International Conference on. Vol. 2. IEEE, 2008.

[6] Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. OReilly Media Inc.

[7] Manning, Christopher D., Surdeanu, Mihai, Bauer, John, Finkel, Jenny, Bethard, Steven J., and McClosky, David. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.

[8] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

[9] PyTeaser, (2015), Github Repository, https://github.com/xiaoxu193/PyTeaser. Retrieved on July 15, 2015.

[10] Wu, Zhibiao, and Martha Palmer. "Verbs semantics and lexical selection." Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1994.

[11] Papasalouros, Andreas, Konstantinos Kanaris, and Konstantinos Kotis. "Automatic Generation Of Multiple Choice Questions From Domain Ontologies." e-Learning. 2008.

[12] Hasan, Yllias Chali Sadid A. "Towards Automatic Topical Question Generation."

[13] Hasan, Kazi Saidul, and Vincent Ng. "Automatic keyphrase extraction: A survey of the state of the art." Proceedings of the Association for Computational Linguistics (ACL), Baltimore, Maryland: Association for Computational Linguistics (2014).