Universidade FUMEC
Faculdade de Ciências Empresariais
Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento

Exploiting Neural Networks in Question Answering to Support the Teaching-Learning Process

Marco Antônio Calijorne Soares

Belo Horizonte
2018

Marco Antônio Calijorne Soares

Exploiting Neural Networks in Question Answering to Support the Teaching-Learning Process

Msc thesis presented to the Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento of FUMEC University, as requirements for the Master's degree.

Research Track: Technology and Information Systems

Supervisor: Prof. Dr. Fernando Silva Parreiras

Belo Horizonte
2018

Dados Internacionais de Catalogação na Publicação (CIP)

S676e Soares, Marco Antônio Calijorne, 1980-

Exploiting neural networks in question answering to support the teaching-learning process / Marco Antônio Calijorne Soares. – Belo Horizonte, 2018.

65 f. : il. ; 29,7 cm

Orientador: Fernando Silva Parreiras

Dissertação (Mestrado em Sistemas de Informação e Gestão do Conhecimento), Universidade FUMEC, Faculdade de Ciências Empresariais, Belo Horizonte, 2018.

1. Recuperação da informação. 2. Processamento de linguagem natural (Computação). 3. Estratégias de aprendizagem. I. Título. II. Parreiras, Fernando Silva. III. Universidade FUMEC, Faculdade de Ciências Empresariais.

CDU: 681.3:37

Ficha catalográfica elaborada pela Biblioteca Universitária-FUMEC

Abstract

Question Answering (QA) systems brought a fresh perspective to the Information Retrieval (IR) research area, enabling humans to ask natural language question to a computational system which can retrieve a single and precise answer. In front of this nature, we understand that this kind of approach could bring a valuable contribution to a Teaching-Learning process. This study aims to analyze how question answering algorithms performs when applied to an educational environment. To achieve this goal, we developed a Systematic Literature Review (SLR) which enlighten the concepts, definitions and patterns of QA field, guiding us on which algorithm, approach and paradigm were more suitable for our needs. After that we created our own educational corpus and tested two approaches with it. As a result, we concluded that, QA systems can be used as a important tool on a teaching-learning process as we could reach a 77% match on factoid answers.

Key-words: question answering systems; natural language processing; information retrieval.

List of Figures

Figura 1 – Architecture of the three macro question answering modules. Question processing module classifies the question by its type and morphology. Answer processing module uses the classification and transformation made in question processing module to extract the answer from the result of the Document Processing module that executes previously to create datasets, indexes or neural models. 15
Figura 2 – Publication Time-Line since 2000 and the average of the results obtained by the approaches implementation. The average oscillation over the years show a constantly seek to better option on QA research. . . . 22
Figura 3 – Relationship between the high accuracy reached and independence language implementations, showing that Natural Language Processing reached best performance. 23
Figura 4 – TOP 15 Techniques, algorithms, frameworks and tools observed in SLR. 24
Figura 5 – QA paradigms and their metrics. 25
Figura 6 – Fields where researcher are implementing their approaches. There is a wide set of fields, besides the amount of implementations in open domain on *www* and medicine on restricted domain. 26
Figura 7 – Natural language processing techniques used in other paradigms besides de NPL based one. 27
Figura 8 – Main architecture of a Question Answering system based on DMN+ approach. The Input Module with the sentence reader and input fusion layer, which differentiates this approach from the traditional DMN implementation. 32
Figura 9 – Process defined to normalize the books files. This task was executed to increase the information quality extracted from all books. 35
Figura 10 – Examples of how the training data was prepared for training the DMN+ approach. 38
Figura 11 – Results achieved by each approach when executed with our corpus. . . 40
Figura 12 – Differences between DMN+ and Seq2Seq when analyzing their performances. 40
Figura 13 – Average Elapse time reached by each approach classified by the answers type. 42
Figura 14 – Average Elapse time reached by each approach classified by the answers type and the question type. 42
Figura 15 – Languages by Paradigm implementation. 63

List of Tables

Tabela 1 – Research Questions. 18
Tabela 2 – Number of papers retrieved in each Digital Library after search strings execution. 19
Tabela 3 – Inclusion and Exclusion Criteria defined for screening. 19
Tabela 4 – Question Assessment Questions. 20
Tabela 5 – 124 Papers addressed in this study. They were categorized by the question answering paradigm implemented. 21
Tabela 6 – Six Categories created to classify the Studies. 22
Tabela 7 – Software Engineering books used to create the system corpus. 34
Tabela 8 – Exclusion criteria used to classify sentences extracted from the software engineering books. One achieved criteria is enough to not use the sentence on the corpus. 35
Tabela 9 – Inclusion and Exclusion criteria used to classify the question according their compatibility to a corpus set. 36
Tabela 10 – Amount of questions by its type used as testing set on our implemented approaches. 37
Tabela 11 – Time spent for each approach on model training. Both models were trained in the same conditions of hardware and basic software settings. 41

Summary

1 INTRODUCTION 11

2 SYSTEMATIC LITERATURE REVIEW 13
2.1 Introduction 13
2.2 Background 14
2.2.1 Question Answering 14
2.2.1.1 Evaluation Methods 16
2.3 Research 17
2.3.1 Data Retrieval 18
2.3.2 Screening of papers 18
2.3.3 Eligibility 19
2.3.4 Data Extraction and Classification 20
2.4 Results and Discussion 20
2.5 Discussion 20
2.5.1 *RQ1* - What is the representativeness of QA paradigm? 20
2.5.2 *RQ2* - Which are the most frequently applied QA techniques? 23
2.5.3 *RQ3* - Which metrics or indicators are used to compare different Question Answering algorithms, techniques or systems? 24
2.5.4 *RQ4* - Which are the fields in which Question Answering systems and Natural Language Processing are being applied? 25
2.5.5 *RQ5* - How the relationship between Question Answering systems and Natural Language Processing is built? 25
2.6 Related Work 26

3 EXPERIMENT 29
3.1 Introduction 29
3.2 Background 29
3.2.1 Question Answering 29
3.2.2 Improved Dynamic Memory Network 30
3.2.3 Sequence To Sequence 31
3.3 Experiments 32
3.3.1 Definition of QA Paradigm and Approaches 33
3.3.2 QA System Domain and Corpus 33
3.3.3 Training and Testing sets 34
3.3.3.1 Training Sets 34
3.3.3.2 Testing Set 36

3.3.4 Implemented Approaches . . . 36
3.3.4.1 Improved Dynamic Memory Network . . . 38
3.3.4.2 Sequence To Sequence . . . 39
3.4 Results . . . 39
3.5 Discussion . . . 41
3.6 Conclusion and Future Work . . . 43

4 CONCLUSION . . . 45

References . . . 47

APPENDIX 61

1 Introduction

Question Answering (QA), a growing field from Information Retrieval (IR), has as its main goal provide precise answers to questions posed in natural language Athenikos and Han (2010). In other words, by employing different techniques, such as Natural Language Processing and Information Extraction, these systems aims to retrieve a single and precise answer for a question, both in natural language, asked by a human, instead of returning long lists of possible documents that could have the answer for the inquirer doubt, as the current search engines usually works.

The research on QA systems had its beginning far back in 1960s, when systems like BASEBALL Green Jr et al. (1961) and LUNAR Woods (1997) were developed but, the scientific community, involved with information retrieval, starts to join this studies only after conferences like TREC[1] begun with tasks in this field. Since then, algorithms, techniques and paradigms have been developed seeking to improve QA systems accuracy and performance.

Question answering systems can be applied in several domain of knowledge, such as, medicine Yu et al. (2007); Terol et al. (2007); Dietze and Schroeder (2009); Athenikos et al. (2009); Cao et al. (2010); Richardson et al. (2011); Cao et al. (2011); Aarabi (2013); Ferrucci et al. (2013); Abacha and Zweigenbaum (2015); Hristovski et al. (2015), turism Ferrández et al. (2009a); Ferrandez et al. (2011); Bhoir and Potey (2014); Hartawan et al. (2015), agriculture Malik et al. (2013); Biswas et al. (2014), information technology Pudaruth et al. (2016) and many others, besides the different languages that they can attend. Watchful to this context, we observed a high adherence of these systems to educational environment, specially for a Portuguese teaching-learning process. With this in mind we raised the following question, ***What are the performances of a Question answering systems when they are applied to a Portuguese educational environment?***

To answer this question, we outline this study, first of all, aiming to understand more about the algorithms, techniques, solutions and paradigms that circles QA systems. Our goal with that, was establish a reliable process for the choice of the algorithms that better fits to our needs. To achieve this goal we developed a systematic literature review (SLR).

The SLR showed us that natural language processing (NLP) paradigms would be better suitable to our needs than the other, once it is enabled to multilingual implementations and achieved great results. The techniques were also enlighten and we could see

[1] http://trec.nist.gov/

that the ones that are based on deep learning that are capable of learning long-term dependencies.

With the results of SLR, we started our experiment, where we tested our educational corpus in two QA algorithms based on NLP, Improved Dynamic Memory Network (DMN+) and Sequence To Sequence (Seq2Seq). The experiment showed to us that QA systems can be used as an important tool in a educational environment, since the approach based on +DMN reach a 77% rate on answer matching for factoid questions and a 63% match when we consider all types of questions.

This work is composed by 4 chapters. Chapter 2 describes our Systematic Literature Review, in Chapter 3 we detailed our experiment and its results and in Chapter 4 we outline our conclusions about the research.

2 Systematic Literature Review

2.1 Introduction

Question Answering (QA) systems have emerged as powerful platforms for automatically answering questions asked by humans in natural language using either a pre-structured database or a collection of natural language documents Echihabi and Marcu (2003); Grau (2006); Sun and Chai (2007); Lin et al. (2007); Chali et al. (2011); Dwivedi and Singh (2013); Ansari et al. (2016); Lende and Raghuwanshi (2016). In other words, QA systems make it possible asking questions and retrieve the answers using natural language queries Abdi et al. (2016) and may be considered as an advanced form of Information Retrieval (IR) Cao et al. (2010).

With the efforts from academic research, Question Answering is a growing research field worldwide Voorhees and Tice (2000); Wang et al. (2000). The demand for this kind of system increases day by day since it delivers short, precise and question-specific answers Pudaruth et al. (2016). Nevertheless, a systematic approach for understanding the algorithms, techniques and systems around Question Answering is lacking so far. Although previous literature reviews have focused on specific aspects of Question Answering like domain Athenikos and Han (2010); Kolomiyets and Moens (2011), information retrieval paradigm Gupta and Gupta (2012) and hybrid based paradigm Kalyanpur et al. (2012), the relationships between domains, algorithms, techniques and systems have not been established.

We provide an a holistic view of QA according to the literature. We performed a systematic mapping study to enlighten the paradigms, technologies, domains, metrics and concepts that surround this field of research. As result, we identified how the research community addresses the theme, the main paradigms observed, how the approaches fit in different kinds of domains, the results obtained by the implementations of these approaches.

We outline this paper as follows: we explained the main concepts about the theme in section 2.2, then we detailed the systematic literature review process in Section 2.3 and presented the results of the SLR in Section 2.4. We also analyzed the related works found on literature in Section 2.6.

2.2 Background

2.2.1 Question Answering

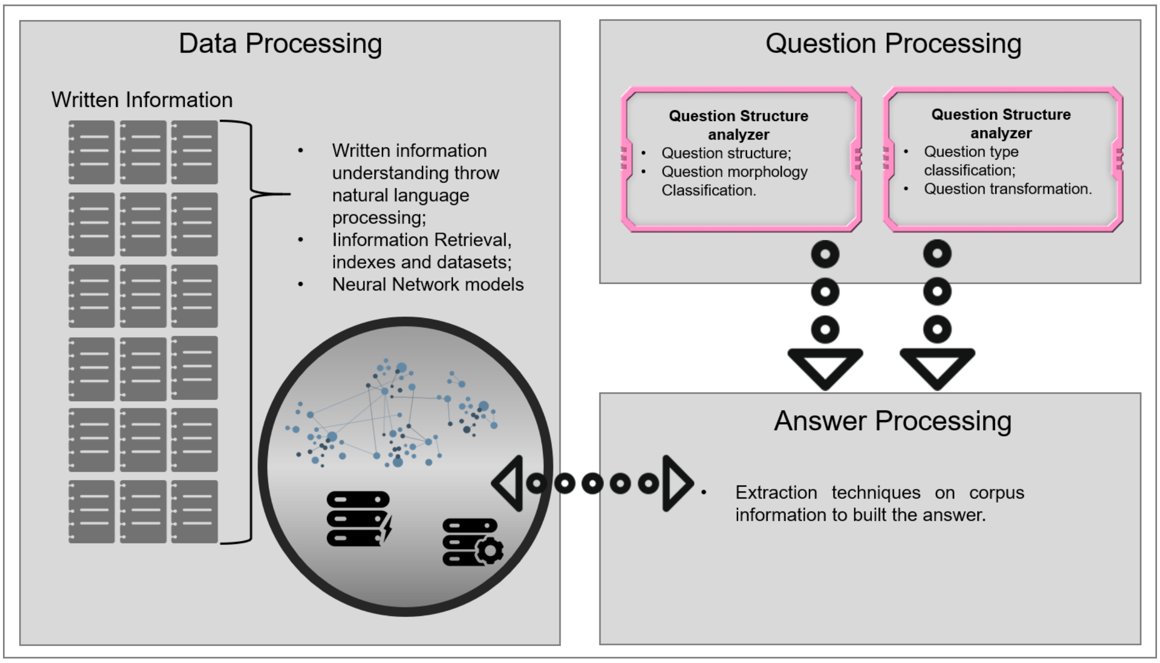
Question Answering systems in information retrieval are tasks that automatically answer the questions asked by humans in natural language using either a pre-structured database or a collection of natural language documents Echihabi and Marcu (2003); Grau (2006); Sun and Chai (2007); Lin et al. (2007); Chali et al. (2011); Dwivedi and Singh (2013); Ansari et al. (2016); Lende and Raghuwanshi (2016). In other words, QA systems enable asking questions and retrieving answers using natural language queries Abdi et al. (2016). Cao et al. (2010) consider QA systems an advanced form of information retrieval. The demand for this kind of system increases on a daily basis since it delivers short, precise and question-specific answers. Pudaruth et al. (2016). With the efforts from academic research, the QA subject has attracts growing interest around the world Voorhees and Tice (2000); Wang et al. (2000) and the main evidence of this is the IBM Watson Ferrucci (2012).

To understand the Question Answering subject, we firstly define the associated terms. A *Question Phrase* is the part of the question that says what is searched. The term *Question Type* refers to a categorization of the question for its purpose. In the literature the term *Answer Type* refers to a class of objects which are sought by the question. *Question Focus* is the property or entity being searched by the question. *Question Topic* is the object or event that the question is about. *Candidate Passage* can broadly be defined as anything from a sentence to a document retrieved by a search engine in response to a question. *Candidate Answer* is the text ranked according to its suitability to as an answer Prager et al. (2007).

Previous studies mostly defined a architecture of Question Answering systems in three macro modules Sucunuta and Riofrio (2010); Vila et al. (2011); Allam and Haggag (2012); Gupta and Gupta (2012); Malik et al. (2013); Bhoir and Potey (2014); Neves and Leser (2015): *Question Processing*, *Document Processing* and *Answer Processing* as showed in Figure 1.

Question Processing receives the input from the user, a question in natural language, to analyze and classify it. The analysis is to find out the type of question, meaning the focus of the question. This is necessary to avoid ambiguities in the answer Malik et al. (2013). The classification is one of the main steps of a QA system. There are two main approaches for question classification, manual and automatic Ray et al. (2010). Manual classification applies handmade rules to identify expected answer types. These rules may be accurate but they are time-consuming, tedious, and non-extensible in nature. There are approaches that classify the question type as *What*, *Why*, *Who*, *How*, *Where* questions and so on Moldovan et al. (2003); Gupta and Gupta (2012). This type of definition helps

Figure 1 – Architecture of the three macro question answering modules. Question processing module classifies the question by its type and morphology. Answer processing module uses the classification and transformation made in question processing module to extract the answer from the result of the Document Processing module that executes previously to create datasets, indexes or neural models.


Data Processing
Written Information
• Written information understanding throw natural language processing;
• Iinformation Retrieval, indexes and datasets;
• Neural Network models
Question Processing
Question Structure analyzer
• Question structure;
• Question morphology Classification.
Question Structure analyzer
• Question type classification;
• Question transformation.
Answer Processing
• Extraction techniques on corpus information to built the answer.

on a better answer detection. Automatic classifications, in contrast, are extensible to new questions types with acceptable accuracy Liang et al. (2007); Ray et al. (2010).

Question processing is divided in two main procedures. The first one is to analyze the structure of the user's question. The second one it to transform the question into a meaningful question formula compatible with QA's domain Hamed and Ab Aziz (2016). Questions can also be defined by the type of answer expected. The types are *factoid*, *list*, *definition* and *complex question* Kolomiyets and Moens (2011). Factoid questions are the ones that ask about a simple fact and can be answered in a few words Heie et al. (2012), for instance, *How far is it from Earth to Mars?*. List Question demands as an answer a set of entities that satisfies a given criteria Heie et al. (2012), *When did Brazil win Soccer World Cups?* illustrates this point clearly. Definition questions expect a summary or a short passage in return Neves and Leser (2015): *How does the mitosis of a cell work?* is a good illustration of it. In contrast, Complex Question is about information in a context. Usually, the answer is a merge of retrieved passages. This merge is implemented using algorithms, such as: *Normalized Raw-Scoring*, *Logistic Regression*, *Round-Robin*, *Raw Scoring* and *2-step RSV* García-Cumbreras et al. (2012).

Different from question processing that is execute on every question asked by the user, *Document Processing* has as its main feature the selection of a set of relevant documents and the extraction of a set of paragraphs depending on the focus of the question or text understanding throw natural language processing Malik et al. (2013). This task can generate a dataset or a neural model which will provide the source for the answer extraction. The retrieved data can be ranked according to its relevance for the question Neves and Leser (2015).

The *Answer Processing* is the most challenging task on a Question Answering system. This module uses extraction techniques on the result of the *Document Processing* module to present an answer Bhoir and Potey (2014). The answer must be a simple answer for the question, but it might require merging information from different sources, summarization, dealing with uncertainty or contradiction.

2.2.1.1 Evaluation Methods

Evaluation methods are part of a Question Answer system. As QA approaches are developed rapidly, reliable evaluation metrics to compare these implementations are needed. According to Yao (2014), the evaluation metrics used in QA are F_1 and accuracy. To understand these measures we have to keep in mind a 2x2 contingency table. For any particular piece of data being evaluated, this table will classify it in two classes, a fragment that was correctly selected (true positive) or was correctly not selected (false negative) and a fragment that was not correctly selected (false positive) or incorrectly not selected (true negative). Using accuracy as a measure metric means the application of the formula 2.1.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + False\ Negative + True\ Negative} \tag{2.1}$$

The issue observed on this metric for Question Answering systems evaluation is the high rates of true negative a system can find, i. e., when a fact question is made, there is one correct answer, anything else would be incorrect and not selected. In this case, a system could have a high calculated accuracy but unmeaningful. To fix this issue, the *f measure* is addressed and it is based on the same 2x2 contingency table and two measures: *Precision* and *Recall*. Precision (Formula 2.2) is the percentage of selected answers that are correct and Recall (Formula 2.3) is the opposite measure, it is the percentage of correct answers selected. Using Precision and Recall, the fact of a high rate of true negative answers is not relevant anymore Kumar et al. (2005); Yao (2014).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{2.2}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{2.3}$$

Analyzing Precision and Recall and knowing that they are opposite, the key concept here is the trade-off researchers must do in each measure, looking for the best metrics to evaluate their systems. Most of fact QA systems should use Recall as a measure metric since it does not matter how high the false positive rates are, if there are high true positive rates, the result will be good. However, for list or definition QA systems, maybe Precision would be better. To balance this trade-off, the *f measure* is presented (Zhang et al., 2008; Yao, 2014) (Formula 2.4).

$$F_1 = \frac{2(Precision.Recall)}{Precision + Recall} \tag{2.4}$$

This measure implements a weighted way of assessing the Precision and Recall trade-off.

There are metrics that can be used to evaluate QA systems, such as Mean Average Precision (MAP) presented on Formula 2.5, which is a standard measure for Information Retrieval, Mean Reciprocal Rank (MRR) showed in Formula 2.7 used to calculate the answer relevance (Li et al., 2007; Zhang et al., 2008; Yao, 2014). This one is mainly used in Information Retrieval paradigms.

$$MAP = \frac{\sum_{Q}^{q=1} AveP(q)}{Q} \tag{2.5}$$

The *AveP*, Average Precision, is given by the equation 2.6:

$$AveP = \frac{\sum_{N}^{k=1} P(k) \times rel(q)}{|\{relevant\}|} \tag{2.6}$$

$$MRR = \frac{1}{N}\sum_{i=1}^{N} RR_{(q_i)} \tag{2.7}$$

2.3 Research

This Systematic Literature Review (SLR) was based on guidelines provides by Okoli and Schabram (2010) and Keele (2007). The review tasks are based on their eight steps, and here we will describe: *Purpose of the Literature Review*, *Searching the Literature*, *Practical Screen*, *Quality Appraisal* and *Data Extraction*.

A exponential growth in written digital information led us to the need for increasingly sophisticated search tools Bhoir and Potey (2014); Pinto et al. (2014). The amount of unstructured data is increasing and it has been collected and stored at unprecedented rates Chali et al. (2011); Bakshi (2012); Malik et al. (2013). The challenge is to create ways to consume this data, extract information and knowledge having an interesting experience in the process. In this context the Question Answering systems emerge, providing a natural language interaction between humans and computers to answer as many questions as possible and enabling the retrieval of these answers from unstructured data sets. We created five research questions, showed in Table 1, to guide this SLR as an attempt to understand how Question Answering systems techniques, tools, algorithms and systems work and perform, how reliable implementing tasks they are and the relationship between QA and Natural Language Processing (NLP).

ID	Question
RQ1	What is the representativeness of each QA paradigm?
RQ2	Which are the QA techniques addressed?
RQ3	Which metrics or indicators are used to compare the different QA algorithms, techniques and systems?
RQ4	Which are the fields in which QA systems and NLP are used?
RQ5	How the relationship between QA systems and NLP is built?

Table 1 – Research Questions.

2.3.1 Data Retrieval

We indexed journals and papers written in English. Besides the language factor, a date filter was applied. Papers published from 2000 up to December in 2016 were accepted. We defined this date criteria based on the observations of Wang et al. (2000). After the first TREC [1] that addressed QA systems in 1999, the number of studies about the theme increased and also represent its evolution. Book chapters were also excluded from the search. The searches were conducted in five digital libraries, ACM Digital Library [2], IEEE Xplore [3], Science Direct - Elsevier [4], Springer Link [5] and Wiley [6]. To execute this task, a conceptual research string was developed containing the main keyword of the theme. We executed the search strings on December 1st 2016 in each digital library and the results are presented in Table 2.

2.3.2 Screening of papers

We applied inclusion and exclusion criteria to be explicit about the studies we considered in our review. We kept in this study papers that satisfy the inclusion and

[1] http://trec.nist.gov/pubs/trec8/t8_proceedings.html
[2] http://dl.acm.org
[3] http://ieeexplore.ieee.org/Xplore/home.jsp
[4] http://www.sciencedirect.com
[5] http://link.springer.com
[6] http://http://onlinelibrary.wiley.com/

Digital Library	Number Of Returned Papers
ACM Digital Library	130
IEEE Xplore	178
Science Direct - Elsevier	860
Springer Link	633
Wiley	21
TOTAL	**1822**

Table 2 – Number of papers retrieved in each Digital Library after search strings execution.

exclusion criteria, described in Table 3. With the practical screening we selected 203 papers from the initial 1822 set.

Inclusion Criteria	Exclusion Criteria
Papers written in English.	Papers written in other languages rather than English.
Academic papers published in conferences or journals.	Duplicated papers found on the digital libraries.
Papers that describe a Question Answering algorithm, technique or system.	Books, thesis, editorials, prefaces, article summaries, interviews, news, reviews, correspondences, discussions, comments, reader's letters and summaries of tutorials, workshops, panels, and poster sessions.
Papers that establish a relationship between Question Answering and Natural Language Processing.	
Papers published until December 1st 2016	
Papers published from 2000 on wards	

Table 3 – Inclusion and Exclusion Criteria defined for screening.

2.3.3 Eligibility

To achieve the commitment to include relevant papers in this review synthesis, we created quality assessment questions (QAQ) shown in Table 4 as proposed by Kitchenham et al. (2009).

Each QAQ was answered with Yes (Y), Partial (P) or No (N). We assigned values to these answers in which Yes = 1.5, Partial = 0.5 and No = 0. From this moment on, 203 papers were read and evaluated according to the Quality Assessment Questions. The articles with scores lower than 0.5 were excluded from the result set. Finally, we selected 124 papers, that are listed in Table 5.

ID	Research Question
QAQ 1	Is the main objective of the paper to deal with the question answering topic?
QAQ 2	Does the paper describe a question answering algorithm, technique or system?
QAQ 3	Is the paper able to establish the relationship between question answering and natural language processing?
QAQ 4	Does the paper describe the results of the proposed QA algorithm, technique or system with an evaluation metric?

Table 4 – Question Assessment Questions.

2.3.4 Data Extraction and Classification

With the studies included in the review identified, we started a data extraction from the papers full readings. We created 6 categories to classify the studies set.

2.4 Results and Discussion

This section describes and discusses the findings from the data extraction and classification activities. The findings are presented in a graphical view and are organized by research question.

Using the categories described in Section 2.3.4, we analyzed the information throw its time line. When we consider this time line of the publications, it is possible to see relevant works published each year in the Question Answering field. When crossing this information with the obtained results these works got by year, showed in Figure 2, we can highlight a variation on its average, which revels that QA paradigm is an open question.

2.5 Discussion

2.5.1 *RQ1* - What is the representativeness of QA paradigm?

To answer this question, we considered the observations brought up by Yao (2014) to classify the papers. From the 124 studies, 33.9% of the papers implemented a knowledge base paradigm, 33.1% implemented a natural language processing paradigm, 29% implemented an information retrieval paradigm and 4% implemented a Hybrid paradigm. This low number of Hybrid Based implementations is in accordance with the definitions by Yao (2014) that says that a Hybrid approach is harder to implement than the others. We observed a slightly higher usage of knowledge base and natural language processing implementations.

As one may notice, knowledge base and natural language processing paradigms are implemented at almost the same rate (Figure 2), so we went further in this relation

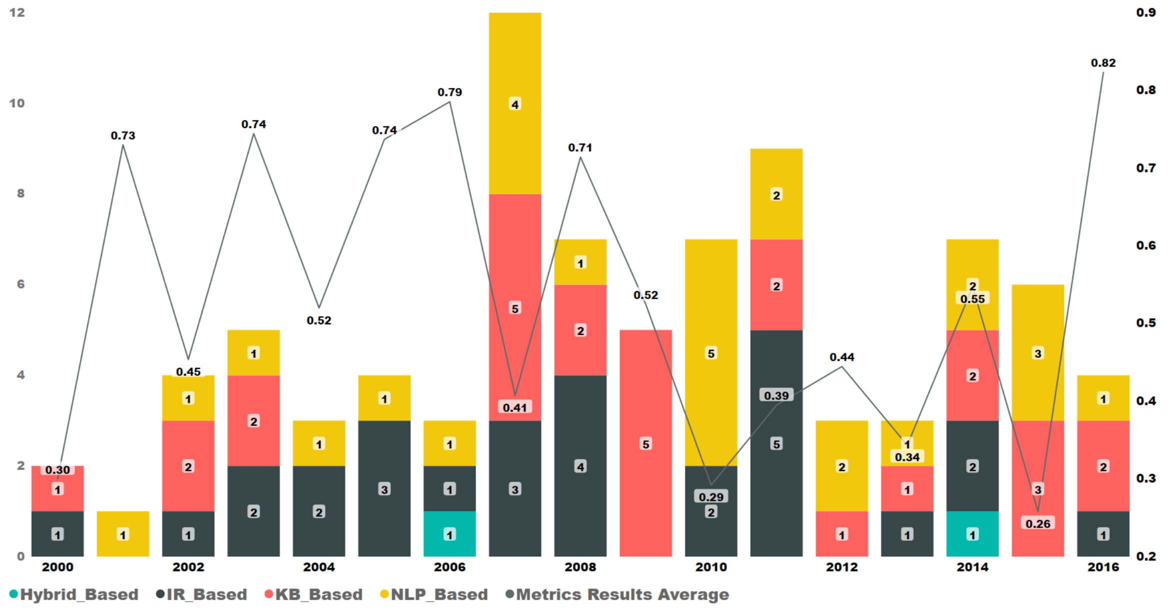
Addressed Paradigm	Papers
NLP	Pack and Weinstein (2001); Li et al. (2002b); Girju (2003); Solorio et al. (2004); Jung et al. (2005); Metzler and Croft (2005); Pustejovsky et al. (2005); Xie and Liu (2005); Saquete et al. (2006); Beg et al. (2007); Liu et al. (2007); Lin et al. (2007); Liang et al. (2007); Ko et al. (2007); Hartawan et al. (2015); Vazquez-Reyes and Black (2008); Mansouri et al. (2008); Wang and Manning (2010); Cao et al. (2010); Surdeanu et al. (2011); Silva et al. (2011); Oh et al. (2011); Moreda et al. (2011); Chali et al. (2011); Pinto et al. (2014); Bhoir and Potey (2014); Bonnefoy et al. (2011); Fikri and Purwarianti (2012); Perera and Perera (2012); Aarabi (2013); Ferrucci et al. (2013); Sagara and Hagiwara (2014); Biswas et al. (2014); Ilvovsky (2014); Chali et al. (2015); Hristovski et al. (2015); Ansari et al. (2016); Lende and Raghuwanshi (2016); Nanda et al. (2016); Pechsiri and Piriyakul (2016); Archana et al. (2016)
IR	Moldovan et al. (2000); Hammo et al. (2002); Moldovan et al. (2003); Dumais (2003); Lin and Katz (2003); Azari et al. (2004); Hammo et al. (2004); Kumar et al. (2005); Li et al. (2005); Radev et al. (2005); Xie and Liu (2005); Yu et al. (2007); Moldovan et al. (2007); Li et al. (2007); Zhang et al. (2008); Wen et al. (2008); Tapeh and Rahgozar (2008); Kosseim and Yousefi (2008); Figueroa and Neumann (2008); Wenyin et al. (2009); Athenikos et al. (2009); Moriceau and Tannier (2010); Balahur et al. (2010); Zong et al. (2011); Richardson et al. (2011); Cao et al. (2011); Vila et al. (2011); Heie et al. (2012); García-Cumbreras et al. (2012); Barskar et al. (2012); Komiya et al. (2013); Kamal et al. (2014); ICA (2014); Pavlić et al. (2015); Bakari et al. (2016); Seena et al. (2016)
KB	Harabagiu et al. (2000); Hsu et al. (2001); Cheng et al. (2002); Li et al. (2002a); Chu-Carroll et al. (2003); Li et al. (2003); Beale et al. (2004); Montero and Araki (2004); Badia (2007); Terol et al. (2007); Lopez et al. (2007); Rabiah et al. (2007); Grau (2006); Rabiah et al. (2008); Guo and Zhang (2009b,a); Guo (2009); Ferrández et al. (2009b,a); Dietze and Schroeder (2009); Varathan et al. (2010); Peng et al. (2010); Furbach et al. (2010); Buscaldi et al. (2010); Malik et al. (2013); Tartir et al. (2011); Al-Nazer and Helmy (2015); Ferrandez et al. (2011); Moré et al. (2012); Kuchmann-Beauger et al. (2013); Spranger and Labudde (2014); Toti (2014); Shekarpour et al. (2015); Abacha and Zweigenbaum (2015); Zayaraz et al. (2015); Yang et al. (2015); Molino et al. (2015); Pudaruth et al. (2016); Hakkoum and Raghay (2016); Abdi et al. (2016)
Hybrid	Delmonte (2006); Frank et al. (2007); Sucunuta and Riofrio (2010); Chandrasekar (2014); Walke and Karale (2013)

Table 5 – 124 Papers addressed in this study. They were categorized by the question answering paradigm implemented.

Category	Description
01	Metadata of the papers, comprising the name of the paper and DOI, the home country of the author, University, the year of publication and the digital library where the paper came from
02	Defined which main approach was selected on the paper. Four paradigms were observed, Natural Language Processing based, Information Retrieval based, Knowledge Base based and Hybrid Based Yao (2014)
03	Was based on natural language processing technique used on the approach Wang et al. (2000); Beg et al. (2007); Cao et al. (2011); Monner and Reggia (2012); Bhoir and Potey (2014); Chandrasekar (2014); Pinto et al. (2014); Chali et al. (2015); Ansari et al. (2016); Hakkoum and Raghay (2016).
04	Dealt with the metrics used to evaluate the accuracy of the question answering systems Solorio et al. (2004); Bonnefoy et al. (2011); Pinto et al. (2014); Hakkoum and Raghay (2016).
05	Determined whether the approach used a training set or which kind of corpora or Data Source was addressed. Examples are, Wordnet[7], Hownet[8], Goi-Taikei[9] and MedLine[10].
06	Detailed classifications such as: language of application, fields of application, usage of a chat bot or dialogue approach to improve the quality of the information from users, information merging algorithms used to answer complex questions and question type Hsu et al. (2001); García-Cumbreras et al. (2012); Biswas et al. (2014); Neves and Leser (2015); Mishra and Jain (2016).

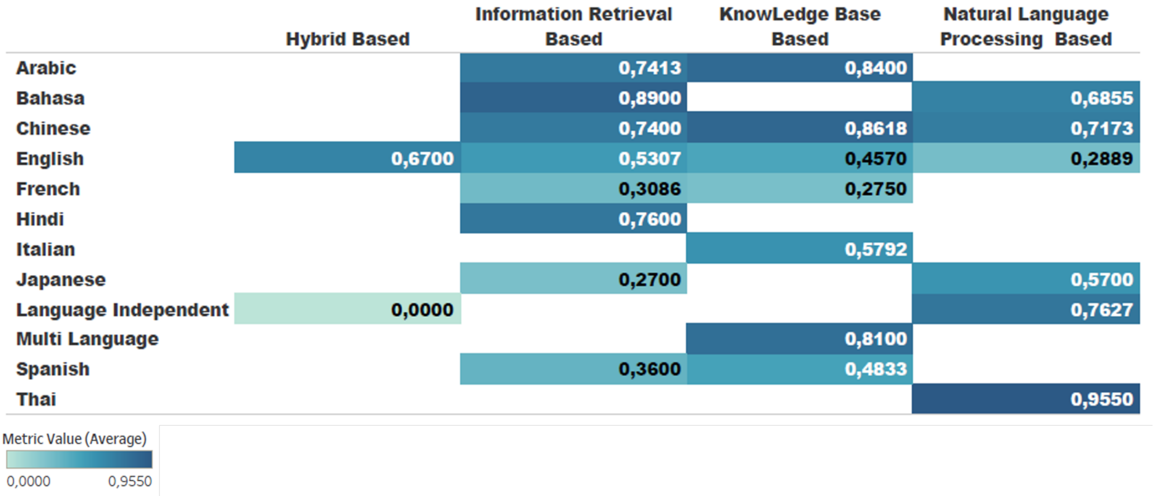
Table 6 – Six Categories created to classify the Studies.

Figure 2 – Publication Time-Line since 2000 and the average of the results obtained by the approaches implementation. The average oscillation over the years show a constantly seek to better option on QA research.



and crossed the amount of implementations of three paradigms (we excluded Hybrid Paradigms due to its lower rates of implementation) against the following dimensions: the first dimension is the language implemented on the approach. The approaches are implemented to understand a natural language question, and answer in the same way. With this analysis, we were able to see that English was the most adopted language and natural language processing paradigms implemented independent language approaches. The second dimension that we had crossed with the representativeness of each QA paradigm was the accuracy each one reaches. We could observe an average accuracy per year higher for natural language processing paradigm in 53% of the cases. Finally, we highlighted in Figure 3 an important finding for this study, which is the good accuracy reached by the implementation when implemented in an independent language approach.

Figure 3 – Relationship between the high accuracy reached and independence language implementations, showing that Natural Language Processing reached best performance.



	Hybrid Based	Information Retrieval Based	KnowLedge Base Based	Natural Language Processing Based
Arabic		0,7413	0,8400	
Bahasa		0,8900		0,6855
Chinese		0,7400	0,8618	0,7173
English	0,6700	0,5307	0,4570	0,2889
French		0,3086	0,2750	
Hindi		0,7600		
Italian			0,5792	
Japanese		0,2700		0,5700
Language Independent	0,0000			0,7627
Multi Language			0,8100	
Spanish		0,3600	0,4833	
Thai				0,9550

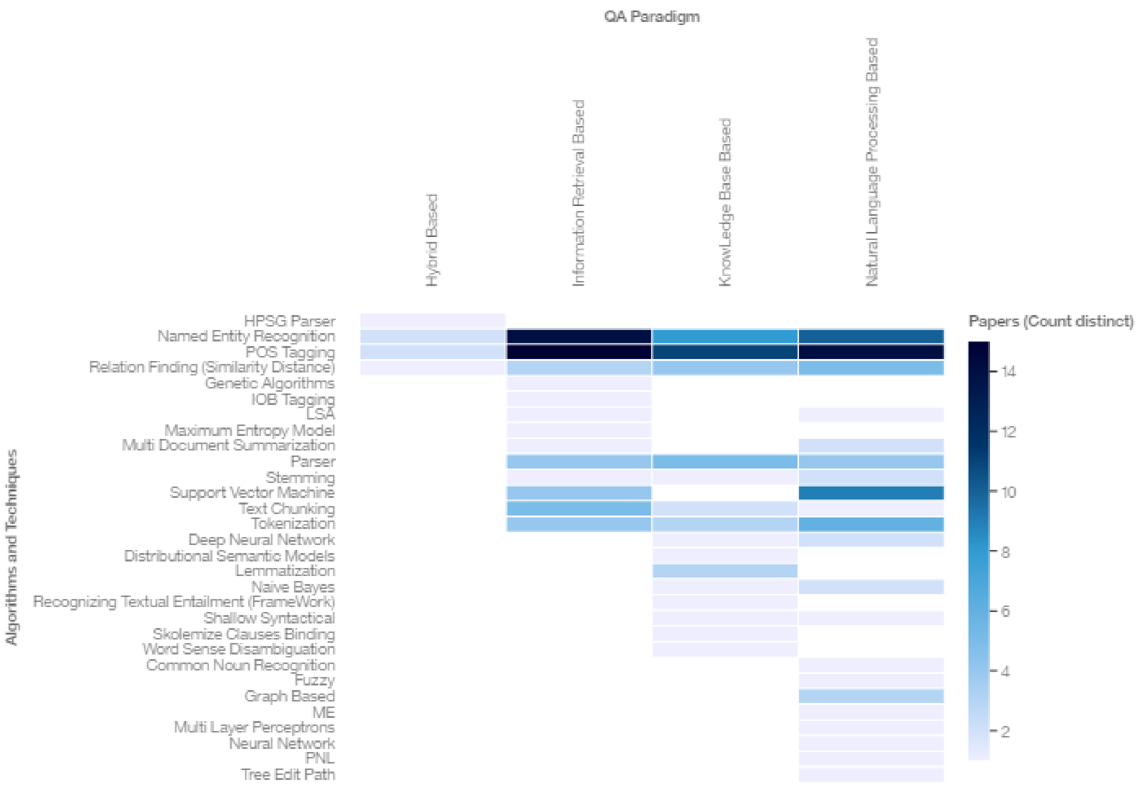
Metric Value (Average)

0,0000 0,9550

2.5.2 *RQ2* - Which are the most frequently applied QA techniques?

To implement the modules of a QA software architecture (*Question Processing*, *Document Processing* and *Answer Processing*), researchers have used techniques, algorithms, frameworks and systems related to information extraction, natural language processing and machine learning. We categorized these tools in order to analyze and summarize their relevance to the study. We clustered, in Figure 4, the 15 most addressed techniques, algorithms, frameworks and tools observed in the SLR by the paradigms they are part of.

Figure 4 – TOP 15 Techniques, algorithms, frameworks and tools observed in SLR.



2.5.3 *RQ3* - Which metrics or indicators are used to compare different Question Answering algorithms, techniques or systems?

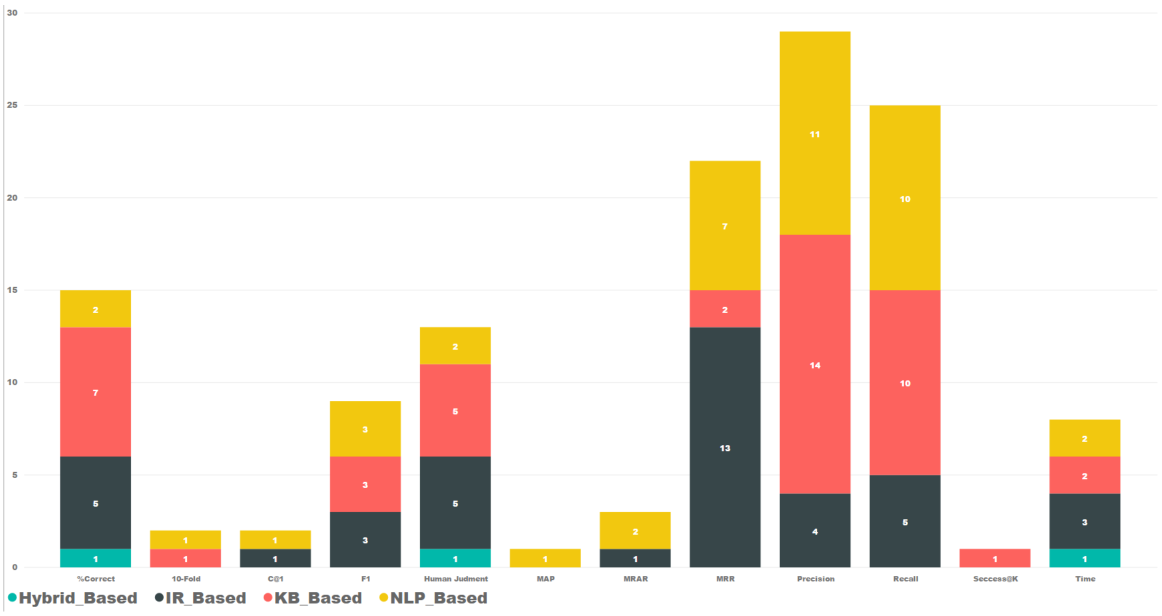
In our research, we extracted the metrics that researchers used to evaluate their implementations in each works. In Figure 5, we clustered the identified metrics by the QA paradigms.

We are able to understand from this analysis that since natural language processing paradigm had the best average evaluation performance over the years, the *Precision* and *Recall* metrics are indicated to evaluate this kind of implementation.

The elapsed time performed by the approach on answer extraction and user delivery was not evaluated by most of researchers, only 6% of papers addressed this metric.

It is important to observe how the evaluation was performed in the approaches when analyzing the obtained results. In order to provide a comparison among implementations, a common dataset for training and testing is crucial. We crossed metrics extracted by the usage of a testing or training set provided by QA conference, such as *TREC* Pack and Weinstein (2001); Moldovan et al. (2003); Chu-Carroll et al. (2003); Girju (2003); Lin and Katz (2003); Li et al. (2005); Lin and Demner-Fushman (2006); Kunichika et al.

Figure 5 – QA paradigms and their metrics.



(2007); Kosseim and Yousefi (2008); Hartawan et al. (2015) and *CLEF* Figueroa and Neumann (2008); Buscaldi et al. (2010) , and we found out that 24% of the approaches used a training or test set like that.

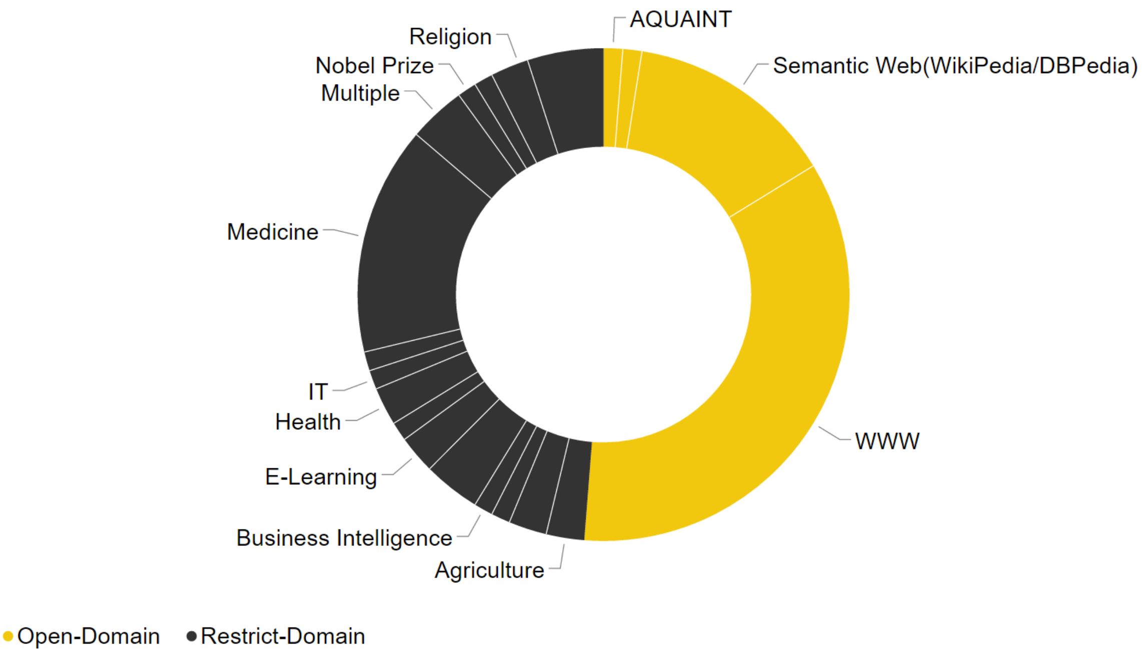
2.5.4 *RQ4* - Which are the fields in which Question Answering systems and Natural Language Processing are being applied?

Question Answering has been implemented in different fields of knowledge. The domains in which QA systems are implemented can be divided in *Open-Domain*, *Restrict-Domain* and *Closed-Domain*. Our research summarized where and in which context researchers implemented their systems. We did not find any applications in *Closed-Domain*. We found out with this analysis that *Open-Domain* based on World Wide Web implementations are the biggest part of the researches and medicine subject is also treated at an important rate, Figure 6 reveals that. From this analysis, we see that question answering intersects with many areas and domains, showing how this systems can be important on knowledge extraction for any kind of user and need.

2.5.5 *RQ5* - How the relationship between Question Answering systems and Natural Language Processing is built?

When we analyze the QA implementations starting from 2000, it is possible to see the paradigms in each approach. The paradigms are named over their main task, but they use natural language processing techniques whether to classify the question or to retrieve

Figure 6 – Fields where researcher are implementing their approaches. There is a wide set of fields, besides the amount of implementations in open domain on *www* and medicine on restricted domain.



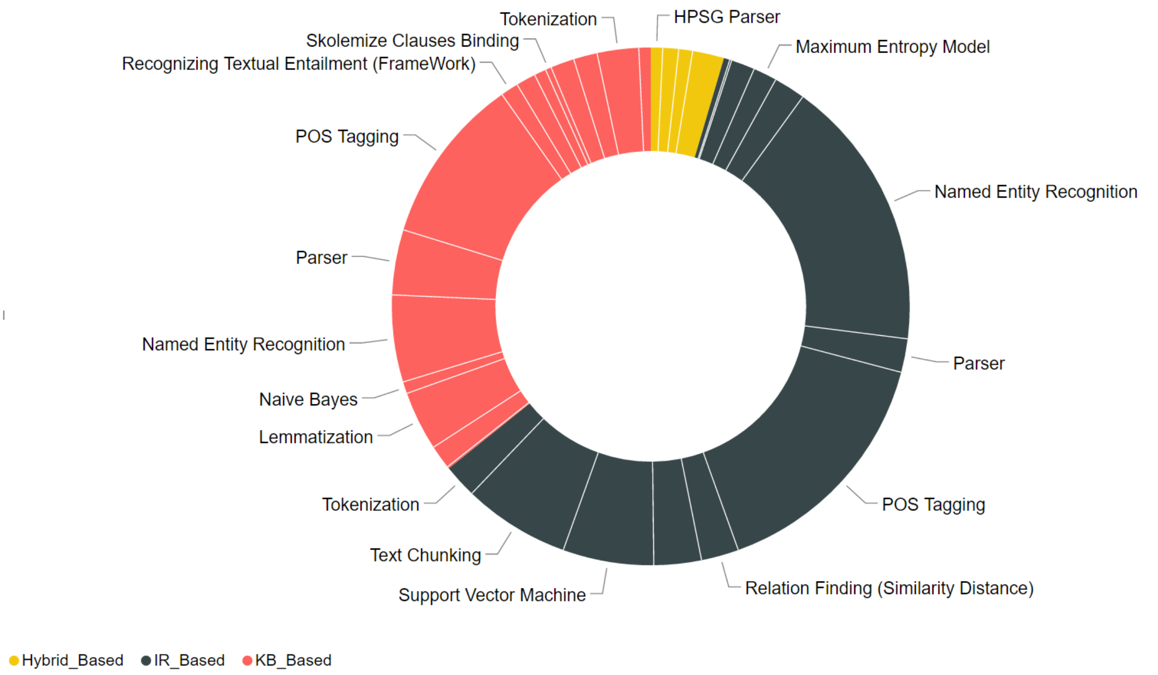
the answer. In other words, NLP paradigm is not the only relationship between QA and natural language processing. Information Retrieval and Knowledge Base make use of NLP techniques that can help on the implementation of these paradigms. Techniques such as POS Tagging, Tokenization, Named Entity Recognition, Semantic Parser and Similarity Distance are described in papers which addressed IR or KB paradigms as we can see on Figure 7. This analysis shows us how natural language processing is important for QA systems, they are essential to create the understanding between the user and the machine and most part of the researcher are using NLP on their approaches.

2.6 Related Work

In the past decades, automated QA has generated interesting of information extraction researchers Maybury (2008). Question Answering systems have become challenging due to the complexity and applicability of these systems. Studies published over the past fifteen years addressed points of view, surveys and reviews about the theme Hamed and Ab Aziz (2016).

From the works we found, we could see how question answering and natural language processing interacts and intersects Hirschman and Gaizauskas (2001) and a comparison on approaches based on natural language processing, information retrieval and question templates, analyzing the differences among the QA approaches, their accuracy

Figure 7 – Natural language processing techniques used in other paradigms besides de NPL based one.



and applicability Andrenucci and Sneiders (2005); Prager et al. (2007).

We found reviews for each kind of paradigm. A knowledge-based approach in order to implement a QA systems in the Bio-medical domain Athenikos and Han (2010); Kolomiyets and Moens (2011), a information retrieval paradigm complete analysis Gupta and Gupta (2012) and a hybrid based paradigm review Kalyanpur et al. (2012). One of the limitations with these studies is that they do not look after all paradigms together, comparing then.

There were works that discuss the techniques, defines the core components and proposes trends for the future of QA field Bouziane et al. (2015); Hamed and Ab Aziz (2016); Mishra and Jain (2016), describes the QA core components and provides a comparison among the Question Answering systems implemented in 16 papers Mishra and Jain (2016).

These works helped on the development of this systematic literature review. This study aims to fill a gap on systematic literature reviews on question answering, analyzing the paradigms and their behavior, making available a evaluation of them to establish when which one can be better used. We considered in this study a classification for the paradigms, their implemented techniques, their metrics and fields of usage, which no other work considered.

3 Experiment

3.1 Introduction

Question Answering (QA) systems are one of the most studied AI subjects in recent years Mollá and Vicedo (2007). Allowing humans to make natural language questions to a computational system that retrieves back a specific answer, this kind of platforms brought a whole new understanding and knowledge to information retrieval subject Echihabi and Marcu (2003); Grau (2006); Sun and Chai (2007); Lin et al. (2007); Chali et al. (2011); Dwivedi and Singh (2013); Ansari et al. (2016); Lende and Raghuwanshi (2016).

When we understand the potential of Question Answering systems, we started to think how they would behave when applied in an educational environment which can benefit greatly of this kind of system. The performance of QA systems can be strongly observed on several studies Kumar et al. (2005); Moldovan et al. (2007); Ferrández et al. (2009a); Cao et al. (2010); Abdi et al. (2016) but there is no study that addressed the Portuguese language. With that in mind, we raised a research question which we intent to solve with this study, What is the performance of a QA system when it is applied to the teaching-learning environment for a Portuguese subject?

Computational techniques have been developed to improve the results of Question Answering systems, in particular with regards to natural language processing. However, there is lack on researches applied to the educational environment.

Section 3.2 details the concepts of Question Answering systems and the used algorithms, Section 3.3 provides the details of our experiment, in Section 3.4 and 3.5 our results and the discussion are enlighten and finally the Section 3.6 brings our conclusion and the future works that should be addressed.

3.2 Background

3.2.1 Question Answering

Question Answering systems are tasks that aim to answer questions asked in natural language using either a pre-structured database or a collection of written information Echihabi and Marcu (2003); Grau (2006); Sun and Chai (2007); Lin et al. (2007); Chali et al. (2011); Dwivedi and Singh (2013); Ansari et al. (2016); Lende and Raghuwanshi (2016). QA systems are an advanced form of information retrieval Cao et al. (2010) and they are usually classified by the paradigm that is implemented and we can define as

Natural Language Processing, *Information Retrieval*, *Knowledge Base* and *Hybrid* Yao (2014).

Overall, there are some studies that provides a basic architecture of question answering systems Sucunuta and Riofrio (2010); Vila et al. (2011); Allam and Haggag (2012); Gupta and Gupta (2012); Malik et al. (2013); Bhoir and Potey (2014); Neves and Leser (2015): *Question Processing*, *Document Processing* and *Answer Processing*.

- *Question Processing:* This module is responsible to analyze question structure and classify its morphology Ray et al. (2010); Malik et al. (2013). Besides that, it classifies the type of the question Moldovan et al. (2003); Liang et al. (2007); Ray et al. (2010); Gupta and Gupta (2012) and perform a question transformation creating a meaningful question formula compatible with QA's domain Hamed and Ab Aziz (2016). The question classifications can be defined as *Factoid*, which are the questions that are arguing about a fact and their answers doesn't use a lot of words Heie et al. (2012), *Definition*, that requires a summary or short passages as a answer Heie et al. (2012), and finally *List* questions that demands for its answer a set of entities that satisfies a given criteria Neves and Leser (2015).

- *Document Processing:* Responsible for written information understanding throw machine learning and deep learning techniques Malik et al. (2013); Neves and Leser (2015).

- *Answer Processing:* Execute extraction techniques on corpus information Bhoir and Potey (2014).

A number of techniques have been developed to implement question answering systems. Our SLR enlighten four paradigms as the bases for the QA systems, natural language processing (NLP), information retrieval (IR), knowledge base (KB) and hybrid. From those results observed in our SLR, we identify the suitable paradigm and approaches for our needs. As we needed an approach for dealing with Portuguese language and high performance on answer retrieval we choose algorithms based on NLP paradigm with approaches based in deep learning.

Once the paradigms and approaches are defined we selected two algorithms do use on our implementation, improved dynamic neural network and sequence to sequence.

3.2.2 Improved Dynamic Memory Network

We begin by outlining where improved dynamic memory networks came from. This approach is an enhancement of Dynamic Memory Network (DMN) Xiong et al. (2016) and represents a neural network architecture improved for QA problems. From a training set of input sequences, that could be a sentence, a story, papers, books and questions, DMN can create episodic memories and use these to postulate consistent answers. Dynamic Memory Networks for question answering is composed by four modules, *Input Module*, *Question Module*, *Episodic Memory Module* and *Answer Module* Anything (2015).

The Input Module is responsible to process the training data. It processes the input vectors associated with a question into a set of vectors termed facts. The module is built using a Gated Recurrent Unit (GRU) that enables the network to learn if the sentence that is being considerate is relevant or there is not related to the answer.

The Question Module process each question word by word, and creates a vector using the same GRU with the same weights as the input module did. At this moment, both facts and questions are encoded as embedding.

Episodic Memory Module has the main goal to retrieve the answer for the question from the input facts. This module is built with two components, the attention mechanism, which is responsible to create a contextual vector, and the memory update mechanism that aims to generate the episode memory based on the contextual vector. Finally the last module is the Answer Module. It is the one responsible for generate an appropriate response.

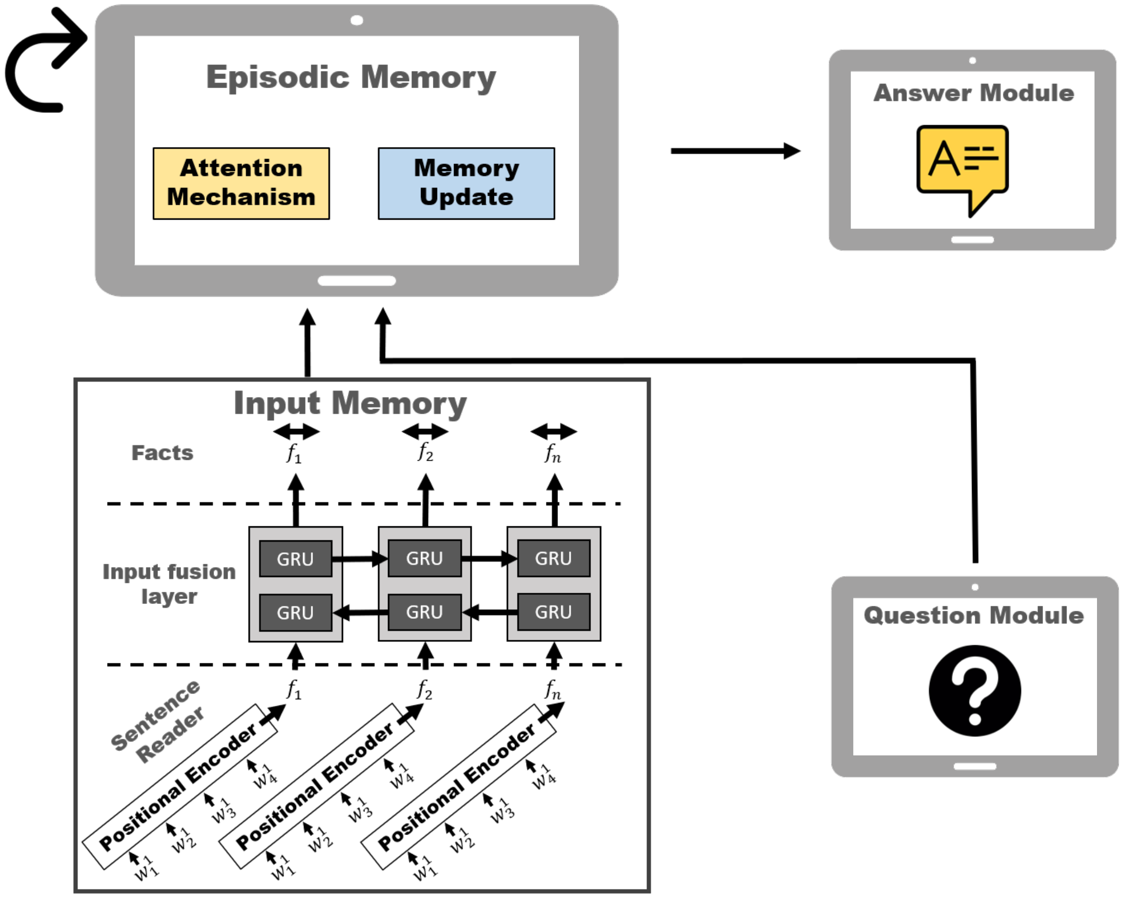
The DMN+ addresses two gaps on DMN. The first one is related to the single GRU. It only allows sentences to have context from sentences before them. The second gap is if the related sentence (which could be the answer for example) is too far away, influencing on the interaction of these distance sentences on the word level GRU. These two issues were treated replacing the single GRU by two new components, a sentence reader and the input fusion layer, making possible the interactions between sentences Xiong et al. (2016). Figure 8 enlightens the main architecture of a DMN+ model.

3.2.3 Sequence To Sequence

Sequence to Sequence is a model based on two recurrent neural networks (RNN), one is the encoder and the other is the decoder.

The encoder acts processing the input sequence and returns its own internal state. The outputs are discarded from the encoder RNN, only recovering the state. This state will serve as the context of the decoder in its own step. The decoder is trained to predict the next characters of the target sequence, given previous characters of the target sequence. Specifically, it is trained to turn the target sequences into the same sequences but offset by one timestep in the future, a training process called "teacher forcing"in this context. Importantly, the encoder uses as initial state the state vectors from the encoder, which is how the decoder obtains information about what it is supposed to generate Stroh et al.; Cho et al. (2014); Sutskever et al. (2014).

Figure 8 – Main architecture of a Question Answering system based on DMN+ approach. The Input Module with the sentence reader and input fusion layer, which differentiates this approach from the traditional DMN implementation.



To use a sequence to sequence model as a basis for a question answering system, it must be trained as follows. The RNN encoder process the story (which is small segments of the corpus), followed by a symbol that determines where the question starts, after that starts the question. Then, another symbol indicates to the network to starts decoding, with the decoder's initial state being the encoder's final. The decoder creates an answer sequence, followed by a STOP symbol which indicates when the processing should end.

3.3 Experiments

The main goal of this study is analyze the performance of question answering approaches when applying it on a teaching-learning process. To achieve this goal we outline this work, taking the following aspects:

- *Understanding and Defining QA Systems*: We developed a systematic literature review which helped us to understand QA definitions, standards and concepts, besides enlighten the paradigms, algorithms and approaches that were better suitable to our research;

- *Domain of Application*: We defined a restrict domain subject where a QA system could be analyzed. We choose Software Engineering subject, a computer science topic. We made this choice due its theoretical nature which fits better to this kind of systems;

- *Approaches Analyze*: We implemented two QA algorithms based on our SLR findings aiming to analyze their performance.

3.3.1 Definition of QA Paradigm and Approaches

The first step in this process was to develop a systematic literature review aiming to clarify question answering concepts and identify which is the most appropriate approach or algorithm that could be used in our research.

In our SLR[1] we found out that approaches based on natural language processing paradigm were more suitable to our needs than the others (Knowledge Base and Information Retrieval paradigms) due its easy customization to different languages and high accuracy rates. Our findings from the SLR leaded us to the bases to select the domain of our system, the QA algorithms used and finally the means to evaluate them.

3.3.2 QA System Domain and Corpus

An essential part of question answering system is the corpus used as its knowledge source for training and testing. This work is based on a restrict domain, focusing it in software engineering field of study. The academic literature on question answering has revealed that the restrict domain gives accurate answer than the open domain QA systems Vila et al. (2011); Lende and Raghuwanshi (2016); Nanda et al. (2016); Tartir et al. (2011). We choose software engineering subject due its theoretical nature which make the question answer system task easier when answering fact, definition or list questions Cheng et al. (2002); Kumar et al. (2005); Tartir et al. (2011); Vila et al. (2011); Moré et al. (2012); Lende and Raghuwanshi (2016).

The experiments were run using a corpus extracted from 11 Software Engineering books described in Table 7. The reasons why these books were selected are:

- They are the most addressed books that Brazilian teachers uses on their Software Engineering lecture;

[1] Submitted in Jan/2018 at *Journal of King Saud University - Computer and Information Sciences*

- Besides concepts, methods and standards of software engineering, there are books dealing with agile methods, a recurrent and important topic on this area;

Book Main Subject	Book
Software Engineering	Wazlawick (2013); PAULA FILHO and PÁDUA; Hirama (2011); Sommerville; Schach (2009); Pressman (1995); PAULA FILHO and PÁDUA
Agile Methods	Poppendieck and Poppendieck (2009); Prikladnicki et al. (2014); Sbrocco and Macedo (2012); Cohn (2000)

Table 7 – Software Engineering books used to create the system corpus.

To the extracted contents of the books we added 4186 questions and answers pairs extracted from internet. The set composed by the books and question answers pairs were used as the training set. Besides that, 90 questions were prepared to be used as the testing set. This set was created by the authors of this study or extracted from teachers questions found at the used books.

3.3.3 Training and Testing sets

3.3.3.1 Training Sets

As described in section 3.3.2, we used as a source of knowledge 11 Software Engineering books, written in Portuguese. To extract the information of each book we defined a process, detailed on Figure 9 seeking to normalize the files, then the data were collected using custom software written in python.

With the books files normalized and ready to information extraction, we executed the custom software built in Python aiming to recover as much information as possible with quality from files. To maximize this task, we established a set of exclusion criteria, designed to guide the software on the content classification, making possible to remove non important information. Table 8 provides the criteria used or our software.

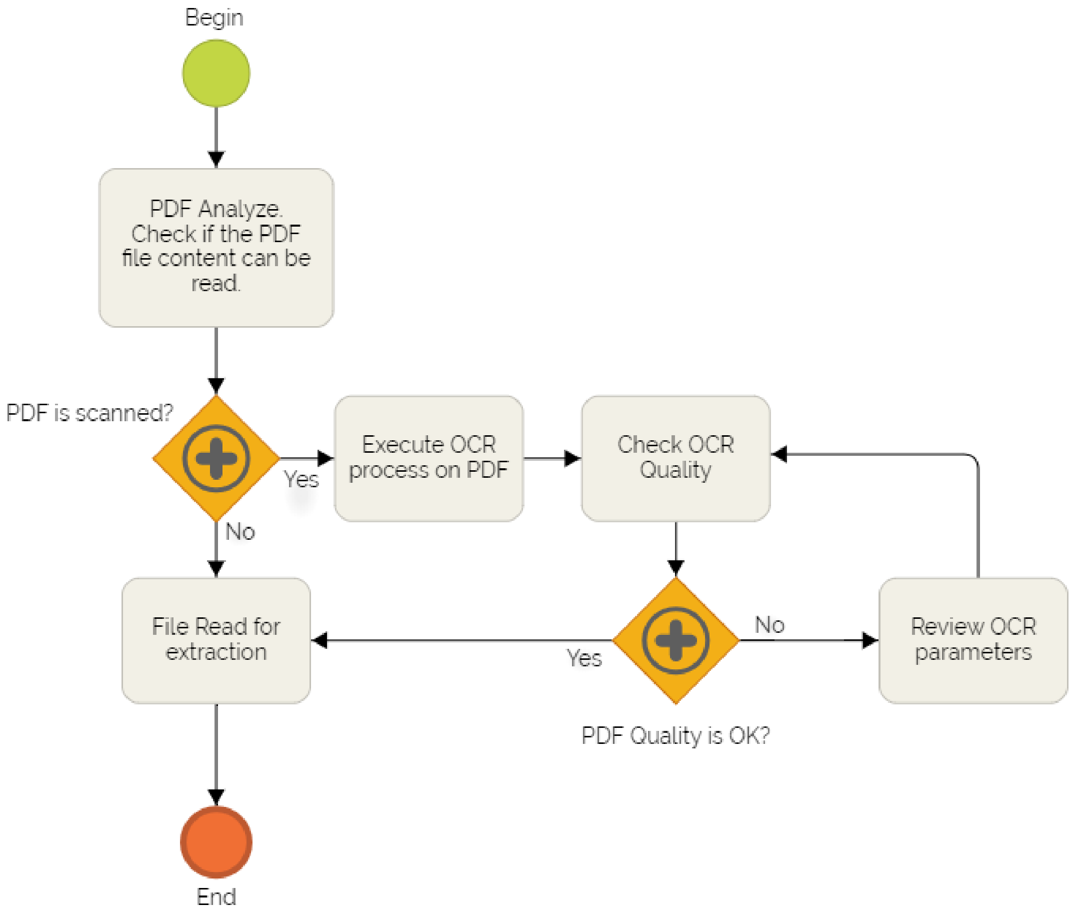
The software was designed to extracted the information in phrases that were saved as a new line in a text file, this made easier our task to deploy a corpus in the specific format requested by each question answering algorithm we used.

By the end of text extraction we ended up with 89198 phrases and a vocabulary composed by 30482.

To improve the corpus quality extracted from the books we designed a training questions set. As we need reliable questions and answers pairs, the data were gathered from a web site [2] that provides Brazilian public tenders content, containing questions

[2] http://www.qconcursos.com

Figure 9 – Process defined to normalize the books files. This task was executed to increase the information quality extracted from all books.



Type of Exclusion	Description
Titles, Headers and Footers	Books names, chapters description, sessions titles, headers and footers information were removed.
Sentences with less than 5 Words	We din not remove all sentences that met this criteria. Phrases that were part of a list such as bullets and numbering and sentences that were a part of the text were kept on the corpus.
Chapter or Section Description	Phrases that describes what each chapter or section talks about were removed.
Summaries and References	All summaries, tables of contents, figure lists and the references were removed.
Others	Page numbers, Proper names (Authors and Co-authors), References and specific phrases (manually identified) were also removed.

Table 8 – Exclusion criteria used to classify sentences extracted from the software engineering books. One achieved criteria is enough to not use the sentence on the corpus.

and answers for several subjects, including software engineering. To download the needed content, we built a second custom software that parsed the pages and save the information in a JSON file that have the description of the question, all possible answers, the correct answer and the metadata of the set. The first step to download the questions and answers was to filter on the web site only the questions related to Information Technology area and Software Engineering subject. After filtering the information we used the resulted URL to start parsing the HTML pages and download the information. On this moment, we retrieved all questions that the web site filtered for us and generated a complete JSON file with all of them. Once the download was finished an other step was initiated to classify the questions regarding their degree of agreement to the needs of our corpus. This step was based in inclusion and exclusion criteria that are listed in Table 9.

Type of Criteria	Description
Inclusion Criteria	Questions that are direct classified as a List, Definition or fact question.
Inclusion Criteria	True or False Questions.
Exclusion Criteria	Questions that have some kind of image analyze on their content.
Exclusion Criteria	Questions with text fragments that required analyze and interpretation.

Table 9 – Inclusion and Exclusion criteria used to classify the question according their compatibility to a corpus set.

By the end of this screening process, we ended up with 4186 questions and answers pairs that were aggregated to the corpus, improving its quality and reliability.

3.3.3.2 Testing Set

In order to test the implemented question answering systems accuracy, we developed a set of 90 testing questions. The aim of this task is to normalize all used data, keeping the differences between the implemented approaches only on its models. The questions were classified according its types Surdeanu et al. (2011); Chali et al. (2011); Bonnefoy et al. (2011); Perera and Perera (2012); Aarabi (2013); Biswas et al. (2014); Ilvovsky (2014); Chali et al. (2015); Hristovski et al. (2015); Ansari et al. (2016); Lende and Raghuwanshi (2016); Nanda et al. (2016); Archana et al. (2016). In Table 10 we described the amount of questions by their types and also provides some examples that we used.

3.3.4 Implemented Approaches

We executed two QA algorithms based on natural language processing paradigms, improved dynamic memory network (DMN+) and sequence to sequence (Seq2Seq). Each

Type Of Question	Amount of Questions	Used Examples
Fact Questions	40	• *"Quem originalmente propos o modelo Espiral?"* • *"Como ficou conhecido o período da década de 1960 até meados da década de 1980?"* • *"Qual é um exemplo da crise de software dos anos 1960?"*
Definition Questions	25	• *"Qual é o papel do Engenheiro de Software?"* • *"O que é modelo de processo?"* • *"O que é projeto?"*
List Questions	25	• *"Quais são os princípios da engenharia de software?"* • *"Quais são as vantagens em definir o desenvolvimento de software como um processo?"* • *"Como são classificados os mitos do software?"*

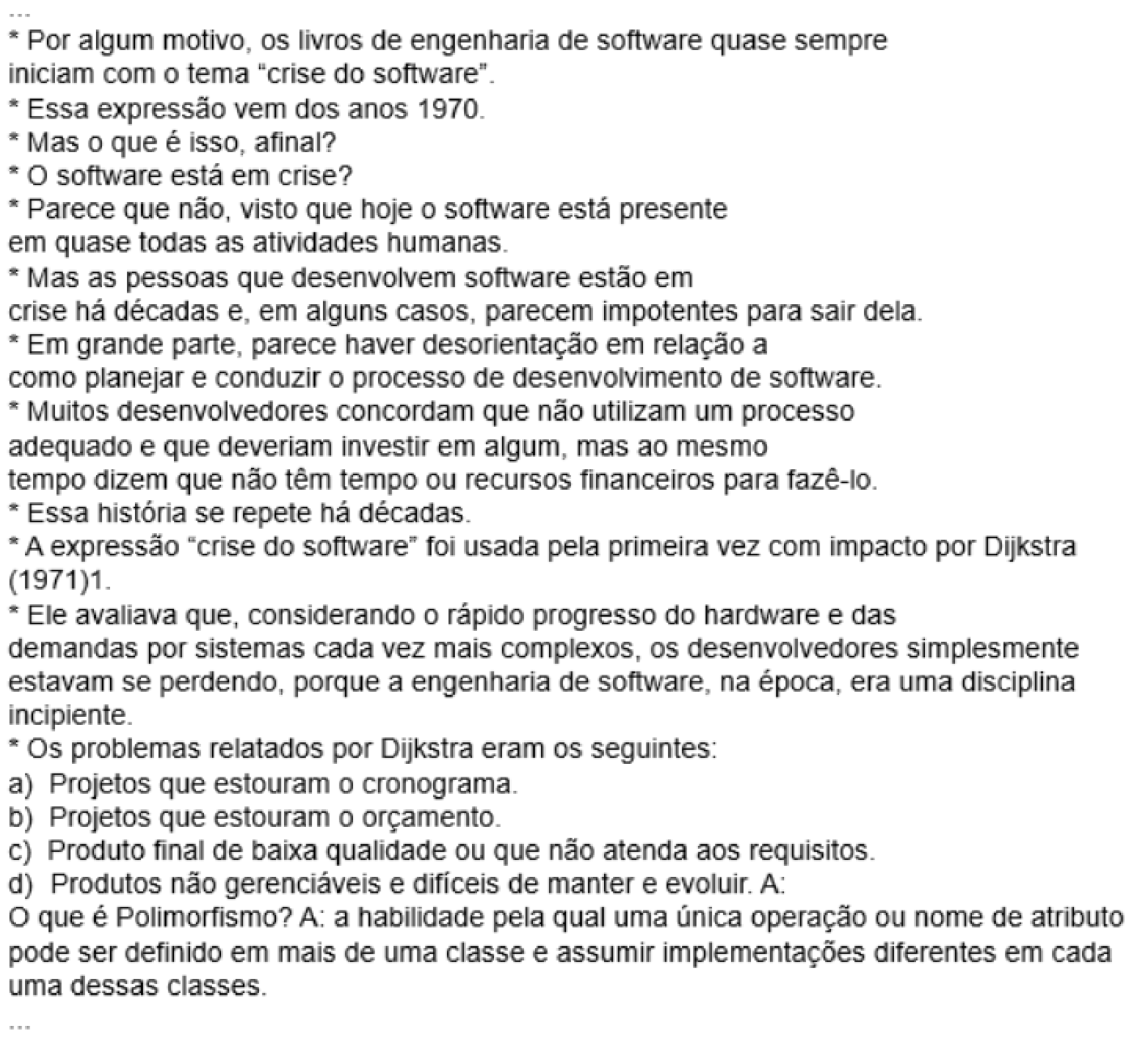
Table 10 – Amount of questions by its type used as testing set on our implemented approaches.

approach needed a specific formatted file for it's training which was created by the authors.

3.3.4.1 Improved Dynamic Memory Network

This is one of the most widely-used algorithms for question answering tasks when addressing a natural language processing Stroh et al.; Xiong et al. (2016); Kumar et al. (2016). For this approach we created a training file based on our extracted corpus and it was loaded to the Input and Question module. On the input module the data was loaded by sentences to be encoded into distributed vector representations. The question module encodes the questions extracted from the training set into a distributed vector representation. Figure 10 displays an examples of how data was prepared for this approach training.

Figure 10 – Examples of how the training data was prepared for training the DMN+ approach.



```
...
* Por algum motivo, os livros de engenharia de software quase sempre
iniciam com o tema “crise do software”.
* Essa expressão vem dos anos 1970.
* Mas o que é isso, afinal?
* O software está em crise?
* Parece que não, visto que hoje o software está presente
em quase todas as atividades humanas.
* Mas as pessoas que desenvolvem software estão em
crise há décadas e, em alguns casos, parecem impotentes para sair dela.
* Em grande parte, parece haver desorientação em relação a
como planejar e conduzir o processo de desenvolvimento de software.
* Muitos desenvolvedores concordam que não utilizam um processo
adequado e que deveriam investir em algum, mas ao mesmo
tempo dizem que não têm tempo ou recursos financeiros para fazê-lo.
* Essa história se repete há décadas.
* A expressão “crise do software” foi usada pela primeira vez com impacto por Dijkstra
(1971)1.
* Ele avaliava que, considerando o rápido progresso do hardware e das
demandas por sistemas cada vez mais complexos, os desenvolvedores simplesmente
estavam se perdendo, porque a engenharia de software, na época, era uma disciplina
incipiente.
* Os problemas relatados por Dijkstra eram os seguintes:
a) Projetos que estouram o cronograma.
b) Projetos que estouram o orçamento.
c) Produto final de baixa qualidade ou que não atenda aos requisitos.
d) Produtos não gerenciáveis e difíceis de manter e evoluir. A:
O que é Polimorfismo? A: a habilidade pela qual uma única operação ou nome de atributo
pode ser definido em mais de uma classe e assumir implementações diferentes em cada
uma dessas classes.
...
```

To run our experiment we used the algorithm implemented in Python based on

Xiong et al. (2016) findings.

3.3.4.2 Sequence To Sequence

This implementation consists of two recurrent neural networks (RNN), one to work as the encoder and another one as the decoder Cho et al. (2014); Sutskever et al. (2014); Bahdanau et al. (2014); Vinyals and Le (2015). On the encoder RNN we loaded the sequences throw a training file formatted with sentences by line, with that, the encoder would ideally capture the semantic summary of the input sequence. Based on this context, the decoder generates the output sequence.

3.4 Results

The standard approach to this section is to present and describe the results in a systematic and detailed way. The results were obtained after the submission of the testing set to each implemented approach. The answers retrieved by them were analyzed manually by the authors and the data were evaluated and detailed here.

The three key results of this experimental are:

- The approach based on Improved Dynamic Memory Networks reach better results than Sequence To Sequence approach.
- Question Answering systems based on natural language processing can reach interesting results.
- Question Answering systems can be used as a tool to support a teaching-learning process.

Figure 11 presents the results obtained by each executed approach regarding the answers classifications we made.

What stands out in this figure is how the approach based on improved dynamic memory networks performed better than Sequence To Sequence based approach. The DMN+ implementation answered 57 question correctly while Sequence To Sequence approach answered correct 41 questions. DMN+ had a better performance on the other classifications either, it answered less non correct questions than the sequence to sequence approach and had retrieved only 7 wrong answers, other than double wrong answers responded by Seq2Seq.

The differences between DMN+ and Seq2Seq are highlighted in Figure 12.

As shown in Figure 12, the DMN+ approach answered correctly 31 fact questions against 21 answered by Seq2Seq. For the definition questions, DMN+ answered 14

Figure 11 – Results achieved by each approach when executed with our corpus.

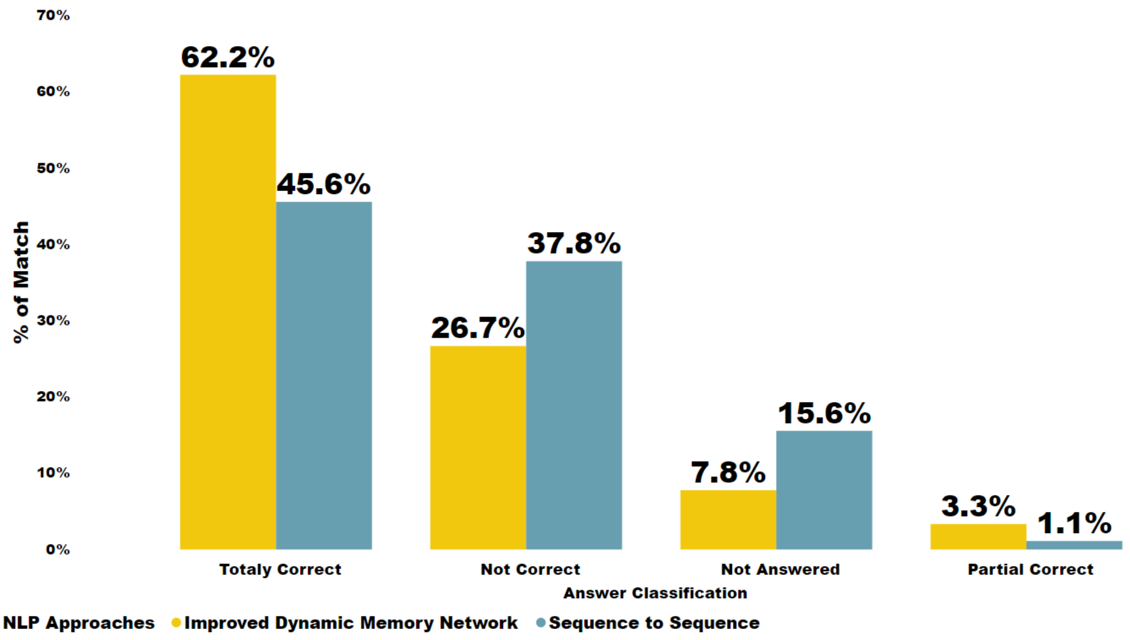
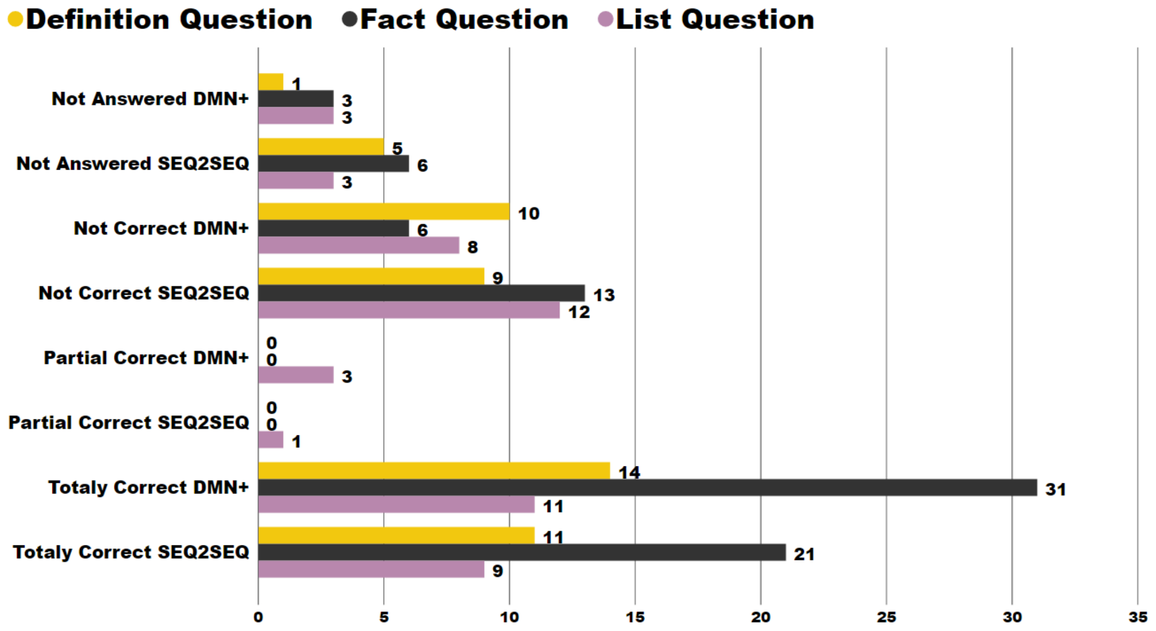


Figure 12 – Differences between DMN+ and Seq2Seq when analyzing their performances.



questions correctly and Seq2Seq answered 11. When we analyze the correct answers for list questions provided by each approach, we can see that DMN+ answered in a correct manner 11 questions and Seq2Seq only 9 questions.

Regarding wrong answers provided by the approaches, we can see that Seq2Seq had a better performance on retrieving less not correct answers when we are dealing with definition questions, on other hand, analyzing the amount of wrong answers retrieved for fact and list questions, DMN+ had better results than Seq2Seq.

Analyzing the amount of questions that didn't retrieve any answer, we also observed that DMN+ performed better, this approach didn't answer 1 definition questions, 3 fact questions and 3 list questions, on the other hand, seq2seq didn't answer 5 definition questions, 6 fact questions and 3 list questions.

Besides the amount of answered questions by their types, we analyzed the elapsed time on each model training and on their answer retrieval.

Regarding the time spent to train the models, Table 11 provides the results obtained on each model training. We can see that DMN+ spent almost 8 hours more to finish the training.

NLP Approach	Time Elapsed (HH:MM:SS)
Improved Dynamic Memory Networks	38:26:53
Sequence To Sequence	30:37:24

Table 11 – Time spent for each approach on model training. Both models were trained in the same conditions of hardware and basic software settings.

Figure 13 displays the average elapsed time that each implemented approach reached. It is possible to observe that DMN+ spent almost 5 times more time than Seq2Seq to retrieve their answers.

When we analyze the average elapse time classified by the question and answer types, it is possible to identify the huge difference among then. In special, the average time took by DMN+ to retrieve correct fact questions in relation to SeqToSeq.

3.5 Discussion

One of the mail goals of this experiment was to analyze the performance of two question answering algorithms, based on natural language processing paradigm when applied to a restrict domain. In our results we could see that Improved Dynamic Memory Networks performed better than Sequence To Sequence model, particularly as concerns when the models are answering questions of fact type.

Figure 13 – Average Elapse time reached by each approach classified by the answers type.

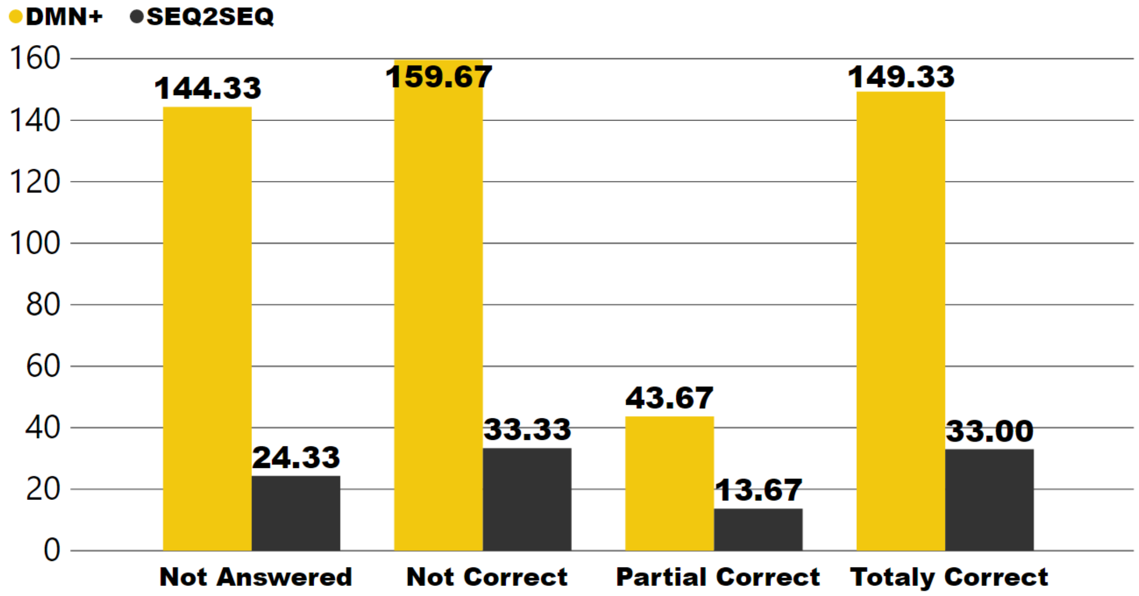
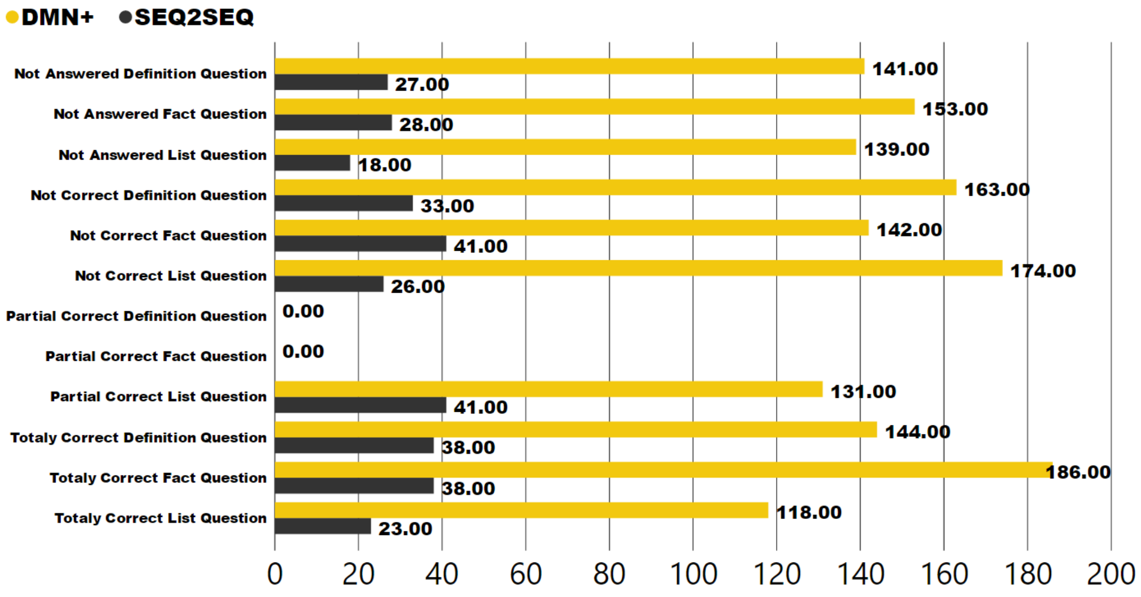


Figure 14 – Average Elapse time reached by each approach classified by the answers type and the question type.



Another important finding was that we could see an important adherence of question answering systems to a educational environment as a tool to support a teaching-learning process.

Other authors analyzed these two models in questions answering systems as well. There are similarities between the results expressed by this work and those described by (Stroh et al.; Bahdanau et al., 2014; Cho et al., 2014; Sutskever et al., 2014; Vinyals and Le, 2015; Xiong et al., 2016; Anything, 2015; Raghuvanshi and Chase; Kumar et al., 2016), however, there is a important difference between their approaches and ours, the data used as Corpus. As we detailed in section 3.3.3, we used as a corpus, written information about software engineering while the other studies based their experiments in Facebook bAbi data[3] and the DAtaset for QUestion Answering on Real-world (DAQUAR).

Analyzing the results obtained by the other studies we could see that they had a better performance on answering fact questions Raghuvanshi and Chase; Xiong et al. (2016); Kumar et al. (2016). We concluded that this difference of performance is related to how the corpus were prepared for the task, the data sets used by then are specially orientated for a fact question answering algorithm. Although we have created our training files based on our corpus, the nature of our data extracted from the books and training questions brought to us difficulties on mounting the tasks for each model. This issue related to the training file development lead us to embrace the DMN+ algorithm, the substitution of the single GRU for the sentence layer and Input fusion layer allowed us to create our tasks with questions that could be answered in a segment that appears before or after it Xiong et al. (2016). The article provide a valuable description of Dynamic Memory Networks and why the improvement was necessary.

The elapse time on model training and answer retrieval was not addressed by the other works, so this is was not compared in this study.

As the creation of training files became an key task on this kind of application, more research on this topic need to be undertaken, aiming to facilitate make available QA systems for different subjects

3.6 Conclusion and Future Work

The aim of this present research was to analyze question answering algorithms based on natural language processing paradigm and determine if they can be used as an appliance to support a teaching-learning process. This is the first study that has used Question answering algorithms that addresses Portuguese language in a tool to support an educational environment.

[3] https://research.fb.com/downloads/babi/

The current study found that in our context, improved dynamic memory networks performed better than sequence to sequence algorithm. Despite the better training and answer retrieving elapse time of Seq2Seq, their results are less interesting than the ones obtained by DMN+, specially when we consider Fact questions.

The findings of this study suggest that question answering system can be used as a tool on a teaching-learning support process. We observed a 63% of match for the DMN+ algorithm when we consider all types of question. If we consider only the questions classified as Fact, the performance is even better, reaching over 77% of match. These research highlight the potential usefulness of question answering system in a educational environment, providing support on accurate knowledge extraction.

There is, therefore, a definite need to create a corpus that can be reliable and complete enough to be used as a source to the QA system. This task is not easy and requires a major effort on retrieve and storage reliable information regarding to the domain selected for the system.

Further studies needs to attempt on the performance of algorithms that aims to answer complex question, the ones that need to be interpreted. There are several approaches which claims to solve this Chali et al. (2011); Neves and Leser (2015); Chali et al. (2015); Pudaruth et al. (2016); Ansari et al. (2016) and they should be analyzed seeking for improvement on a tool that can be used in a educational environment. Besides that, more research on the corpus creation should be addressed seeking to make easier the availability of QA systems for other subjects.

4 Conclusion

The purpose of the current study was to enlighten the Question Answering subject in terms of its algorithms, approaches, paradigms, evaluation metrics, fields of application and others dimensions. Besides of that, we wanted to analyze how these question answering approaches would perform when applied to an educational environment.

These study has shown that question answering systems have better accuracy when they are based on natural language processing paradigm and also, this paradigm is easier adapted to different languages. This is an important result since we didn't find any study in Portuguese that could help us in our work. From the NLP paradigms, we could see that algorithms based on deep learning could bring better results, specially the ones capable of learning long-term dependencies. From that, we chose Improved Dynamic Memory Networks (DMN+) and Sequence To Sequence (Seq2Seq) models to analyze their performance with our corpus. Aiming to evaluate these two approaches in a educational approach, we select software engineering subject, a field from Computer Science as our source of knowledge. We selected this subject due its theoretical nature, where we could extract questions that could be classified as factoid, definition or list.

After the execution of the approaches with our software engineering corpus we analyzed the results and the findings clearly indicates that DMN+ model performed better than Seq2Seq model. If we analyze all types of questions, DMN+ reached a 63% match on correct answers while Seq2Seq reached 45%. Regarding only to fact questions, DMN+ answered correct 77% of the questions, against 52% for Seq2Seq. With this results, and the difference between the correct answers of DMN+ and Seq2Seq, we could concluded that the improvement made on Dynamic Memory Network substituting its Input Module by a Sentence Reader and the Input Fusion layer which gave to the approach a broadly analysis on the corpus made the difference on its performance.

During our study, we identify a limitation for these kind of approaches. The creation of a formatted corpus is a difficult task, either because it requires vast and reliable data sources or because of the high effort involved to format this data source into training files. We suggest as future works studies that could make easier the development of the corpus, these would bring agility when making QA systems available for other subjects. Another possible future work could be on answering complex questions, the ones that need interpretation to be retrieved.

References

(2014). Shallow parsing natural language processing implementation for intelligent automatic customer service system.

Aarabi, P. (2013). Virtual cardiologist—a conversational system for medical diagnosis. In *Electrical and Computer Engineering (CCECE), 2013 26th Annual IEEE Canadian Conference on*, pages 1–4. IEEE.

Abacha, A. B. and Zweigenbaum, P. (2015). Means: A medical question-answering system combining nlp techniques and semantic web technologies. *Information Processing & Management*, 51(5):570–594.

Abdi, A., Idris, N., and Ahmad, Z. (2016). Qapd: an ontology-based question answering system in the physics domain. *Soft Computing*, pages 1–18.

Al-Nazer, A. and Helmy, T. (2015). Personalizing health and food advices by semantic enrichment of multilingual cross-domain questions. In *GCC Conference and Exhibition (GCCCE), 2015 IEEE 8th*, pages 1–6. IEEE.

Allam, A. M. N. and Haggag, M. H. (2012). The question answering systems: A survey. *International Journal of Research and Reviews in Information Sciences (IJRRIS)*, 2(3).

Andrenucci, A. and Sneiders, E. (2005). Automated question answering: Review of the main approaches. In *Information Technology and Applications, 2005. ICITA 2005. Third International Conference on*, volume 1, pages 514–519. IEEE.

Ansari, A., Maknojia, M., and Shaikh, A. (2016). Intelligent question answering system based on artificial neural network. In *Engineering and Technology (ICETECH), 2016 IEEE International Conference on*, pages 758–763. IEEE.

Anything, A. M. (2015). Dynamic memory networks for natural language processing. *Kumar et al. arXiv Pre-Print*.

Archana, S., Vahab, N., Thankappan, R., and Raseek, C. (2016). A rule based question answering system in malayalam corpus using vibhakthi and pos tag analysis. *Procedia Technology*, 24:1534–1541.

Athenikos, S. J. and Han, H. (2010). Biomedical question answering: A survey. *Computer methods and programs in biomedicine*, 99(1):1–24.

Athenikos, S. J., Han, H., and Brooks, A. D. (2009). A framework of a logic-based question-answering system for the medical domain (loqas-med). In *Proceedings of the 2009 ACM symposium on Applied Computing*, pages 847–851. ACM.

Azari, D., Horvitz, E., Dumais, S., and Brill, E. (2004). Actions, answers, and uncertainty: a decision-making perspective on web-based question answering. *Information processing & management*, 40(5):849–868.

Badia, A. (2007). Question answering and database querying: Bridging the gap with generalized quantification. *Journal of Applied Logic*, 5(1):3–19.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bakari, W., Bellot, P., and Neji, M. (2016). Aqa-webcorp: Web-based factual questions for arabic. *Procedia Computer Science*, 96:275–284.

Bakshi, K. (2012). Considerations for big data: Architecture and approach. In *Aerospace Conference, 2012 IEEE*, pages 1–7. IEEE.

Balahur, A., Boldrini, E., Montoyo, A., and Martínez-Barco, P. (2010). Going beyond traditional qa systems: challenges and keys in opinion question answering. In *Proceedings of the 23rd international conference on computational linguistics: Posters*, pages 27–35. Association for Computational Linguistics.

Barskar, R., Ahmed, G. F., and Barskar, N. (2012). An approach for extracting exact answers to question answering (qa) system for english sentences. *Procedia Engineering*, 30:1187–1194.

Beale, S., Lavoie, B., McShane, M., Nirenburg, S., and Korelsky, T. (2004). Question answering using ontological semantics. In *Proceedings of the 2nd Workshop on Text Meaning and Interpretation*, pages 41–48. Association for Computational Linguistics.

Beg, M. S., Thint, M., and Qin, Z. (2007). Pnl-enhanced restricted domain question answering system. In *Fuzzy Systems Conference, 2007. FUZZ-IEEE 2007. IEEE International*, pages 1–7. IEEE.

Bhoir, V. and Potey, M. (2014). Question answering system: A heuristic approach. In *Applications of Digital Information and Web Technologies (ICADIWT), 2014 Fifth International Conference on the*, pages 165–170. IEEE.

Biswas, P., Sharan, A., and Malik, N. (2014). A framework for restricted domain question answering system. In *Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on*, pages 613–620. IEEE.

Bonnefoy, L., Bellot, P., and Benoit, M. (2011). The web as a source of evidence for filtering candidate answers to natural language questions. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 63–66. IEEE Computer Society.

Bouziane, A., Bouchiha, D., Doumi, N., and Malki, M. (2015). Question answering systems: Survey and trends. *Procedia Computer Science*, 73:366–375.

Buscaldi, D., Rosso, P., Gómez-Soriano, J. M., and Sanchis, E. (2010). Answering questions with an n-gram based passage retrieval engine. *Journal of Intelligent Information Systems*, 34(2):113–134.

Cao, Y., Liu, F., Simpson, P., Antieau, L., Bennett, A., Cimino, J. J., Ely, J., and Yu, H. (2011). Askhermes: An online question answering system for complex clinical questions. *Journal of biomedical informatics*, 44(2):277–288.

Cao, Y.-g., Cimino, J. J., Ely, J., and Yu, H. (2010). Automatically extracting information needs from complex clinical questions. *Journal of biomedical informatics*, 43(6):962–971.

Chali, Y., Hasan, S. A., and Joty, S. R. (2011). Improving graph-based random walks for complex question answering using syntactic, shallow semantic and extended string subsequence kernels. *Information Processing & Management*, 47(6):843–855.

Chali, Y., Hasan, S. A., and Mojahid, M. (2015). A reinforcement learning formulation to the complex question answering problem. *Information Processing & Management*, 51(3):252–272.

Chandrasekar, R. (2014). Elementary? question answering, ibm's watson, and the jeopardy! challenge. *Resonance*, 19(3):222–241.

Cheng, J., Kumar, B., and Law, K. H. (2002). A question answering system for project management applications. *Advanced engineering informatics*, 16(4):277–289.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Chu-Carroll, J., Czuba, K., Prager, J., and Ittycheriah, A. (2003). In question answering, two heads are better than one. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 24–31. Association for Computational Linguistics.

Cohn, M. (2000). *Desenvolvimento de software com Scrum: aplicando métodos ágeis com sucesso*. Bookman.

Delmonte, R. (2006). Hybrid systems for information extraction and question answering. In *Proceedings of the Workshop on How Can Computational Linguistics Improve Information Retrieval?*, pages 9–16. Association for Computational Linguistics.

Dietze, H. and Schroeder, M. (2009). Goweb: a semantic search engine for the life science web. *BMC bioinformatics*, 10(10):S7.

Dumais, S. (2003). Data-driven approaches to information access. *Cognitive Science*, 27(3):491–524.

Dwivedi, S. K. and Singh, V. (2013). Research and reviews in question answering system. *Procedia Technology*, 10:417–424.

Echihabi, A. and Marcu, D. (2003). A noisy-channel approach to question answering. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 16–23. Association for Computational Linguistics.

Ferrández, O., Izquierdo, R., Ferrández, S., and Vicedo, J. L. (2009a). Addressing ontology-based question answering with collections of user queries. *Information Processing & Management*, 45(2):175–188.

Ferrandez, O., Spurk, C., Kouylekov, M., Dornescu, I., Ferrandez, S., Negri, M., Izquierdo, R., Tomas, D., Orasan, C., Neumann, G., et al. (2011). The qall-me framework: A specifiable-domain multilingual question answering architecture. *Web semantics: Science, services and agents on the world wide web*, 9(2):137–145.

Ferrández, S., Toral, A., Ferrández, Ó., Ferrández, A., and Muñoz, R. (2009b). Exploiting wikipedia and eurowordnet to solve cross-lingual question answering. *Information Sciences*, 179(20):3473–3488.

Ferrucci, D., Levas, A., Bagchi, S., Gondek, D., and Mueller, E. T. (2013). Watson: beyond jeopardy! *Artificial Intelligence*, 199:93–105.

Ferrucci, D. A. (2012). Introduction to “this is watson”. *IBM Journal of Research and Development*, 56(3.4):1–1.

Figueroa, A. G. and Neumann, G. (2008). Genetic algorithms for data-driven web question answering. *Evolutionary computation*, 16(1):89–125.

Fikri, A. and Purwarianti, A. (2012). Case based indonesian closed domain question answering system with real world questions. In *Telecommunication Systems, Services, and Applications (TSSA), 2012 7th International Conference on*, pages 181–186. IEEE.

Frank, A., Krieger, H.-U., Xu, F., Uszkoreit, H., Crysmann, B., Jörg, B., and Schäfer, U. (2007). Question answering from structured knowledge sources. *Journal of Applied Logic*, 5(1):20–48.

Furbach, U., Glöckner, I., Helbig, H., and Pelzer, B. (2010). Logic-based question answering. *KI-Künstliche Intelligenz*, 24(1):51–55.

García-Cumbreras, M., Martínez-Santiago, F., and Ureña-López, L. (2012). Architecture and evaluation of bruja, a multilingual question answering system. *Information retrieval*, 15(5):413–432.

Girju, R. (2003). Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12*, pages 76–83. Association for Computational Linguistics.

Grau, B. (2006). Finding an answer to a question. In *Proceedings of the 2006 international workshop on Research issues in digital libraries*, page 7. ACM.

Green Jr, B. F., Wolf, A. K., Chomsky, C., and Laughery, K. (1961). Baseball: an automatic question-answerer. In *Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference*, pages 219–224. ACM.

Guo, Q. and Zhang, M. (2009a). Question answering based on pervasive agent ontology and semantic web. *Knowledge-Based Systems*, 22(6):443–448.

Guo, Q.-l. (2009). A novel approach for agent ontology and its application in question answering. *Journal of Central South University of Technology*, 16(5):781–788.

Guo, Q.-l. and Zhang, M. (2009b). Semantic information integration and question answering based on pervasive agent ontology. *Expert Systems with Applications*, 36(6):10068–10077.

Gupta, P. and Gupta, V. (2012). A survey of text question answering techniques. *International Journal of Computer Applications*, 53(4).

Hakkoum, A. and Raghay, S. (2016). Semantic q&a system on the qur'an. *Arabian Journal for Science and Engineering*, 41(12):5205–5214.

Hamed, S. K. and Ab Aziz, M. J. (2016). A question answering system on holy quran translation based on question expansion technique and neural network classification. *Journal of Computer Science*, 12(3):169–177.

Hammo, B., Abu-Salem, H., and Lytinen, S. (2002). Qarab: A question answering system to support the arabic language. In *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*, pages 1–11. Association for Computational Linguistics.

Hammo, B., Abuleil, S., Lytinen, S., and Evens, M. (2004). Experimenting with a question answering system for the arabic language. *Computers and the Humanities*, 38(4):397–415.

Harabagiu, S. M., Paşca, M. A., and Maiorano, S. J. (2000). Experiments with open-domain textual question answering. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 292–298. Association for Computational Linguistics.

Hartawan, A., Suhartono, D., et al. (2015). Using vector space model in question answering system. *Procedia Computer Science*, 59:305–311.

Heie, M. H., Whittaker, E. W., and Furui, S. (2012). Question answering using statistical language modelling. *Computer Speech & Language*, 26(3):193–209.

Hirama, K. (2011). *Engenharia de software: qualidade e produtividade com tecnologia*. Elsevier Brasil.

Hirschman, L. and Gaizauskas, R. (2001). Natural language question answering: the view from here. *natural language engineering*, 7(04):275–300.

Hristovski, D., Dinevski, D., Kastrin, A., and Rindflesch, T. C. (2015). Biomedical question answering using semantic relations. *BMC bioinformatics*, 16(1):6.

Hsu, W.-L., Wu, S.-H., and Chen, Y.-S. (2001). Event identification based on the information map-infomap. In *Systems, Man, and Cybernetics, 2001 IEEE International Conference on*, volume 3, pages 1661–1666. IEEE.

Ilvovsky, D. (2014). Using semantically connected parse trees to answer multi-sentence queries. *Automatic Documentation and Mathematical Linguistics*, 48(1):33–41.

Jung, H., Yi, E., Kim, D., and Lee, G. G. (2005). Information extraction with automatic knowledge expansion. *Information processing & management*, 41(2):217–242.

Kalyanpur, A., Boguraev, B. K., Patwardhan, S., Murdock, J. W., Lally, A., Welty, C., Prager, J. M., Coppola, B., Fokoue-Nkoutche, A., Zhang, L., et al. (2012). Structured data and inference in deepqa. *IBM Journal of Research and Development*, 56(3.4):10–1.

Kamal, A. I., Azim, M. A., and Mahmoud, M. (2014). Enhancing arabic question answering system. In *Computational Intelligence and Communication Networks (CICN), 2014 International Conference on*, pages 641–645. IEEE.

Keele, S. (2007). Guidelines for performing systematic literature reviews in software engineering. In *Technical report, Ver. 2.3 EBSE Technical Report. EBSE*. sn.

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., and Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology*, 51(1):7–15.

Ko, J., Nyberg, E., and Si, L. (2007). A probabilistic graphical model for joint answer ranking in question answering. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 343–350. ACM.

Kolomiyets, O. and Moens, M.-F. (2011). A survey on question answering technology from an information retrieval perspective. *Information Sciences*, 181(24):5412–5434.

Komiya, K., Abe, Y., Morita, H., and Kotani, Y. (2013). Question answering system using q & a site corpus query expansion and answer candidate evaluation. *SpringerPlus*, 2(1):396.

Kosseim, L. and Yousefi, J. (2008). Improving the performance of question answering with semantically equivalent answer patterns. *Data & Knowledge Engineering*, 66(1):53–67.

Kuchmann-Beauger, N., Brauer, F., and Aufaure, M.-A. (2013). Quasl: A framework for question answering and its application to business intelligence. In *Research Challenges in Information Science (RCIS), 2013 IEEE Seventh International Conference on*, pages 1–12. IEEE.

Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., and Socher, R. (2016). Ask me anything: Dynamic memory networks for natural language processing. In *International Conference on Machine Learning*, pages 1378–1387.

Kumar, P., Kashyap, S., Mittal, A., and Gupta, S. (2005). A hindi question answering system for e-learning documents. In *Intelligent Sensing and Information Processing, 2005. ICISIP 2005. Third International Conference on*, pages 80–85. IEEE.

Kunichika, H., Honda, M., Hirashima, T., and Takeuchi, A. (2007). A method for evaluating answers by comparing semantic information in a question and answer interaction. *Systems and Computers in Japan*, 38(7):84–97.

Lende, S. P. and Raghuwanshi, M. (2016). Question answering system on education acts using nlp techniques. In *Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), World Conference on*, pages 1–6. IEEE.

Li, H., Fan, X., and Li, L. (2003). A new approach to design the universal chinese question parser. In *Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003 International Conference on*, pages 671–676. IEEE.

Li, P., Wang, X. L., Guan, Y., and Zhao, Y. M. (2005). Extracting answers to natural language questions from large-scale corpus. In *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on*, pages 690–694. IEEE.

Li, S., Zhang, J., Huang, X., Bai, S., and Liu, Q. (2002a). Semantic computation in a chinese question-answering system. *Journal of Computer Science and Technology*, 17(6):933–939.

Li, W., Srihari, R. K., Li, X., Srikanth, M., Zhang, X., and Niu, C. (2002b). Extracting exact answers to questions based on structural links. In *proceedings of the 2002 conference on multilingual summarization and question answering-Volume 19*, pages 1–9. Association for Computational Linguistics.

Li, X., Hu, D., Li, H., Hao, T., Chen, E., and Liu, W. (2007). Automatic question answering from web documents. *Wuhan University Journal of Natural Sciences*, 12(5):875–880.

Liang, Z., Lang, Z., and Jia-Jun, C. (2007). Structure analysis and computation-based chinese question classification. In *Advanced Language Processing and Web Information Technology, 2007. ALPIT 2007. Sixth International Conference on*, pages 39–44. IEEE.

Lin, H., Yang, Z., and Zhao, J. (2007). Question-answering system based on concepts and statistics. *Frontiers of Electrical and Electronic Engineering in China*, 2(1):23–28.

Lin, J. and Demner-Fushman, D. (2006). Methods for automatically evaluating answers to complex questions. *Information Retrieval*, 9(5):565–587.

Lin, J. and Katz, B. (2003). Question answering from the web using knowledge annotation and knowledge mining techniques. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 116–123. ACM.

Liu, C.-h., Wang, Y.-c., Zheng, F., and Liu, D.-r. (2007). Using lsa and text segmentation to improve automatic chinese dialogue text summarization. *Journal of Zhejiang University-SCIENCE A*, 8(1):79–87.

Lopez, V., Uren, V., Motta, E., and Pasin, M. (2007). Aqualog: An ontology-driven question answering system for organizational semantic intranets. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(2):72–105.

Malik, N., Sharan, A., and Biswas, P. (2013). Domain knowledge enriched framework for restricted domain question answering system. In *Computational Intelligence and Computing Research (ICCIC), 2013 IEEE International Conference on*, pages 1–7. IEEE.

Mansouri, A., Affendey, L. S., Mamat, A., and Kadir, R. A. (2008). Semantically factoid question answering using fuzzy svm named entity recognition. In *Information Technology, 2008. ITSim 2008. International Symposium on*, volume 2, pages 1–7. IEEE.

Maybury, M. (2008). New directions in question answering. In *Advances in open domain question answering*, pages 533–558. Springer.

Metzler, D. and Croft, W. B. (2005). Analysis of statistical question classification for fact-based questions. *Information Retrieval*, 8(3):481–504.

Mishra, A. and Jain, S. K. (2016). A survey on question answering systems with classification. *Journal of King Saud University-Computer and Information Sciences*, 28(3):345–361.

Moldovan, D., Clark, C., Harabagiu, S., and Hodges, D. (2007). Cogex: A semantically and contextually enriched logic prover for question answering. *Journal of Applied Logic*, 5(1):49–69.

Moldovan, D., Harabagiu, S., Pasca, M., Mihalcea, R., Girju, R., Goodrum, R., and Rus, V. (2000). The structure and performance of an open-domain question answering system. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 563–570. Association for Computational Linguistics.

Moldovan, D., Paşca, M., Harabagiu, S., and Surdeanu, M. (2003). Performance issues and error analysis in an open-domain question answering system. *ACM Transactions on Information Systems (TOIS)*, 21(2):133–154.

Molino, P., Lops, P., Semeraro, G., de Gemmis, M., and Basile, P. (2015). Playing with knowledge: A virtual player for "who wants to be a millionaire?" that leverages question answering techniques. *Artificial Intelligence*, 222:157–181.

Mollá, D. and Vicedo, J. L. (2007). Question answering in restricted domains: An overview. *Computational Linguistics*, 33(1):41–61.

Monner, D. and Reggia, J. A. (2012). Neural architectures for learning to answer questions. *Biologically Inspired Cognitive Architectures*, 2:37–53.

Montero, C. A. and Araki, K. (2004). Information-demanding question answering system. In *Communications and Information Technology, 2004. ISCIT 2004. IEEE International Symposium on*, volume 2, pages 1177–1182. IEEE.

Moré, J., Climent, S., and Coll-Florit, M. (2012). An answering system for questions asked by students in an e-learning context. *International Journal of Educational Technology in Higher Education*, 9(2):229–239.

Moreda, P., Llorens, H., Saquete, E., and Palomar, M. (2011). Combining semantic information in question answering systems. *Information Processing & Management*, 47(6):870–885.

Moriceau, V. and Tannier, X. (2010). Fidji: using syntax for validating answers in multiple documents. *Information retrieval*, 13(5):507–533.

Nanda, G., Dua, M., and Singla, K. (2016). A hindi question answering system using machine learning approach. In *Computational Techniques in Information and Communication Technologies (ICCTICT), 2016 International Conference on*, pages 311–314. IEEE.

Neves, M. and Leser, U. (2015). Question answering for biology. *Methods*, 74:36–46.

Oh, H.-J., Sung, K.-Y., Jang, M.-G., and Myaeng, S. H. (2011). Compositional question answering: A divide and conquer approach. *Information Processing & Management*, 47(6):808–824.

Okoli, C. and Schabram, K. (2010). A guide to conducting a systematic literature review of information systems research. *Sprouts Work. Pap. Inf. Syst*, 10(26).

Pack, D. and Weinstein, C. (2001). The use of dynamic segment scoring for language-independent question answering. In *Proceedings of the first international conference on Human language technology research*, pages 1–5. Association for Computational Linguistics.

PAULA FILHO, W. and PÁDUA, D. Engenharia de software; fundamentos, métodos e padrões, 2003. *Rio de Janeiro, Editora LTC.*

Pavlić, M., Han, Z. D., and Jakupović, A. (2015). Question answering with a conceptual framework for knowledge-based system development "node of knowledge". *Expert systems with applications*, 42(12):5264–5286.

Pechsiri, C. and Piriyakul, R. (2016). Developing a why–how question answering system on community web boards with a causality graph including procedural knowledge. *Information Processing in Agriculture*, 3(1):36–53.

Peng, X., Chen, Y., and Huang, Z. (2010). A chinese question answering system using web service on restricted domain. In *Artificial Intelligence and Computational Intelligence (AICI), 2010 International Conference on*, volume 1, pages 350–353. IEEE.

Perera, R. and Perera, U. (2012). Question answering through unsupervised knowledge acquisition. In *International Conference on Advances in ICT for Emerging Regions (ICTer2012).*

Pinto, D., Gómez-Adorno, H., Vilarino, D., and Singh, V. K. (2014). A graph-based multi-level linguistic representation for document understanding. *Pattern Recognition Letters*, 41:93–102.

Poppendieck, M. and Poppendieck, T. (2009). *Implementando o desenvolvimento Lean de Software: do conceito ao dinheiro.* Bookman Editora.

Prager, J. et al. (2007). Open-domain question–answering. *Foundations and Trends® in Information Retrieval*, 1(2):91–231.

Pressman, R. S. (1995). *Engenharia de software*, volume 6. Makron books Sao Paulo.

Prikladnicki, R., Willi, R., and Milani, F. (2014). *Métodos ágeis para desenvolvimento de software*. Bookman Editora.

Pudaruth, S., Boodhoo, K., and Goolbudun, L. (2016). An intelligent question answering system for ict. In *Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on*, pages 2895–2899. IEEE.

Pustejovsky, J., Knippen, R., Littman, J., and Saurí, R. (2005). Temporal and event information in natural language text. *Language resources and evaluation*, 39(2):123–164.

Rabiah, A. K., Sembok, T., and Halimah, B. (2008). Utilization of external knowledge to support answer extraction from restricted document using logical reasoning. In *Information Technology, 2008. ITSim 2008. International Symposium on*, volume 2, pages 1–8. IEEE.

Rabiah, A. K., Sembok, T. M. T., and Halimah, B. (2007). Towards skolemize clauses binding for reasoning in inference engine. In *Computational Science and its Applications, 2007. ICCSA 2007. International Conference on*, pages 273–282. IEEE.

Radev, D., Fan, W., Qi, H., Wu, H., and Grewal, A. (2005). Probabilistic question answering on the web. *Journal of the Association for Information Science and Technology*, 56(6):571–583.

Raghuvanshi, A. and Chase, P. Dynamic memory networks for question answering.

Ray, S. K., Singh, S., and Joshi, B. P. (2010). A semantic approach for question classification using wordnet and wikipedia. *Pattern Recognition Letters*, 31(13):1935–1943.

Richardson, K. D., Bobrow, D. G., Condoravdi, C., Waldinger, R., and Das, A. (2011). English access to structured data. In *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*, pages 13–20. IEEE.

Sagara, T. and Hagiwara, M. (2014). Natural language neural network and its application to question-answering system. *Neurocomputing*, 142:201–208.

Saquete, E., Munoz, R., and Martínez-Barco, P. (2006). Event ordering using terseo system. *Data & Knowledge Engineering*, 58(1):70–89.

Sbrocco, J. and Macedo, P. (2012). Metodologias ágeis-engenharia de software sob medida. *São Paulo: Editora Érica Ltda.*

Schach, S. R. (2009). *Engenharia de Software-: Os Paradigmas Clássico e Orientado a Objetos*. AMGH Editora.

Seena, I., Sini, G., and Binu, R. (2016). Malayalam question answering system. *Procedia Technology*, 24:1388–1392.

Shekarpour, S., Marx, E., Ngomo, A.-C. N., and Auer, S. (2015). Sina: Semantic interpretation of user queries for question answering on interlinked data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 30:39–51.

Silva, J., Coheur, L., Mendes, A. C., and Wichert, A. (2011). From symbolic to sub-symbolic information in question classification. *Artificial Intelligence Review*, 35(2):137–154.

Solorio, T., Pérez-Coutino, M., Montes-y Gémez, M., Villasenor-Pineda, L., and López-López, A. (2004). A language independent method for question classification. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1374. Association for Computational Linguistics.

Sommerville, I. Engenharia de software-8ª edição (2007). *Ed Person Education*.

Spranger, M. and Labudde, D. (2014). Establishing a question answering system for forensic texts. *Procedia-Social and Behavioral Sciences*, 147:197–205.

Stroh, E., Student, S., and Mathur, P. Question answering using deep learning.

Sucunuta, M. E. and Riofrio, G. E. (2010). Architecture of a question-answering system for a specific repository of documents. In *Software Technology and Engineering (ICSTE), 2010 2nd International Conference on*, volume 2, pages V2–12. IEEE.

Sun, M. and Chai, J. Y. (2007). Discourse processing for context question answering based on linguistic knowledge. *Knowledge-Based Systems*, 20(6):511–526.

Surdeanu, M., Ciaramita, M., and Zaragoza, H. (2011). Learning to rank answers to non-factoid questions from web collections. *Computational linguistics*, 37(2):351–383.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Tapeh, A. G. and Rahgozar, M. (2008). A knowledge-based question answering system for b2c ecommerce. *Knowledge-Based Systems*, 21(8):946–950.

Tartir, S., Arpinar, I. B., and McKnight, B. (2011). Semanticqa: exploiting semantic associations for cross-document question answering. In *Innovation in Information & Communication Technology (ISIICT), 2011 Fourth International Symposium on*, pages 1–6. IEEE.

Terol, R. M., Martínez-Barco, P., and Palomar, M. (2007). A knowledge based method for the medical question answering problem. *Computers in biology and medicine*, 37(10):1511–1521.

Toti, D. (2014). Aqueos: a system for question answering over semantic data. In *Intelligent Networking and Collaborative Systems (INCoS), 2014 International Conference on*, pages 716–719. IEEE.

Varathan, K. D., Sembok, T. M. T., and Kadir, R. A. (2010). Automatic lexicon generator for logic based question answering system. In *Computer Engineering and Applications (ICCEA), 2010 Second International Conference on*, volume 2, pages 349–353. IEEE.

Vazquez-Reyes, S. and Black, W. J. (2008). Evaluating causal questions for question answering. In *Computer Science, 2008. ENC'08. Mexican International Conference on*, pages 132–142. IEEE.

Vila, K., Maźon, J.-N., and Ferŕandez, A. (2011). Model-driven adaptation of question answering systems for ambient intelligence by integrating restricted-domain knowledge. *Procedia Computer Science*, 4:1650–1659.

Vinyals, O. and Le, Q. (2015). A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Voorhees, E. M. and Tice, D. M. (2000). Building a question answering test collection. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 200–207. ACM.

Walke, P. P. and Karale, S. (2013). Implementation approaches for various categories of question answering system. In *Information & Communication Technologies (ICT), 2013 IEEE Conference on*, pages 402–407. IEEE.

Wang, M. and Manning, C. D. (2010). Probabilistic tree-edit models with structured latent variables for textual entailment and question answering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1164–1172. Association for Computational Linguistics.

Wang, W., Auer, J., Parasuraman, R., Zubarev, I., Brandyberry, D., and Harper, M. (2000). A question answering system developed as a project in a natural language processing course. In *Proceedings of the 2000 ANLP/NAACL Workshop on Reading comprehension tests as evaluation for computer-based language understanding sytems-Volume 6*, pages 28–35. Association for Computational Linguistics.

Wazlawick, R. (2013). *Engenharia de software: conceitos e práticas*, volume 1. Elsevier Brasil.

Wen, D., Jiang, S., and He, Y. (2008). A question answering system based on verbnet frames. In *Natural Language Processing and Knowledge Engineering, 2008. NLP-KE'08. International Conference on*, pages 1–8. IEEE.

Wenyin, L., Hao, T., Chen, W., and Feng, M. (2009). A web-based platform for user-interactive question-answering. *World Wide Web*, 12(2):107–124.

Woods, W. A. (1997). Conceptual indexing: A better way to organize knowledge.

Xie, N. and Liu, W. (2005). An answer fusion model for web-based question answering. In *Semantics, Knowledge and Grid, 2005. SKG'05. First International Conference on*, pages 8–8. IEEE.

Xiong, C., Merity, S., and Socher, R. (2016). Dynamic memory networks for visual and textual question answering. In *International Conference on Machine Learning*, pages 2397–2406.

Yang, M.-C., Lee, D.-G., Park, S.-Y., and Rim, H.-C. (2015). Knowledge-based question answering using the semantic embedding space. *Expert Systems with Applications*, 42(23):9086–9104.

Yao, X. (2014). Feature-driven question answering with natural language alignment.

Yu, H., Lee, M., Kaufman, D., Ely, J., Osheroff, J. A., Hripcsak, G., and Cimino, J. (2007). Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians. *Journal of biomedical informatics*, 40(3):236–251.

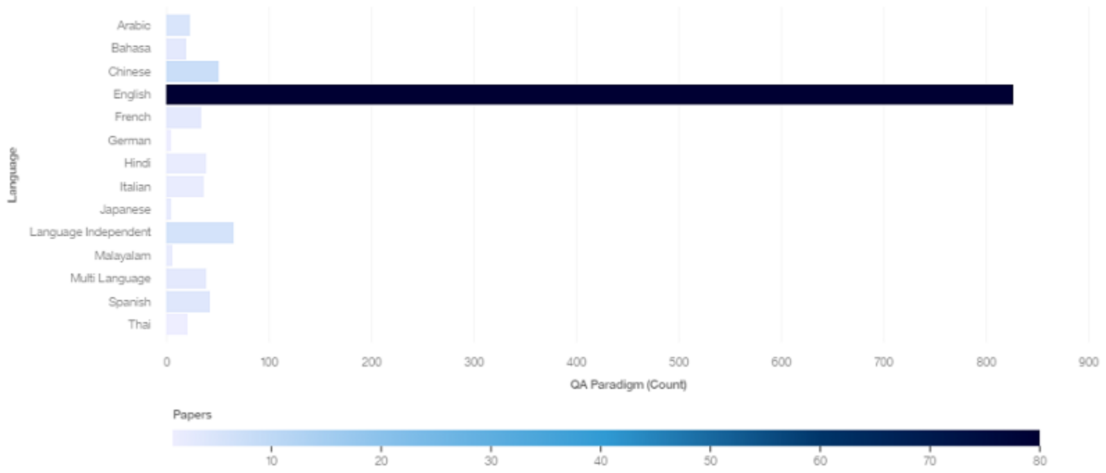
Zayaraz, G. et al. (2015). Concept relation extraction using naïve bayes classifier for ontology-based question answering systems. *Journal of King Saud University-Computer and Information Sciences*, 27(1):13–24.

Zhang, X., Hao, Y., Zhu, X.-Y., and Li, M. (2008). New information distance measure and its application in question answering system. *Journal of Computer Science and Technology*, 23(4):557–572.

Zong, H., Yu, Z., Guo, J., Xian, Y., and Li, J. (2011). An answer extraction method based on discourse structure and rank learning. In *Natural Language Processing andKnowledge Engineering (NLP-KE), 2011 7th International Conference on*, pages 132–139. IEEE.

Appendix

Figure 15 – Languages by Paradigm implementation.


Arabic
Bahasa
Chinese
English
French
German
Hindi
Italian
Japanese
Language Independent
Malayalam
Multi Language
Spanish
Thai
Language
QA Paradigm (Count)
Papers