The Thirty-Second International Florida
Artificial Intelligence Research Society Conference (FLAIRS-32)

Towards Predicting Difficulty of Reading Comprehension Questions

Takshak Desai, Dan I. Moldovan
Department of Computer Science
The University of Texas at Dallas
Richardson TX
{takshak, moldovan} @ hlt.utdallas.edu

Abstract

We present a corpus and approach to deduce the difficulty of questions asked in a reading comprehension test. A feature-driven model is designed that associates each question with a difficulty level. This eliminates the laborious task of manually annotating questions in a computerized testing environment. Experiments performed on our corpus show that our model can classify questions with a micro F-score of 0.68.

Introduction

Reading comprehension is widely used in classroom and testing environments to gauge student understanding; it requires a reader to identify high-level semantic relations that hold between text components, and have a deep understanding of the content (Brooks, Arnold, and Iacobbo 1977). In order to generate inferential questions for reading comprehension, we had designed a rule-based system (Desai, Dakle, and Moldovan 2018) that applies a set of syntactic transformations on relation triples to obtain question-answer pairs. These relations, which include Cause, Solutionhood, Background, etc. are described by the Rhetorical Structure Theory or RST (Mann and Thompson 1988). They illustrate how text spans are functionally related to each other.

Questions generated are of varying lengths and scopes, with some derived from implicit coherence relations; and some requiring the reader to give a detailed reply. Unlike factoid questions which expect a simple scan through the text to look for the correct response, these questions are more meaningful and require a deeper understanding of the text.

As a representative example, consider a of text and the questions that follow:

*Kidder, Peabody & Co. is trying to struggle back. [Only a few months ago, the 124-year-old securities firm seemed to be on the verge of a meltdown, racked by internal squabbles and defections.]$_1$ [Its relationship with parent General Electric Co. had been frayed **since** a big Kidder insider-trading scandal]$_3$ ..... [More than 20 new managing directors and senior vice presidents have been hired since January. The firm's brokerage force has been trimmed and its mergers-and-acquisitions staff increased to a record 55 people ...]$_3$*

Question 1: Why was Kidder on the verge of a meltdown a few months ago?
Question 2: How has Kidder tried to fight back following the issues it was facing months ago?
Question 3: What frayed the relationship between General Electric Co. and Kidder?

Here, Question 1 is an example of an intra-sentential question while Question 3 is an example of an inter-sentential question. Question 2 is derived from an explicit relation made apparent by the use of keyword 'since'.

To understand how relatively complex the questions are with respect to each other, we came up with a labelled dataset; and a novel feature-driven approach that automatically classifies these questions into their difficulty levels. We consider a rich set of syntactic and semantic features that takes into account the question-answer pairs and contextual information from the passage to perform classification. Our model gave an F-score of 0.68 against the corpus.

Related Work

Measuring Question Difficulty

The task of measuring question difficulty is inherently subjective: perception of difficulty is influenced by factors (Torgesen 2004) such as age, reasoning and inferential skills, extent of conceptual knowledge, ability to perform accurate and fluent reading, etc. Further, native speakers find questions easier compared to those for whom the language of the text is a second language (Van Gelderen et al. 2004). It was also shown (Nunan and Keobke 1995) that student perception of question difficulty differs from reality: A student may find a question more demanding if he/she is intimidated by the task and does not put in appropriate effort to attempt it. Likewise, the question may seem easy if he/she incorrectly assumes an aspect to be the task's key aspect.

Despite the psychological barriers to question difficulty analysis, several techniques have been proposed over the years to measure question complexity. One approach suggested classifying questions via the scope of the document

from which they are generated (Mannem, Prasad, and Joshi 2010): a general level question focuses on almost the entire paragraph, a medium level question concentrates on multiple clauses or sentences, and specific level questions are derived from single sentences.

Likewise, psychometrics (the discipline of study in psychology and education concerned with testing, measurement, assessment and related activities) suggests that the difficulty of a multiple-choice question can be statistically gauged by the proportion of test takers who answered it correctly: the value is estimated empirically by conducting a study before the actual test (Holland and Thayer 1985).

Models for Measuring Question Difficulty

Several statistical models have been built over the years to predict question difficulty. Linear SVMs were used (Yahya and Osama 2011) to automatically classify questions into six classes. It used an expert-curated dataset and a set of lexical features to achieve an accuracy of 0.85. Another study proposed using associative cellular neural network (Namba 2012) to classify questions from a Java programming course into three levels: easy, standard and difficult. Similar features were used (Hutzler et al. 2014) to design an automated ranking system for an Intelligent Tutoring System. A regressor that makes use of textual features was built (Sheehan, Flor, and Napolitano 2013) that measures the difficulty of listening comprehension items. The use of word-embeddings with a CNN-based neural network was also suggested (Hsu et al. 2018) to estimate the difficulty of multiple-choice questions.

Previous research work done in question difficulty prediction has focused merely on extracting naive textual features from the question: our work is significantly different from these approaches as we take into account both the question, the answer and its context in the paragraph for measuring how complex the question is: this allows us to achieve better accuracy in the classification task.

Problem Definition and Dataset

Formally, we define the problem of question difficulty classification as follows:

Given a set of question-answer pairs $(q_i, A_i) \in Q$ generated for a document d, where the answer A_i is obtained from d; and each question is associated with a difficulty level $c_j \in C = \{1, 2, 3\}$, build a model that approximates the function $c_j = f(q_i, A_i)$. The smaller the value of the difficulty level associated with a question, the easier it is.

Coherence Relations

Rhetorical Structure Theory or RST was introduced (Mann and Thompson 1988) to describe how textual elements are related to each other via discourse relations. Discourse parsing typically involves first segmenting the content into pieces of text called Elementary Discourse Units (EDUs): EDUs may be complete sentences, clauses or just words. These are then arranged as the nodes of a discourse graph: relations between nodes are represented by labelled arcs. In the context of a rhetorical relation, text spans are of two types: Nucleus and Satellite. Nucleus typically represents the text whose understanding is being facilitated by the Satellite.

For example, in the sentence *[Only a few months ago, the 124-year-old securities firm seemed to be on the verge of a meltdown]*$_1$ *[,racked by internal squabbles and defections.]*$_2$, EDU 1 is the Nucleus and EDU 2 is the Satellite that are related via the Cause relation. Some may be multi-nuclear in nature, for example, Contrast where both text spans are opposing views or facts.

In our case, every nucleus-satellite pair represents a candidate question-answer pair, where one text span represents a question, and the other represents the answer. Since these relations hold between arbitrary text spans, it allows us to generate both inter- and intra-sentential questions that test the understanding of different types of relations.

Generating Questions

We used the RST-DT corpus (Carlson, Marcu, and Okurowski 2003) that contains 385 Wall Street Journal articles; each article is accompanied by a discourse graph that describes the coherence relations holding between text spans. We make use of hand-designed templates (Desai, Dakle, and Moldovan 2018) to craft questions. The system parses through the document to identify coherence relation pairs and applies a set of syntax transformations to convert them into questions. Templates are defined for different types of relations such as Evidence, Evaluation, Condition, Circumstance, etc.

Due to the arbitrary nature of discourse, some of the generated questions were erroneous. Common reasons for this included grammatical and/or semantic incorrectness, redundancy in question due to superfluity of language, ambiguity in question meaning, etc. Likewise, some questions made no sense, and some were semantically identical to others. An evaluation of the questions revealed that 30% of the generated questions had extraneous use of language and 15% of the questions were ambiguous. 9% of the questions were semantically incorrect: an investigation of the sources of errors revealed the major reasons to be parsing errors, inability to handle direct speech, subordinate clause - main clause rearrangement, etc.

To reduce the number of instances of questions having poor quality, we modified some of the generated questions to make them sound more fluent and natural. In general, it was observed that inter-sentential questions had lots of superfluity as templates designed did not account for this superfluity of language. Our assessment showed that 55% of the generated questions had to be modified while 19% of the questions were discarded because they made no sense or were semantically similar to other generated questions.

We sampled 125 documents from the dataset and generated questions using coherence relations for each document. Out of the 1109 questions generated, we considered 894 questions (with/without modification) in our dataset.

Annotating Questions

The task of question annotation was performed by a team of 2 annotators who carefully perused each question and gave

it a difficulty rating. To increase the κ value, we ensured that all annotators belonged to the same age group and spoke English as a second language (Van Gelderen et al. 2004).

Thus, each question $q_i \in Q$ associated with document d is associated with a difficulty level c_i on a scale of $1-3$. Our dataset finally contains the documents, the coherence relation tuples, the question derived from the relation tuples and a difficulty rating. Table 1 provides some statistics on the corpus and the inter-annotator agreement. To measure the inter-annotator reliability, we use Cohen's kappa and Pearson coefficient measures. We achieved a reasonable agreement of $\kappa = 0.91$ and $\rho = 0.89$.

Dataset and Question Sizes	
No. of documents	125
No. of Questions selected	894
Average No. of words per Question	12.65
Class Distribution	
Ratio of questions with class 1	0.44
Ratio of questions with class 2	0.37
Ratio of questions with class 3	0.19
Inter-Annotator agreement	
Cohen's kappa κ	0.91
Pearson coefficient ρ	0.89

Table 1: Corpus and inter-annotator agreement statistics

Example

As an example, consider an article from our corpus with the associated questions shown in Figure 1. Each question is associated with a difficulty rating as shown. Briefly, the passage begins by stating that Mobil Corp. cut down the size of its work force. Then it provides additional details and background information about staff reductions. Later, it outlines the reasons behind job cuts; and possible outcomes of said cuts. Coherence relations allow us to model this logical flow of ideas: questions generated from these relations would test a broad, high-level understanding of the discourse. From the sample questions shown, one can infer that they are capable of assessing a variety of concepts such as cause-effect relations, cohesion of and correlation between ideas, association between two paragraphs, etc.

Classification

Models for classification and Implementation

We train three classifiers: 1. Logistic Regression with L2 regularization, 2. Linear SVM and 3. Random Forest Classifier with 100 decision tree estimators. We have used three different classifiers to test if the effect of features on classification accuracy is independent of the classifier used.

We made use of the scikit-learn library [1] to implement the classifiers. Classification is performed using 10-fold cross validation: micro-averages of the performance metrics are reported for each classifier and feature vector.

[1] http://scikit-learn.org/

Enhancing the baseline

As a baseline, we consider the bag-of-word (Yahya and Osama 2011) and tf-idf representation of questions as baselines. An analysis of the feature space showed that the classifier ended up depending on irrelevant words such as 'farmer', 'arena' and 'bullock' which were probably hurting the accuracy. To identify important words that could improve classification accuracy, we tokenized the dataset to identify all unique unigrams and computed their frequencies of occurrence. To reduce the dependency of classifier on irrelevant words, we compiled a list of frequently occurring tokens (that are not stop words) to include in the bag-of-words and tf-idf vectors. The following tokens were identified as relevant features: *why, what, how, circumstance, condition, evidence, when, solution, cause, result*. As one can see, these words seem to be fairly important as they reveal the intent of the question, for example, a question containing the word 'cause' is most likely to test a student on the cause-effect relationship.

In Table 2, we provide the results for the baseline feature representations. We observe an improvement in F-score for some representation-model combinations with feature reduction. We observed small improvements in performance for categorizing questions into classes 1 and 2; however accuracy of performance for class 3 remained virtually the same. A qualitative analysis of the obtained results revealed that a large fraction of questions that contain the identified keywords belonged to classes 1 and 2. Therefore, an improvement in accuracy of classification for these classes was expected. However, for class 3, we need special features such as length of the answer, nature of coherence relation, similarity between question and the sentence(s) from which the question is derived, etc. The next sub-section describes these features and their effect on performance.

Features

To improve upon existing systems that measure the difficulty of a question, we consider the following features that are appended to each of the four representations described in the previous Section:

1. **Question Length**: This is an integer given by the number of words in the question. A longer question is expected to be easier to understand than a shorter one (Cannell, Miller, and Oksenberg 1981).
2. **Count of complex syntactic structures**: The ability to apply appropriate parsing and inference rules to comprehend a question's meaning may depend on the sentence's syntactic structure. Here, we count the number of clauses and prepositional phrases in the question (Cannell, Miller, and Oksenberg 1981). We used the Stanford Parser (Socher et al. 2013) for counting.
3. **Discourse connectives in the answer**: This is a binary value that indicates whether the coherence relation used to derive the question is implicit or explicit. Discourse connectives in the answer such as 'but', 'since', 'as a result', etc. signal explicit coherence (Taboada 2006). Questions derived from explicit relations are expected to be easier to answer as opposed to those generated from implicit ones.

Passage:

Mobil Corp. is preparing to slash the size of its work force in the U.S., possibly as soon as next month, say individuals familiar with the company's strategy. The size of the cuts isn't known, but they'll be centered in the exploration and production division, which is responsible for locating oil reserves, drilling wells and pumping crude oil and natural gas. Employees haven't yet been notified. Sources said that meetings to discuss the staff reductions have been scheduled for Friday at Mobil offices in New Orleans and Denver.
Mobil's latest move could signal the beginning of further reductions by other oil companies in their domestic oil-producing operations. [In yesterday's third-quarter earnings report, the company alluded to a $40 million provision for restructuring costs involving U.S. exploration and production operations. The report says that the restructuring will take place over a two-year period and will principally involve the transfer and termination of employees in our U.S. operations. A company spokesman, reached at his home last night, would only say that there will be a public announcement of the reduction program by the end of the week.]$_3$ [Most oil companies, including Mobil, have been reporting lower third-quarter earnings, largely as a result of lower earnings from chemicals as well as refining and marketing businesses.]$_4$ Individuals familiar with Mobil's strategy say that [Mobil is reducing its U.S. work force because of declining U.S. output.]$_1$ [Yesterday, Mobil said domestic exploration and production operations had a $16 million loss in the third quarter, while comparable foreign operations earned $234 million.]$_4$ [Industry wide, oil production in this country fell by 500,000 barrels a day to 7.7 million barrels in the first eight months of this year. Daily output is expected to decline by at least another 500,000 barrels next year.]$_2$ Some Mobil executives were dismayed that a reference to the cutbacks was included in the earnings report before workers were notified.

Sample Questions:

No.	Question	Class
1	Why is Mobil Corp. reducing its U.S. work force?	1
2	Evaluate the situation of oil production in the Unites States.	2
3	What structural changes is Mobil Corp. undergoing?	2
4	Why is Mobil Corp. reporting lower third-quarter earnings?	3

Figure 1: A representative example from the corpus: we show the passage, generated questions and their difficulty levels

Baselines	Logistic Regression			Linear SVM			Random Forest		
	P	R	F	P	R	F	P	R	F
bow_q	0.45	0.48	0.46	0.44	0.46	0.45	0.46	0.50	0.48
tfidf_q	0.38	0.48	0.42	0.44	0.49	0.46	0.48	0.51	0.49
bow_q*	0.53	**0.54**	0.53	0.53	**0.55**	0.53	0.53	**0.55**	**0.54**
tfidf_q*	**0.54**	**0.55**	**0.54**	**0.54**	**0.55**	0.54	**0.54**	**0.55**	**0.54**

Table 2: Performance metrics for all feature representations and classifiers. Here, bow_q: bag-of-words representation of questions, tfidf_q: tf-idf representation of questions, bow_q* and tfidf_q*: corresponding representations with most frequent tokens.

Features	Logistic Regression			Linear SVM			Random Forest		
	P	R	F	P	R	F	P	R	F
bow_q_f	0.62	0.63	0.62	0.62	0.63	0.63	0.65	0.64	0.64
tfidf_q_f	**0.67**	**0.68**	0.67	**0.67**	0.67	0.67	0.63	0.64	0.64
bow_q_f*	0.66	0.67	0.66	0.67	**0.68**	**0.68**	0.62	0.62	0.62
tfidf_q_f*	0.64	0.65	0.65	0.64	0.65	0.64	0.63	0.64	0.63

Table 3: Performance metrics for all feature representations: Model abbreviations are identical to those given in Table 2.

4. **Similarity ratio between Question and its Source**: Some of the generated questions were summarized versions of sentences from which they are derived. All co-referents were resolved by replacing them with the concepts they were referencing. Further, some required a complete restructuring of the sentence as they were ambiguous. Identifying such semantic relations (coreference resolution, paraphrase detection, etc.) is particularly useful in gauging the breadth of knowledge of a language (Araki et al. 2016). To quantify the semantic similarity between the question and question stem, we computed the cosine distance between their vector representations.

5. **Nature of question**: This is a binary feature where we check whether the question is derived from multiple sentences; or from a single sentence or clause, allowing us

to determine how well one can identify and differentiate inter- and intra-sentential discourse relations and draw logical connections between ideas that may be displaced far apart from each other in the text.

6. **Nature of answer**: This is a binary feature where we check whether the expected answer is a single sentence or clause; or comprises multiple sentences. Questions that require a long response are expected to be more difficult than those requiring short and direct answers. The reason is that many of such questions require readers to interpret evaluations and assessments of opinions, identify multiple causes or evidences of an event, or detail a solution to some problem or issue.

The experimental results for our dataset are shown in Table 3. When we append features to each of the baseline feature representations, the performance improves considerably. An important reason identified was the significant improvement in F-scores for Class 3. Our hand-designed features were capable of distinguishing tougher questions from the easier ones as now the system recognizes the nature of coherence relations, questions and answers; and the paraphrasing of sentences yielding differently worded questions. The best F-score was observed for the linear SVM classifier with bag-of-words model containing the most frequent tokens. We investigate the effectiveness of each feature by incorporating them one at a time into the model: Table 4 shows the results for the best feature-classifier combination.

Features	Precision	Recall	Type F1
Baseline	0.53	0.55	0.53
+QL	+0.02	+0.00	+0.01
+PC	+0.005	-0.01	+0.005
+DC	+0.05	+0.05	+0.05
+CR	+0.03	+0.02	+0.03
+NQ	+0.03	+0.02	+0.03
+NA	+0.02	+0.01	+0.02

Table 4: Improvements in F-scores for each class: each cell indicates how the F-score increased with the incorporation of features: Here QL: Question Length, PC: Count of complex syntactic structures, DC: Presence of discourse connectives, CR: Similarity ratio between the question and its source, NQ: Nature of question; and NA: Nature of answer

Qualitative Analysis

Table 5 highlights some of the question-answer pairs and how they fared against our models. Each of the mentioned example explain how the feature-driven classifiers are able to identify the properties of question-answer pairs such as nature of question, nature of answer, nature of relation, paraphrasing, etc. to correctly predict the class value of a question as opposed to baselines that misclassify them.

On Differentiating between Classes 2 and 3

It was observed that the F-score for class 3 was generally lower than the F-scores for classes 1 and 2. Some reasons for this poor performance are:

1. The class-distribution is skewed. As seen in Table 1, the dataset contains many examples from classes 1 and 2, however only 19% were from class 3. The classifiers did not have enough data to perform reasonably well.
2. Classifiers found it difficult to differentiate between classes 2 and 3. While both classes differed from class 1 in the sense that either most questions were inter-sentential or were derived from implicit relations; annotators revealed that differentiating between classes 2 and 3 was challenging as it required them to make several judgement calls such as differentiating between how deep the semantics of a relation is.

For example, in Table 5, example 4 shows an instance of misclassification: our features misclassified this as class 2. Annotators revealed that they had classified this question as class 3 because not only was the question inter-sentential and the relation implicit in nature, but it required a much deeper understanding of the text to arrive at the answer: our features are not able to capture this effectively.

Conclusions and Future Work

Our contributions in this paper can be briefly summarized as follows:

1. We present a novel reading comprehension corpus in which the questions are annotated with a difficulty level.
2. A feature-driven classifier is presented that classifies questions according to their difficulty level. We used a rich set of semantic features to perform this task.
3. As opposed to considering factoid questions, we considered high-level meaningful questions that test a student's understanding of discourse coherence and semantics.

There are several avenues for further research. We have considered reading comprehension questions for generic documents: one can consider other sources of data also such as science textbooks, political discourse, medical literature, etc. Likewise, there are techniques that use approaches other than coherence relations to generate questions from discourse (Du, Shao, and Cardie 2017): such questions can also be analyzed for complexity. It would also be interesting to test the efficacy of word embeddings and neural networks in performing question difficulty analysis.

References

Araki, J.; Rajagopal, D.; Sankaranarayanan, S.; Holm, S.; Yamakawa, Y.; and Mitamura, T. 2016. Generating questions and multiple-choice answers using semantic analysis of texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1125–1136.

Brooks, P. H.; Arnold, D. J.; and Iacobbo, M. 1977. Some cognitive aspects of reading comprehension. *Peabody Journal of Education* 54(3):146–153.

Cannell, C. F.; Miller, P. V.; and Oksenberg, L. 1981. Research on interviewing techniques. *Sociological methodology* 12:389–437.

#	Class	Question	Expected Answer	Comments
1	2	How has Kidder tried to struggle back following the issues it was facing months ago?	More than 20 new managing directors and senior vice presidents have been hired since January. The firm's brokerage force has been trimmed and its mergers-and-acquisitions staff increased to a record 55 people	All baselines misclassified this as 1 probably because they did not realize the answer is a detailed one. However, with features, all classifier correctly classified it.
2	2	Under what conditions could Kidder's hiring binge backfire?	Kidder's hiring binge involving executive-level staffers, some with multiple-year contract guarantees, could backfire unless there are results.	The baselines classified this as 3. However, our features correctly identified the presence of 'unless' in the answer stem and predicted this to be an easier question.
3	2	What refueled speculation that Kidder is getting out of the brokerage business?	Mr. Carpenter this month sold off Kidder's eight brokerage offices in Florida and Puerto Rico to Merrill Lynch & Co.	The baselines predicted the class as 1. However, the question was a summarized version of a much larger nucleus in the relation. Our features used the cosine similarity to correctly predict the class as 2.
4	3	How has GE capital started to exploit the synergy between itself and Kidder Peabody?	The Kidder units have worked on 37 investment banking deals this year	Both the baselines and feature-driven classifiers erroneously classified this as class 2.

Table 5: Qualitative Analysis of Results.

Carlson, L.; Marcu, D.; and Okurowski, M. E. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*. Springer. 85–112.

Desai, T.; Dakle, P.; and Moldovan, D. 2018. Generating questions for reading comprehension using coherence relations. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, 1–10.

Du, X.; Shao, J.; and Cardie, C. 2017. Learning to ask: Neural question generation for reading comprehension. *arXiv preprint arXiv:1705.00106*.

Holland, P. W., and Thayer, D. T. 1985. An alternate definition of the ets delta scale of item difficulty. program statistics research.

Hsu, F.-Y.; Lee, H.-M.; Chang, T.-H.; and Sung, Y.-T. 2018. Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. *Information Processing & Management* 54(6):969–984.

Hutzler, D.; David, E.; Avigal, M.; and Azoulay, R. 2014. Learning methods for rating the difficulty of reading comprehension questions. In *Software Science, Technology and Engineering (SWSTE), 2014 IEEE International Conference on*, 54–62. IEEE.

Mann, W. C., and Thompson, S. A. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse* 8(3):243–281.

Mannem, P.; Prasad, R.; and Joshi, A. 2010. Question generation from paragraphs at upenn: Qgstec system description. In *Proceedings of QG2010: The Third Workshop on Question Generation*, 84–91.

Namba, M. 2012. Intelligent tutoring system with associative cellular neural network. In *E-Learning-Organizational Infrastructure and Tools for Specific Areas*. InTech.

Nunan, D., and Keobke, K. 1995. Task difficulty from the learner's perspective: Perceptions and reality. *Hong Kong papers in linguistics and language teaching* 18:1–12.

Sheehan, K. M.; Flor, M.; and Napolitano, D. 2013. A two-stage approach for generating unbiased estimates of text complexity. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, 49–58.

Socher, R.; Bauer, J.; Manning, C. D.; et al. 2013. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 455–465.

Taboada, M. 2006. Discourse markers as signals (or not) of rhetorical relations. *Journal of pragmatics* 38(4):567–592.

Torgesen, J. 2004. Adolescent literacy, reading comprehension & the fcat. In *CLAS Conference, Naples, FL. Retrieved*, volume 3, 07.

Van Gelderen, A.; Schoonen, R.; De Glopper, K.; Hulstijn, J.; Simis, A.; Snellings, P.; and Stevenson, M. 2004. Linguistic knowledge, processing speed, and metacognitive knowledge in first-and second-language reading comprehension: A componential analysis. *Journal of educational psychology* 96(1):19.

Yahya, A. A., and Osama, A. 2011. Automatic classification of questions into bloom's cognitive levels using support vector machines.