Indian J.Sci.Res. 14 (1): 83-90, 2017
ISSN: 0976-2876 (Print)
ISSN: 2250-0138 (Online)

READING COMPREHENSION SYSTEM – A REVIEW

K.M. ARIVUCHELVAN[a1] AND K. LAKAHMI[b]

[a]Research Scholar, Periyar Maniammai University, Thanjavur, India
[b]Professor, Periyar Maniammai University, Thanjavur, India

ABSTRACT

Reading Comprehension (RC) Systems are to understand a given text and return answers in response to questions about the text. Reading Comprehension can be viewed as single document question answering system. Machine reading comprehension becomes more vital to get the required information in no time. Many researchers are working in the area of machine reading comprehension since 1990s. Maturity of Natural Language Processing and AI has lead to the re-search in machine reading comprehension. Main aim of this paper is to review the various approaches adopted in Reading Comprehension system and to discuss the issues which are to be addressed.

KEYWORDS: Machine Learning, Pattern Matching, Reading Comprehension, Textual Entailment

N the modern era, information growth is exponential. Reading all the text that are generated is a time consuming process. The main aim of reading is to understand the text, but the level of understanding always differs from one reader to another reader. So an automated system must be devised to understand the text given. Automatic understanding of text helps in text summarization, question answering system and many more NLP applications.

Anselmo Penas et al. (2011) defined Machine Reading (MR) as a task that deals with the automatic understanding of texts. Evaluation of this "automatic understanding" can be approached in two ways: the first one is translating the text into formal language representation and evaluate those using structured queries. This approach is used for Information extraction. The second one understands the given text and evaluates it through natural language questions.

Machine Reading Comprehension System is a system that understands knowledge about the content of text given and generates answers for the questions queried. In this paper Reading Comprehension System and Machine Reading Comprehension are used interchangeably to refer Machine Reading Comprehension System. Reading Comprehension System provides computational solution for the query raised either by the user or auto generated by machine for any given comprehension/text. The research towards machine reading comprehension has started during the late 1990's and still it becomes an open ended research for researchers to achieve better results.

Earlier systems introduced for Reading Compression attempted simple approach based on pattern matching (bag-of-words) or through some handcrafted rules. But the level of understanding was not good enough to answer all the questions raised. To provide better answer good understanding is needed, so researcher developed many methods to improve the understanding level of the system. In this paper we have discussed various methods that produce surface level understanding to deeper level understanding with their evaluation results.

The paper is organized as follows: Section 2 gives an overview of Reading Comprehension System. Section 3 describes different methodologies involved in RCS. Finally, conclusions are given in Section 4.

READING COMPREHENSION SYSTEM (RCS)

Reading comprehension is the ability to read text, process it, and understand its meaning. Reading comprehension is a dynamic and an interactive process. To understand a text, the reader needs to recognize each word and retrieve its meaning, combine this information with syntactic knowledge to make meaningful sentences and integrate the meanings of each sentence to construct representation of the state of affairs described by the text. However the level of understanding differs from reader to reader. To evaluate their understanding levels, reading comprehension tests are proposed. Such tests ask reader to read a story and to demonstrate his/her understanding of that story by answering questions about it.

Reading Comprehension System is required since millions and millions of documents are generated every day. It is tedious for human to read and understand each and every document manually. Reading Comprehension system alleviates this problem. Fig. 1 depicts the general block diagram of reading comprehension system.

[1]Corresponding author

The block diagram shows four main blocks. The blocks comprehension text and set of questions are given as input by the user. The block final answer will receive answers for the questions queried from the central block. The central block is the heart of the Reading Comprehension System where NLP/AI techniques are applied to the text for understanding. Researchers propose different methodology to analyze the text and produce the relevant answer for the question given. Many methods are used to evaluate the performance of Reading Comprehension System. Simple method uses bag of words method for representing the text. Questions given are compared with the bag of words and the relevant answers are extracted from the text. Answers extracted are compared with the correct answer and the system is evaluated.

Fig. 2 shows an example story and set of questions to be answered.

Library of Congress Has Books for Everyone

(WASHINGTON, D.C., 1964) - It was 150 years ago this year that our nation's biggest library burned to the ground. Copies of all the written books of the time were kept in the Library of Congress. But they were destroyed by fire in 1814 during a war with the British.

That fire didn't stop book lovers. The next year, they began to rebuild the library. By giving it 6,457 of his books, Thomas Jefferson helped get it started.

The first libraries in the United States could be used by members only. But the Library of Congress was built for all the people. From the start, it was our national library.

Today, the Library of Congress is one of the largest libraries in the world. People can find a copy of just about every book and magazine printed.

Libraries have been with us since people first learned to write. One of the oldest to be found dates back to about 800 years B.C. The books were written on tablets made from clay. The people who took care of the books were called "men of the written tablets."

1. Who gave books to the new library?

2. What is the name of our national library?

3. When did this library burn down?

4. Where can this library be found?

5. Why were some early people called "men of the written tablets"?

Figure 2: Sample Remedia Reading Comprehension Story and Questions

This paper covers most of the important methods used for text understanding in Reading Comprehension System.

Evaluation Method

This section briefs various evaluation methods used in Reading Comprehension System. Reading Comprehension tests are considered to be one of the best evaluation methods for machine reading. When the machine reading system understands the text/story given, the system is evaluated based on answers it return for the question given. The system is tested with the available answer key. Returned answers are compared with the answer key to validate its correctness. Based on the number of questions queried and the number of correct answers the accuracy is evaluated as in (1).

$$accuracy = \frac{n_R}{n} \tag{1}$$

where

n_R: number of questions correctly answered.

n: total number of questions.

Anselmo Penas et al. (2010) came up with a new idea for evaluation that the system need not answer the question if it does not find a correct answer. System can leave a question unanswered in case it was not sure about the correct answer to that question. The objective was to reduce the incorrect answers while keeping the correct ones, by leaving some questions unanswered. The evaluation measure proposed was c@1.

Anselmo Peñas and Alvaro Rodrigo (2011) used the new accuracy measure (c@1) and demonstrated how this measure was able to reward systems that maintain the same number of correct answers and at the same time decrease the number of incorrect ones, by leaving some questions unanswered. This measure is well suited for tasks such as Reading Comprehension tests, where multiple choices per question are given, but only one is correct. The formulation of c@1 is given in (2)

$$c@1 = \frac{1}{n}\left(n_R + n_U \frac{n_R}{n}\right) \tag{2}$$

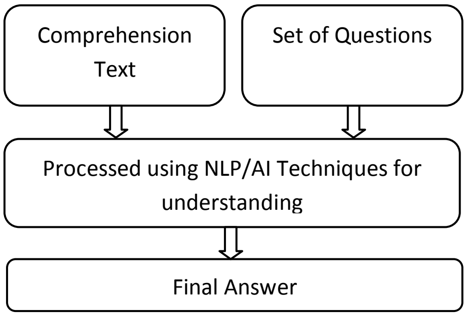

Figure 1: Simple Block Diagram of Reading Comprehension System.

where,

n_R: number of questions correctly answered.

n_U: number of questions unanswered.

n: total number of questions

c@1 acknowledges returning NoA answers in the proportion that a system answers questions correctly, which is measured using the traditional accuracy. (the proportion of questions correctly answered). Thus, a higher accuracy over answered questions would give more value to unanswered questions, and therefore, a higher final c@1 value. This measure will encourage the development of systems able to check the correctness of their responses because NoA answers add value to the final value, while incorrect answers do not.

There is another secondary measure called accuracy used in traditional QA (3).

$$accuracy = \frac{n_R + n_{UR}}{n} \quad (1)$$

where,

n_R: number of questions correctly answered.

n_{UR}: number of unanswered questions whose candidate answer was correct.

n: total number of questions.

The following section will give a detailed discussion on different methodologies and their evaluation results in Reading Comprehension System.

LITERATURE REVIEW

In this literature survey, we had explained various methodologies developed for RCS, the dataset used and their achieved results are shown.

L Hirschman et al. (1999) at MITRE Corporation had proposed an automated reading comprehension system called DeepRead. This system accepts comprehension text as input and finds answers in the text for the question queried by the user. The technique used in DeepRead is pattern matching (bag-of-words) technique to retrieve the sentences (both question and text) are represented as the set of words, and then the information content is extracted from them. Different methods are applied for extraction and those are removal of function or stop words, stemming, name tagger, noun classification and finally noun resolution system. After the extraction of information content search task is performed to find the best match between the word set representing the question and the sets representing the sentences in the task. The corpus consisting of 60 stories of remedial reading materials for grade 3-6 is used as dataset for evaluation and the result achieved by this system is about 30% - 40%.

Ellen Riloff and Michael Thelen (2000) had developed a rule based question answering system called QUARC (QUestion Answering for Reading Comprehension). QUARC used set of heuristic for each 'Wh' question type (Who, What, When, Where, Why). QUARC takes comprehension text (story) and a question as input and process them to find the correct answer for the question from the given text. This system parses the question and all sentences of the story using a parser called Sundance. It uses morphological analysis, part of speech tagging, and semantic class tagging and entity recognition. Apart from 'Wh' rules a special set of rules is formed for dateline, which will be helpful for answering when and where questions. The rules are applied to each and every sentence of the story and each rule is awarded a point and those points were keys for finding the answer to the question. QUARC uses the same dataset used by DeepRead and achieves 40% accuracy.

Hwee Tou Ng et al. (2000) had developed a new QA system named AQUAREAS (Automated QUestion Answering upon Reading Stories) using machine learning approach. This system is independent of handcrafted rules followed in previous approaches. Here they represent each question-sentence pair as feature vector. This feature vector representation helps learning algorithm to build five classifiers automatically for each question type. The machine learning approach used here was comprised of two steps. First, a set of features was designed to capture the information that in turn helps to distinguish answer sentences from non-answer sentences. Second step is to generate a classifier for each question type from the training examples using learning algorithm. The learning algorithm used was C5. The tested the same dataset used by DeepRead. The approach achieved competitive results on answering questions for reading comprehension.

Tiphaine Dalmas et al. (2003) had developed and evaluated robust Question Answering (NLQA) methods. The corpus used by them is CBC4Kids and in that corpus they added a XML annotation layers for tokenization, lemmatization, stemming, semantic classes, POS tags and best ranking syntactic parsers to support experiments with semantic answer retrieval and inference. Due to the enhancements made into the corpus they proposed this would be a standard resource for inference based techniques to come in future. Also the corpus was tested with DeepRead and found the method performs slightly better than on the REMEDIA corpus.

Ben Wellner et al. (2006) presented an automated system called ABCs (Abduction Based Comprehension system). This system understands the role of various linguistic components in reading comprehension with respect to its questions. ABCs had an abductive inference engine with three main capabilities: (1) first-order logical representation, (2) graceful degradation and (3) system transparency. In first-order logical representation entities relations and events in the text and inference rules are represented. Inclusion of abduction in the reasoning engine helps knowledge representation and reasoning systems. Abductive inferences provides cue to where the system is performing poorly and to where the existing knowledge is inaccurate or new knowledge is provided. Few subcomponents are not automated and still the system achieves 35% to 45% accuracy and on few question types like who it achieves 50% accuracy.

Juan Martinez-Romo and Lourdes Araujo (2011) had developed a system which constructs a co-occurrence graph with words. In their system architecture they had focused on four main modules as Background preprocessing, Co-occurrence graph, Detecting communities and Question Answering. In background preprocessing, they used one reference corpus consisting of about 30,000 un-annotated documents related to the topic. GENIA tagger was used for PoS tagging and it is done only on nouns and verbs. In co-occurrence graph, aim is to create a link joining every two words sharing a common meaning. For doing this they had extracted nouns and verbs, further stemming was done using porter algorithm. In detecting communities, WalkTrap program is used to compute communities in large networks using random walk. In question answering, detected communities are treated as different context of a question in the corpus. Each question is assigned to a community based on their similarity. Similarly for answers, each response is assigned to a community and selected the answer based on highest similarity. The c@1 of 0.27 was obtained.

Suzan Verberne (2011) had proposed retrieval based question answering system for machine reading evaluation. In QA4MRE, they followed a relatively knowledge poor approach based on Information Retrieval techniques. It involves two steps: (1) retrieval of relevant fragments from the document for the input question (2) matching of the multiple choice answer candidates against the retrieved fragments in order to select the most likely answer. For retrieval of fragments they followed two information expansion methods: (1) Statistical expansion (2) question to fact expansion. They concluded that statistical expansion gives better results over question to fact expansion and there need further improvement to achieve better results. The overall c@1 of 0.37 was obtained in this method.

Detmar Meurers et al. (2011) had presented CoMiC-DE (Comparing Meaning in Context - DE), the first content assessment system for German. Content assessment supports the integration of context and task information into analysis. The comparison of student answers and target answer is based on an alignment of tokens, chunks, and dependency triples between the student and the target answer at different levels of abstraction. It is not sufficient to align only identical surface forms. In student answers there is chance for lexical and syntactic variation hence alignment would support different levels of abstraction. Here, the different question types and the ways in which the information asked for is encoded in the text. Then analyze the role of the question. The surface-based account of information given in the question should be replaced with the answer in the context of the question. The experiments are tested on the Corpus of Reading comprehension Exercises in German called CREG. CoMiC-DE performs on a competitive level of accuracy at 84.6%.

Adrian Iftene et al. (2012) participated in QA4MRE 2012 evaluation task and proposed their new method which is based on textual entailment. They constructed the Text and the Hypothesis for initial test Data. The test data is organized in the form of tags, <document> tag used to build the text whereas <question> and <answer> tags were used to build the hypothesis. Both Text (T) and Hypothesis (H) are given to Textual entailment system to get the partial and global scores per answer. The test data and background knowledge are related to four topics: AIDS, Climate Change, Music and Society and Alzheimer. The test is conducted for both Romanian and English. The c@1 was 0.28 obtained for English and 0.25 for Romanian.

Pinaki Bhaskar et al. (2012) had developed a QA system for QA4MRE @ CLEF 2012. In their system first they form the Hypothesis(H) by combining the question and each answer option. Using Lucene stop words are removed from each H and query words are identified to retrieve the most relevant sentences from the associated document. For retrieving relevant sentences they used TF-IDF. Each retrieved sentence defines the text T. Each T-H pair is assigned a ranking score based on textual entailment principle. Using ranking score, weight is automatically assigned to each answer options. Further each sentence is assigned an inference score with respect to each answer pattern, which is then multiplied with validate weight based on their ranking to find the highest selection score. The identified selection score is considered to be the answer to the given question. The results are evaluated for 3 datasets, 2 with domain knowledge and 1 without domain knowledgebase. The datasets with domain knowledgebase are producing satisfactory results and the result of dataset without domain knowledgebase is very poor. They proved that domain knowledgebase had a strong effect. The test data taken for evaluation is same one which is used in QA4MRE 2012 track.

Peter Clark et.al (2012) proposed an Entailment Based approach for the QA4MRE Challenge. This approach estimates the likelihood of textual entailment between sentences S in text and the question Q and each candidate answer A_i. The entire approach is divided into two important task: entailment assessment and implication assessment. In entailment assessment, the candidate answer A_i had to be found from sentence S, to do so first the sentence S is created by means of formal representation. It is difficult, but the author used natural logic approach to achieve it. Once the candidate answer A_i is found in S, the next step implication assessment is processed. In implication assessment, it is mandatory to validate the candidate answer A with Question Q. The author had investigated the syntactic connection between the Q-A pair. It may difficult in some cases due to the indirect connection of Q-A pair. To resolve this, the author found the closest pair by measuring the distance between the sentences. However the end result achieves only 40% accuracy. The author concluded that the accuracy can be improved when the knowledge problem and reasoning problem achieves good result.

Michael Hahn and Detmar Meurers (2012) had proposed a semantic based approach for reading comprehension questions. They presented CoSeC-DE system for evaluating the content of answers to reading comprehension. The dataset used for evaluation is CREG. Here they use Lexical Resource Semantics (LRS) representation for the student answer, the target answer and the question are automatically derived on the basis of the part of speech tags assigned by tree tagger and the dependency parser by MaltParser. After LRS representation then alignment takes place both with local criteria and global criteria. The aligning meaning representation supports the integration of important information structural differences in a way that is closely related to the information structure research in formal semantics and pragmatics. The result shows CoSeC-DE outperform the earlier system called CoMiC-DE on the same dataset.

Helena Gomez-Adorno et al. (2013) had presented a methodology for handling the question answering system for reading comprehension tests. The developed system accepts a document as input and it answers multiple choice questions about it. Pre processing works were done through Lucene information retrieval engine. The proposed system architecture is organized into four main modules: document processing, information extraction, answer validation, answer selection. To determine the performance of the system they used the corpora provided in the QA4MRE task at CLEF 2011 and 2012. The average overall best run obtained in 2011 is outperformed in 2012.

Somnath Banerjee et al. (2013) focused on Multiple Choice Question answering system for entrance examination. In their system, first they generated answer pattern by combining the question and each answer option to form the hypothesis (H). Next, they removed stop words and interrogative words from each H. Using Lucene the most relevant sentences are retrieved from the associated document with respect to the query word. Each retrieved sentences defines text T and each T-H pair is assigned a ranking score calculated based on textual entailment principle. After calculating the ranking score the matching score was assigned to answer options. Thereafter the inference score was found for each sentence with respect to each answer pattern. The inference score and matching score for each answer option is added. Finally the answer option that gets highest selection score is selected as answer for the given question. The test set chosen from the Japanese center test which is conducted for Japanese University admissions. This system achieves overall c@1 of 0.42 is achieved.

Xinjian Li et al. (2013) had proposed a QA system for entrance exams in QA4MRE at CLEF 2013. They used three components namely character resolver, Sentence extractor and Recognizing Textual Entailment. The character resolver is used to identify the characters who were involved in the story and are assigned with an ID. The sentence extractor would extract the related sentences for each question, the extracted sentences are then used to create a T|H pair. Finally this T|H pair is given as input for the RTE system which will produce answer. The test data for evaluation in the entrance exams task is from Japanese university entrance examination. This system obtained a c@1 of 0.35 during evaluation.

Simon Ostermann et al. (2014) presented a system in CLEF QA Track 2014 Entrance Exam. The system is designed to correctly answer multiple choice reading comprehension exercises. The system is originally designed for scoring short answers given by language learners to reading comprehension questions. This was implemented by two step procedure. In the first step, the sentence which best matched with question is selected. In the second step, the selected sentences are compared with four possible answer choices to find their similarity score. The choice with highest similarity score is returned as correct answer. Preprocessing are done through all standard NLP tools such as sentence splitting (OpenNLP), tokenization (Stanford CoreNLP), PoS Tagging and stemming (TreeTagger) synonym extraction (WordNet). Using alignment model the similarity score is calculated. The system achieved c@1 of 0.25 on the given dataset.

Helena et al. (2014) in their approach presented that the given document and multiple choice answers are transformed into graph based representation which contains lexical, morphological and syntactic features. After the construction of graph, it was traversed into different paths both in texts and in the answer choices to find the syntactic features of the graph. This results in construction of several feature vectors. Finally the cosine similarity is calculated for feature vector to rank the multiple choice answers. The feature that rank achieved highest rank results as correct answer. The dataset used for evaluation is Japanese Center Test. The system achieved a c@1 of 0.375 in the evaluation.

Neil Dhruva et al. (2014) presented a open domain Reading comprehension System to text understanding evaluation. It is based on text similarity measures, textual entailment and coreference resolution. The text is represented in XML document. It was then preprocessed using stanford NLP tools. For retrieving sentence this system uses three similarity measures ie, lexical similarity, ESA-based similarity and PoS similarity. Based on the results of similarity measures textual entailment is calculated for T-H pairs. After calculating textual entailment answer similarity is computed. For answer selection the entailment confidence score and the answer similarity scores are used to calculate the correctness score. Finally the answer option corresponding to the T-H pair with the highest correctness score was selected as correct answer for a given question. This system achieved c@1 score of 0.375 for the given track dataset.

So far we had discussed different methods on MCQ based reading comprehension system. But here Martin Gleize et al. (2014) proposed a method to invalidate the answer options to find correct answer. In this system first the relevant passage for the question was retrieved. Then it generates Predicate Alignment Structure (PAS) to each answer options. Likewise PAS was generated to the retrieved passage. To remove wrong answer a new rule is proposed by the author and its goal was to eliminate as many possible answer options without removing the correct answer option. Finally the answer option is returned based on their alignment score. Tokyo University entrance exam dataset was used for evaluation and this system achieved random baseline score c@1 of 0.25 after submitting several runs.

Dominique Laurent et al. (2015) had proposed an entrance exam evaluation task at CLEF 2015. In this task they had used a special structure to save the results called CDS (Clause Description Structure). The main components of the structure are descriptions of a clause, a subject, a verb and an object. Apart from this the structure also allows indirect object, temporal context, spatial context and so on. For the evaluation a dedicated module is added to compare the CDS from task questions and answers. This module measures the degree of correspondence between the elements. The evaluation result shows 52 good answers out of 89 questions in English and 50 good answers out of 89 in French.

Martin Gleize et al. (2015) had presented a methodology called LIMSI. It selected set of passages as graphs and made enhancement through external sources. The main aim is to reduce the gap between human and computer for extracting knowledge from the text. From there modifications were carried out to get candidate graph from passage graphs. Then they applied classifiers for validation and rejection. Finally, the final score is calculated as difference of validation score and rejection score. The system achieved c@1 of 0.36 during evaluation.

Ramon Ziai and Bjorn Rudzewitz (2015) had proposed a new method called CoMic. In this method the text segment identification was carried out for segmenting the text into meaningful paragraphs. Later it was compared with the question to be answered using similarity metric. It results in ordering the meaningful partition with the questions based on their similarity. Similarity features were then extracted for each candidate answer to each paragraph. For ranking the features the author had chosen a machine learning approach called SVM. This approach achieved c@1 of 0.29 for the given entrance exam task.

We had discussed various methodologies that dealt with text understanding evaluation using Reading Comprehension System.

CONCLUSION

Automatically answering reading comprehension questions is a challenging task and still it is an open ended research for the research community in NLP. In this paper we have seen methods that involve surface level understanding to deeper level understanding. In surface level understanding sentence will be extracted as answer, whereas in deeper level understanding the given text is analyzed through textual entailment technique, answer validation and text similarity measures. Here the evaluation is done using multiple choice question answers. The highest evaluation result achieved for c@1 is 0.42 to the best of our knowledge. This clearly shows still there is huge gap to be filled to obtain more accurate results.

REFERENCES

Penas A., Hovy E.H., Forner P., Rodrigo A., Sutcliffe R. F.E., Forascu C. and Sporleder C., 2001. “Overview of qa4mre at clef 2011: Question answering for machine reading evaluation”, CLEF (Notebook Papers/Labs/Workshop), pp. 1-20.

Anselmo Penas, Pamela Forner, Richard Sutcliffe, ˜Alvaro ´ Rodrigo, Corina Forascu, Inaki Alegria, Danilo Gi- ˜ampiccolo, Nicolas Moreau, and Petya Osenova, "Overview of ResPubliQA 2009: Question Answering Evaluation over European Legislation", In Multilingual Information Access Evaluation I. Text Retrieval Experiments, CLEF 2009, Revised Selected Papers, volume 6241 of Lecture Notes in Computer Science, Springer, 2010.

Anselmo Peñas and Alvaro Rodrigo, "A Simple Measure to Assess Non-response", In Proceedings of 49th Annual Meeting of the Association for Computational Linguistics - Human Language Technologies (ACL-HLT 2011). Portland. Oregon. USA. June 19-24, 2011.

Hirschman L., Light M., Breck E. and Burger J.D., 1999. "Deep read: a reading comprehension system", In: Proceedings of the 37th meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA.

Riloff E. and Thelon M., 2000. "A rule-based question answering system for reading comprehension tests", In: Proceedings of the 2000 ANLP/NAACL Workshop on Reading comprehension tests as evaluation for computer-based language understanding sytems, Stroudsburg, PA, USA.

Ng H.T., Teo L.H., Lai J. and Kwan J.L.P., 2000. "A machine learning approach to answering questions for reading comprehension tests", In: Proceedings of EMNLP/VLC.

Dalmas T., Leidner J.L., Webber B., Grover C. and Bos J., 2003. "Generating Annotated Corpora for Reading Comprehension and Question Answering Evaluation", In EACL-2003 workskhop on Natural Language Processing (NLP) for Question-Answering. Budapest, Hungary.

Wellner B., Ferro L., Greiff W. and Hirschman L., 2006. "Reading comprehension tests for computer-based understanding evaluation", Nat. Lang. Eng., **12**(4):305-334.

Juan Martinez-Romo and Lourdes Araujo, "Graph-based Word Clustering Applied to Question Answering and Reading Comprehension Tests", CLEF 2011 Labs and Workshop - Notebook Papers. 19-22 September, Amsterdam, The Netherlands. Online Proceedings, 2011.

Helena Gomez-Adorno, David Pinto, Darnes Vilarino "A Question Answering System for Reading Comprehension Tests", Springer link, Pattern Recognition, Volume 7914 of the series Lecture Notes in Computer Science pp 354-363, 2013.

S. Verberne, "Retrieval-based Question Answering for Machine Reading Evaluation", CLEF. In CLEF 2011 Labs and Workshop, Notebook Papers., Amsterdam, 2011.

Detmar Meurers, Ramon Ziai, Niels Ott, and Janina Kopp, "Evaluating Answers to Reading Comprehension Questions in Context: Results for German and the Role of Information Structure", In Proceedings of the TextInfer 2011 Workshop on Textual Entailment, EMNLP 2011.

Iftene A, Gˆınsca A-L, Moruz M.A, Trandabat D, Husarciuc M, Boros E, "Enhancing a Question Answering System with Textual Entailment for Machine Reading Evaluation", CLEF 2012 (Online Working Notes/Labs/Workshop), 2012.

Bhaskar P, Pakray P, Banerjee S, Banerjee S, Bandyopadhyay S, Gelbukh A, "Question answering system for QA4MRE@CLEF 2012", In: CLEF 2012 Workshop on Question Answering For Machine Reading Evaluation (QA4MRE).

Peter Clark, Phil Harrison and Xuchen Yao, "An Entailment-Based Approach to the QA4MRE Challenge", Proc. CLEF 2012 (Conference and Labs of the Evaluation Forum) - QA4MRE Lab 2012.

Hahn, M., Meurers, D, "Evaluating the meaning of answers to reading comprehension questions: A semantics-based approach", In: Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-7) at NAACL-HLT 2012. pp. 94-103. Montreal, 2012.

Gómez-Adorno, Helena, David Pinto, and Darnes Vilarino, "A Question Answering System for Reading Comprehension Tests", Mexican Conference on Pattern Recognition. Springer Berlin Heidelberg, 2013.

Banerjee, S., Bhaskar, P., Pakray, P., Bandyopadhyay, S., Gelbukh, A, Multiple Choice Question (MCQ) Answering System for Entrance Examination, Question Answering System for QA4MRE@CLEF 2013. CLEF 2013 Evaluation Labs and Workshop Online Working Notes, ISBN 978-88-904810-5-5, ISSN 2038-4963, Valencia -Spain, 23 -26, 2013.

Li, X., Ran, T., Nguyen, N.L.T., Miyao, Y., Aizawa, A, "Question Answering System for Entrance Exams in QA4MRE", CLEF 2013 Evaluation Labs and Workshop Online Working Notes, ISBN 978-88-904810-5-5, ISSN 2038-4963, Valencia -Spain, 23 -26, 2013.

Simon Ostermann, Nikolina Koleva, Alexis Palmer and Andrea Horbach, "CSGS: Adapting a short answer scoring system for multiple-choice reading comprehension exercises", CLEF 2014 Working Notes, Sheffield, 2014.

Helena Gómez-Adorno, Grigori Sidorov, David Pinto and Alexander Gelbukh, "Graph Based Approach for the Question Answering Task Based on Entrance Exams", CLEF 2014 Working Notes, Sheffield, 2014.

Neil Dhruva, Oliver Ferschke and Iryna Gurevych, "Solving Open-Domain Multiple Choice Questions with Textual Entailment and Text Similarity Measures", CLEF 2014 Working Notes, Sheffield, 2014.

Martin Gleize, Anne-Laure Ligozat and Brigitte Grau, "LIMSI-CNRS@CLEF 2014: Invalidating Answers for Multiple Choice Question Answering", CLEF 2014 Working Notes, Sheffield, 2014.

Dominique Laurent, Baptiste Chardon, Sophie Negre, Camille Pradel and Patrick Seguela, "Reading Comprehension at Entrance Exams 2015", CLEF 2015 Working Notes, Toulouse, 2015.

Ramon Ziai, "CoMiC: Exploring Text Segmentation and Similarity in the English Entrance Exams Task" CLEF 2015 Working Notes, Toulouse, 2015.

Martin Gleize and Brigitte Grau, "LIMSI-CNRS@CLEF 2015: Tree Edit Beam Search for Multiple Choice Question Answering", CLEF 2015 Working Notes, Toulouse, 2015.