Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Programa de Pós-Graduação em Ciência da Computação

Tiago Faceroli Duque

Graph Based Approach for Question Answering - Improving Efficiency in Natural Language Processing for Small Corpora

Juiz de Fora

2019

Tiago Faceroli Duque

Graph Based Approach for Question Answering - Improving Efficiency in Natural Language Processing for Small Corpora

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Juiz de Fora, na área de concentração em Ciência Exatas, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Advisor: Fernanda Cláudia Alves Campos

Co-Advisor: Wagner Antônio Arbex

Co-Advisor: Ely Edison da Silva Matos

Juiz de Fora

2019

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Duque, Tiago Faceroli.
Graph Based Approach for Question Answering - Improving Efficiency in Natural Language Processing for Small Corpora / Tiago Faceroli Duque. – 2019.
80 f. : il.

Advisor: Fernanda Cláudia Alves Campos
Co-Advisor: Wagner Antônio Arbex
Dissertação (Mestrado) – Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Programa de Pós-Graduação em Ciência da Computação, 2019.

1. Sistema de Perguntas e Respostas. 2. Processamento de Linguagem Natural. 3. Grafo de Conhecimento. 4. Gado de Leite I. Campos, Fernanda Cláudia Alves, orient. II. Arbex, Wagner Antônio, coorient. III. Matos, Ely Edison da Silva, coorient. IV. Graph Based Approach for Question Answering - Improving Efficiency in Natural Language Processing for Small Corpora.

Tiago Faceroli Duque

Graph Based Approach for Question Answering - Improving Efficiency in Natural Language Processing for Small Corpora

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Juiz de Fora, na área de concentração em Ciência Exatas, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Aprovada em: 30/07/2019

BANCA EXAMINADORA

Prof. Dr. Fernanda Cláudia Alves Campos - Orientador
Universidade Federal de Juiz de Fora

Prof. Dr. Wagner Antônio Arbex - Coorientador
Universidade Federal de Juiz de Fora

Dr. Ely Edison da Silva Matos - Coorientador
Universidade Federal de Juiz de Fora

Prof. Dr. Tiago Timponi Torrent
Universidade Federal de Juiz de Fora

Dr. Claudio Napolis Costa
EMBRAPA Gado de Leite

"You are worthy, Jehovah our God, to receive the glory and the honor and the power, because you created all things, and because of your will they came into existence and were created." (Revelation 4:11)

ACKNOWLEDGEMENTS

Most importantly, I must thank Jehovah God, the Creator of all things, including human brain. A masterpiece!

Also, I thank my loved wife, who served me as a backup for so long!

To my parents, I thank for so many careful counsels and cheering.

I thank my friends as well, who even while not knowing gave me strength to focus on what is more important.

It is also important to express gratitude to my teachers, who built me up by emptying their own knowledge up onto me. That includes Ely, the *guru* of this work, Wagner, its kickstarter, and Fernanda, who had a lot of courage to jump on when emergencies appeared!

Additionally, I thank IF Sudeste MG and my colleagues, whose institutional programs and personal support allowed for the development of this research while still working.

Gratitude also for the FrameNetBR team, who provided many of the underlying ideas.

Finally, a thanks to EMBRAPA Gado de Leite and many of its workers who gave the start and means to end this research.

M-Bot was silent. Then, in a smaller voice — somehow vulnerable — he asked, “Why do humans fear death?”

I frowned toward the console, where I knew the camera was. “Is that another attempt at humor?”

“No. I want to understand.” Skyward, Brandon Sanderson

ABSTRACT

This work presents the process of building a Question/Answer System (QAS) for Brazilian Portuguese based on a corpus extracted from a tech book on the dairy farming domain. Throughout the work, One common problem in Question/Answer Systems is identified and analyzed, namely, the cases where there's a distance between the terms present in the question and the terms present in the answer - this distance is called "question answering gap". To elaborate a Question/Answer System capable of overcoming such distance, a query expansion mechanism was developed through the use of a domain Knowledge Network, providing an effective way to give answers even though they do not share the same lexicon of the questions.

Keywords: Question Answering System. Natural Language Processing. Knowledge Graph. Dairy Farming.

RESUMO

Este trabalho apresenta o processo de construção de um Sistema de Perguntas e Respostas para o Português Brasileiro basedo em um corpus extraido de um livro técnico sobre o domínio de Gado Leiteiro. Ao longo deste trabalho, identifica-se e analisa-se um problema comum em sistemas de pergunta e resposta, onde ocorre uma distância entre os termos da pergunta e os termos da resposta - essa distância é nomeada "lacuna entre pergunta e resposta". Para elaborar um sistema de pergunta e resposta capaz de sobrepujar tal distância, um mecanismo de expansão de consultas foi desenvolvido através da confecção e uso de uma rede de conhecimento de domínio, fornecendo um meio efetivo de prover respostas mesmo que não compartilhem o mesmo léxico das perguntas.

Palavras-chave: Sistema de Perguntas e Respostas. Processamento de Língua Natural. Grafo de Conhecimento. Gado de Leite.

LIST OF FIGURES

Figure 1 – Question 411 and its correct answer. 16
Figure 2 – Methodological steps of this research. 18
Figure 3 – An excerpt of some concepts proposed in the Common Dairy Ontology (VERHOOSEL; SPEK, 2016). 30
Figure 4 – An Euler Diagram summarizing the discussions on the distinction between Ontologies, Knowledge Graphs and Knowledge Bases. 32
Figure 5 – The result of the reasoning activity in the LUDI Framework for "*livro pesado*" (heavy book). 34
Figure 6 – Histograms about keyword and size distribution. 7(a) presents the relation between intersecting keywords (appear both in question and answer) and total question keywords. 7(b) presents the number of intersecting keywords over total answer keywords. 7(c) presents the distribution of intersecting keywords and 7(d) presents the distribution of size relation between questions and answers in characters. 40
Figure 7 – An example of a simple "Match the Column" activity. Note that without previous knowledge of what Cow, Milk or Pasture actually are, no matching could be made aside from random guessing. 41
Figure 8 – The developed system architecture, with a focus on the modules. . . . 43
Figure 9 – Part of the graph developed in the first attempt to model domain knowledge. 45
Figure 10 – A simple depiction of the Tokenization process. 47
Figure 11 – Two figures representing the Part of Speech Tagging process. 48
Figure 12 – A simple depiction of the Lemmatizing process. 49
Figure 13 – Concepts retrieved using Wikipedia. Note that while not qualified, all the described relations are plausible. 51
Figure 14 – A visualization of the spreading process over Question 400. The order is Left-Right, Top-Down. Numbered dots are answer nodes. 60

LIST OF TABLES

Table 1 – Distribution of questions among Sections in the book used as corpus. . . 15
Table 2 – Criteria for QAS classification based on (MISHRA; JAIN, 2016), a similar table with more details can be found in (MISHRA; JAIN, 2016, p. 349,350). 22
Table 3 – State of The Art Results for QASs, including the best results for English, Brazilian Portuguese and European Portuguese. 28
Table 4 – Topic classification results using four Machine Learning Techniques one the questions. 38
Table 5 – Syntax patterns used to extract knowledge from an earlier corpus. . . . 45
Table 6 – Relation types according to the SIMPLE Ontology project. (LENCI et al., 2000) and applied to this research context. 53
Table 7 – Result of the questions selected for testing after using the best parameters found (2 depth, 0.3 decay and 0.05 Threshold). Questions with * are questions that would clearly fit the description of the "Question Answer Gap" - note that only one of them (408) no answer could be found even among the top scoring ones. 61
Table 8 – Results with some of the used parameters. Decay and Threshold parameters were fixed because they displayed better results at earlier testing and did not interfere with the depth chosen. Best results are marked in bold. 62

LIST OF ACRONYMS

NLP Natural Language Processing

NLU Natural Language Understanding

IR Information Retrieval

NER Named Entity Recognition

ML Machine Learning

BoF Bag-of-Features

BoW Bag-of-Words

API Aplication Programming Interface

QAS Question Answering System

TREC Text Retrieval Conference

ACL Association for Computational Linguistics

MAP Mean Average Precision

P@k Precision at k

ACL Association for Computer Linguistics

KG Knowledge Graph

CLEF Cross-Language Evaluation Forum

CONTENTS

1	**INTRODUCTION**	**12**
1.1	MOTIVATION	13
1.2	PROBLEM DEFINITION	14
1.3	OBJECTIVES	16
1.4	METHODOLOGY	17
1.5	RESEARCH OUTLINE	18
2	**DEFINITIONS AND STATE OF THE ART**	**20**
2.1	QUESTION ANSWERING	20
2.1.1	**Question Answering Systems State of the Art**	26
2.2	KNOWLEDGE REPRESENTATION MODELS	29
2.3	NATURAL LANGUAGE PROCESSING AND CONNECTIONISM	32
2.3.1	**Connectionism**	33
3	**A GRAPH BASED APPROACH FOR QUESTION ANSWERING**	**37**
3.1	INTRODUCTION	37
3.1.1	**An Evaluation of Traditional Methods**	37
3.1.2	**The Question Answering Gap**	39
3.2	A GRAPH BASED APPROACH FOR QUESTION ANSWERING: THE PROPOSAL	42
3.2.1	**Development steps**	42
3.2.2	**Approach Architecture**	42
3.3	APPROACH DEVELOPMENT	44
3.3.1	**Extracting Knowledge from Corpus**	44
3.3.2	**Representing Knowledge Through Graphs**	51
3.3.3	**Spreading Activation**	57
3.4	RESULTS AND COMPARISONS	59
3.4.1	**Restrictions**	62
4	**FINAL REMARKS AND FUTURE WORKS**	**64**
	REFERENCES	**67**
	APPENDIX A – Questions and Answers Used	**73**

1 INTRODUCTION

Having computers react to human language has been one of the core discussions in applied Computer Science since Alan Turing's days (TURING, 1959). Therefore, the activities of processing, understanding and producing human (or natural) language by computers are some of the most sought-after tasks in modern-day Computer Science. It is to such an extent that several fields of research have risen from it, such as Natural Language Processing (NLP), Speech Recognition, Natural Language Generation and, as an effort to deepen researches in human language and reasoning, Natural Language Understanding (NLU).

The past decades have experienced several outbreaks in NLP and Speech Recognition due to the development of Machine Learning (ML) techniques allied to more powerful and accessible computer hardware. Several NLP tasks, such as Dependency Parsing and Part of Speech (POS) Tagging are considered by some as solved in a few languages such as English (DOZAT; MANNING, 2017), while others, such as Named Entity Recognition (STRUBELL et al., 2017) and Machine Translation, have seen deep development in the last years[1]. These advancements caused many people to consider NLP a somewhat dealt with problem (MATOS, 2014, p. 19), since it is commonplace for anyone today to access indexed documents by a query passed to Google Search Engine[2] or to instantly translate text to over 100 distinct languages [3].

However there are still many challenges in the field of NLP and its derivatives. For example, disambiguation is still a problem, even with the application of novel approaches such as vectorization (MIKOLOV et al., 2013) and the use of Deep Learning techniques (YOUNG et al., 2018) (MATOS, 2014).

Related to the problem of disambiguation there are several other challenges, mainly linked to having the computer really understand, and not just estimate, what a natural language entry is really trying to "say". In this recently bustling field of research, called NLU, computational linguists and computer scientists are trying to propose techniques and models that not only treat natural language as a set of symbols that can be transformed into numbers and vectors, but rather a much more complex structure that needs a more specific approach to be treated (ABEND; RAPPOPORT, 2017).

Even with the mentioned problems, the application of NLP and its children techniques compose a very powerful and popular field that is being engulfed by enterprise

[1] The Association for Computer Linguistics (ACL) keeps an updated web page with current State of the art results for many NLP tasks - this can be viewed in `https://aclweb.org/aclwiki/State_of_the_art` - Last accessed May 27th, 2019.

[2] `www.google.com` - Last Accessed May 27th, 2019.

[3] For more information, visit `https://translate.google.com/intl/en/about/languages/` - Last accessed May 27th, 2019.

application and technologies. To name only a couple, there's been an outburst in the use of Chat-bots for easy and cheap communication with customers and a swarm of sentiment analysis techniques for customer complaint and market trends tracking (THANAKI, 2017, p. 8-10).

1.1 MOTIVATION

This work was motivated by an attempt of applying NLP techniques to automate the process of customer service trough a hybrid approach of Question/Answer (QA) and Information Retrieval (IR) for the dairy farming domain in the Brazilian Portuguese language.

The originally proposed solution involved the use of a large question/answer corpus on dairy farming accumulated over the years by Embrapa Gado de Leite, a Brazilian dairy farming research institute[4].

Due to some confidentiality, privacy and intellectual property issues faced during the corpus selection, in order to properly employ the time used in acquiring knowledge about dairy farming and NLP techniques, it was decided to remain with the original idea, but experimenting with other sources of information for the corpus buildup, such as public forums on dairy farming and the book "O Produtor Pergunta, a Embrapa Responde: Gado de Leite"[5], a free book edited by Embrapa with answers for the 500 most common dairy farming questions in Brazil (CRUZ et al., 2011).

After evaluating the available resources, it was opted to focus on using this book as the corpus for the project, in an opposite direction to the current trends. That because by current standards, a book can hardly be considered a corpus (THANAKI, 2017, p. 20). The chosen corpus has another specificity: every single one of its 500 questions are a closed context, referring to a single, specific answer - there are no questions whose answer is repeated. Therefore, this small corpus poses a challenge, since the most popular and advanced techniques today use large and clustered data as input.

In order to confront these problems, it was proposed to apply a set of techniques derived from connectionist models, such as a type of Knowledge Graph, one of many graph-based approaches. With this proposal, it was attempted to strike on another problem, which is represented by the seemingly nonexistent statistical relation between the wording of some questions and their correct answers.

A final challenge confronted during this research is related to the very weakness of Brazilian Portuguese NLP tools. Since the popularity of NLP in Brazil has risen in the last decade, only a few resources are available. And even though Portuguese is in most

[4] `https://www.embrapa.br/gado-de-leite` - Last accessed May 27th, 2019.
[5] The Producer Asks, Embrapa Answers: Dairy Cattle

open API's list of supported languages, all of them rely on a single annotated corpora aged almost 20 years and built on top of news texts (FREITAS et al., 2008), which causes an impact in the efficiency and reliability of the tools. These tools, even if considered state of the art in English, display several problems in some NLP pipeline stages. This last challenge has caused the need to use creativity to accomplish several steps of the preprocessing pipeline in NLP.

1.2 PROBLEM DEFINITION

Traditional NLP algorithms rely on large annotated corpora to solve problems such as classification and topic extraction. Derived tasks such as Question Answer and Chatbots are also data-heavy. This fact is related to the standard use of syntax based Supervised Machine Learning Techniques (CAMBRIA; WHITE, 2014). These techniques are bound to word frequencies organized in "bags of words"(BoW) or, more recently, Word Embeddings (MIKOLOV et al., 2013), which goes a step further, but works in a similar manner to the prior techniques.

In simple terms, to get acceptable results, most Machine Learning techniques require many iterations of input weight balancing (a process usually called *training*) to be done, usually comparing the obtained result to the expected result and attempting to correct wrong weightings until a certain threshold is met. In order to avoid the so called "overfitting", these iterations have to be done on top of diversified sets of previously annotated data, with (in the optimal scenario) an equally distributed amount of representatives for each class (in the case of a classifying task). In most Machine Learning techniques, the result is a statistical model (in others it is a deterministic auto-generated data structure like a graph or a tree, but the production process of this structure is similarly bound to statistical weighting) which, after receiving a given input, returns the most likely output.

As described in the Motivation section, the main data source for this research was chosen to be the book "O Produtor Pergunta, a Embrapa Responde: Gado de Leite"(CRUZ et al., 2011), part on a book series published free of cost by Embrapa. It is divided in 11 sections, each of which encompasses areas pertinent to the dairy cattle management context. Table 1 describes the sections of the book.

One approach that seems fit to the corpus is just using the available data to do topic classification in a statistical way. However, approaching the problem this way doesn't offer a question/answering solution, but rather a topic classification one. This type of approach is interesting and can be labor saving in an attempt to select the best answer for a question, but doesn't offer a solution to the proposed problem.

One interesting detail of this corpus is related to how the questions and answers are

Table 1 – Distribution of questions among Sections in the book used as corpus.

"500 Perguntas 500 Respostas - Gado de Leite- List of Sections		
Section Number	**Question Numbers**	**Section Name (self provided translation)**
1	1-40	"Cria e Recria de Bezerras e Novilhas"(Creation and Reproduction of Calfs and Steers)
2	41-134	"Alimentação e Manejo de Vacas e Touros"(Feeding and Management of Cows and Bulls)
3	135-289	"Recursos Forrageiros"(Forage Resources)
4	290-339	"Reprodução"(Reproduction)
5	340-379	"Melhoramento Genético Animal"(Animal Genetic Improvement)
6	380-414	"Saúde Animal"(Animal Health)
7	415-449	"Mastite e Qualidade do Leite"(Mastitis and Milk Quality)
8	450-457	"Produção Orgânica de Leite"(Organic Milk Production)
9	458-468	"Gerenciamento da Atividade Leiteira"(Milk Activity Management)
10	469-482	"Bem-estar Animal"(Animal Wellfare)
11	483-500	"Instalações, Ambiência e Manejo de Dejetos"(Facilities, Ambience and Waste Management)

Source: Created by author (2019).

linked. As a book made in the style of the famous "Dummy" franchise, the answers rarely use the same wording as the questions. Take, for instance, the question numbered 411, "O que é brucelose?" (What is brucellosis?). In its whole answer, the word that represents the question "core", "brucelose" (*brucellosis*), doesn't appear once. Therefore, many related questions, such as "How to know if my cows have brucellosis?" or "Is brucellosis a bad thing?", would not find a match in the correct answer using rule based and statistical based approaches because the main question feature is missing in the answer. Figure 1 displays the full question and its answer.[6]. This absence of words (addressed in this work as a "gap") is one of the problems that will be addressed in the development of this work.

Another fact about the corpus, is that there is a big difference in length between the number of words in the questions and the number of words in the answers. This is a phenomenon expected to happen in some contexts, since questions usually intend to obtain more information than is available. It was noted that as an average, the answers are about 8 times larger (in number of words) than the questions[7].

Therefore, while the question may contain one or two keywords, the answer not rarely contains more than a dozen.

How, then, to find a relation between the questions and the answers? How to deal with the diminutive corpus without promoting overfitting? Can these problems be tackled in a computational way? These are some of the questions that drive this research.

[6] Except when explicitly mentioned, all the translations from Brazilian Portuguese provided for this work were done by the author. Keep in mind that, in order to preserve some of the original arguments, the translation process was more focused on being literal than easy to read.

[7] This can be contrasted to the characteristic mentioned by CRISCUOLO (2017) regarding the originally proposed corpus, where the "Consumer Questions", as called by the author, tends to be accompanied by a story that helps defining the question scope. In that case, the difference in length is usually smaller.

Figure 1 – Question 411 and its correct answer.

411. O que é *brucelose*?
É uma doença infectocontagiosa, causada por bactéria do gênero Brucella e caracterizada por distúrbios de fertilidade nos machos e fêmeas. O diagnóstico deve ser feito por exame laboratorial específico, realizado pelo menos uma vez ao ano.
Para a prevenção, devem ser vacinadas e marcadas as bezerras, entre o 3º e o 8º mês de idade, com a vacina B-19. Deve-se adquirir somente animais com resultado negativo para o teste, mantê-los isolados em quarentena antes de sua incorporação ao rebanho, e realizar novo teste após 30 dias.
A ingestão de leite cru, proveniente de animal doente, e o contato com suas secreções corporais podem levar à instalação da doença no homem.

Translation:

411. What is *brucellosis*?
It is an infectious disease, caused by bacteria of the Brucella genus and characterized by fertility disturbs in male and female [animals]. The diagnosis can be done through a specific laboratory test, realized at least once every year.
For the prevention, the heifer must be vaccinated and marked, between the 3rd and 8th month of age, with B-19 vaccine. Only acquire animals with negative result in the test, keep them isolated in quarantine before gathering them with the flock, and then realize another test after 30 days.
The consumption of raw milk, obtained from sick animals, and the contact with their body secretions can install the disease on man.

Source: Created by author (2019).

1.3 OBJECTIVES

All of the previous mentioned questions are still a mater of profound research. Dealing with the "gap" between what is symbolized by text and what it really means is the central topic within Natural Language Understanding. Several lines of thought and proposals are being developed by computational linguists, psychologists and philosophers. One of these is called "Connectionism" and has a common origin with Artificial Neural Networks (WASKAN, b), the "biomimetic"[8] approach of the human Neuron for "computer reasoning".

According to the Connectionist approach, knowledge can be represented as a network, where the nodes are parts of this knowledge which, when interconnected, compose the "big picture". In a network, one can get from one node to its neighbors and recursively on. The same way, biology have accepted for a long time that the nervous system, composed by neurons, is a complex network where knowledge can be somewhat stored by

[8] Biomimetics is a term coined in the late 1950's by the biophysicist Otto Schmitt and relates to the "transfer of ideas and analogues from biology to technology" (VINCENT et al., 2006). In this case, the human neuron and its behavior has been copied by techniques such as the Artificial Neuron.

inter-neuron connections.

With this in mind, it can be proposed that an artificial network that represents the ideas (or concepts) and their interconnections could provide a "map" to the knowledge world. This approach could offer a solution for the problem mentioned in the previous section, where questions and answers have different wordings, but refer to the same ideas and concepts. Therefore, one intermediary objective of this work is to try to describe the building process of a "Knowledge Graph" (or network) about the dairy farming domain.

With the network built, the following and central objective is to attempt to built a proof of concept able to correctly answer questions posed in Natural Language using the network as a basis, with a special focus on those questions that does not have the same wording as the correct answer.

A final objective is to provide access to the corpus after it has been organized, since its original display format was a single document in the PDF format. The purpose is to allow further attempts of surpassing the mentioned challenges posed by the corpus characteristics. This could help further development of Brazilian Portuguese Natural Language Processing and Understanding.

1.4 METHODOLOGY

Evaluating the effectiveness of an Information Retrieval system usually relies on mathematical evaluation techniques like Precision and Recall. Question Answering, on the other hand, applies some other evaluation derived techniques, some of which will be presented in Chapter 3.

Yet, evaluating the effectiveness of this proposal goes a few steps further from just numerical values. Hence, to properly check if the objectives were met or accomplished, it is first necessary to establish some methodological steps.

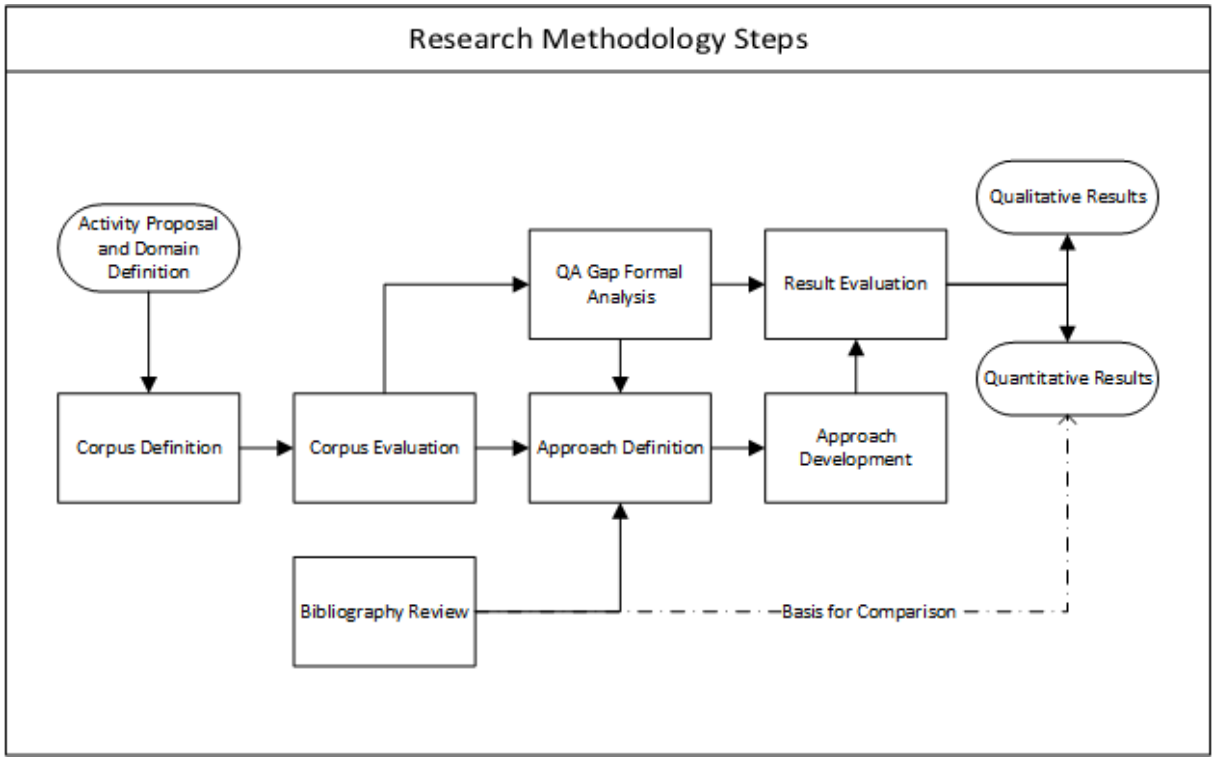
In order to understand the "question/answer gap", systematic, qualitative and quantitative studies were made using both statistics and evaluation of current research models. The results of this step defined which direction this research should take.

Next up, with the proper definition of the "question/answer gap", an extensive analysis was done to ensure that the selected model is able to provide a solution to the problem. This was done, in part, through the study of related works and bibliography.

The appropriateness of the model was quantitatively evaluated, trying to establish a direct relation between the methods and the numerical results obtained. This quantitative analysis is mainly discussed in Section 3.4.

Finally, knowing that this work is an exploratory research, the limitations of the developed model are discussed, in order to provide a baseline for future researches and

Figure 2 – Methodological steps of this research.



Source: Created by author (2019).

developments in the subject.

Figure 2 attempts to summarize the methodological steps to develop and evaluate the proposed research. Also, several other methodological appointments are spread around this study, each one related in the specific topic being discussed. The following outline section gives a roadmap of this research content.

1.5 RESEARCH OUTLINE

To better present the problem, the methodology adopted and the solution proposed, this work is divided as follows: first, to contextualize the field of study by providing classifications, state-of-the-art approaches and inspirations, Chapter 2 will present ongoing works and results achieved in related researches. These include Information Retrieval and Question/Answer researches, a Knowledge Networks and related models overview and some general ideas on NLP, NLU and Connectivism.

The development process of the proposed approach, including previous attempts, methodology and a deeper analysis of the "Question/Answer Gap" are presented in Chapter 3. This chapter discuss graph-based approaches and the algorithm used, Spreading Activation. It also presents the reasoning logic that provoked the approach definition, along with the research results, contributions and limitations.

Finally, chapter 4 will present this research remarks and developments, as well as point out future works and discuss improvements for the proposed approach.

2 DEFINITIONS AND STATE OF THE ART

None of the technologies or approaches used in this work are innovative *per se*. Question/Answering is a long debated topic, maybe as old as Natural Language Processing itself and have several ramifications from the original problem (KOLOMIYETS; MOENS, 2011)(MISHRA; JAIN, 2016).

Regarding Knowledge Network representations, Ontologies are promising structures that compose the core of Web 3.0 and its descendants, being an active field of research with many attempts to shape the web and intense discussions on how to structure data in a way that computers can "understand" it (GRUBER, 1993). Knowledge Graphs, used in the proposed solution, are a simplification of the more formal Ontologies, which compose the nucleus of today's most famous search engines (CIMIANO; PAULHEIM, 2016).

As many other fields of research, Dairy Farming is now filled with attempts of applying computerized approaches to solve its problems and provide better ways of doing things. There are some notable researches in Ontology Building for Dairy Farming and proposals to solve similar problems of the context's Natural Language Processing tasks (VERHOOSEL; SPEK, 2016).

Finally, while still shy, Connectionist model based approaches are getting ever more popular in computational linguistics, following failed attempts to solve several Natural Language related tasks with the traditional, statistical approach (SINHA, 2008).

The purpose of this chapter is to describe the current state-of-the-art proposals for each of these areas, as well as to point out some definitions and successful new attempts, trying to assemble each contribution, point of comparison and inspiration that each one brings to this work.

In the end, as will be noticed, while the approaches used by this work are not innovative in isolation, the approach proposed by this research does not find a match when considering all methodologies and techniques together, causing it to be an innovation as a connectionist model-based QAS, also being one among a few proposals for Brazilian Portuguese .

2.1 QUESTION ANSWERING

The problem tackled by this work fits in between the domain of Information Retrieval (IR) and Question Answering (QA). Relating to the distinction between the two domains, KOLOMIYETS; MOENS (2011) describe Question Answering as a "sophistication" of Information Retrieval, where more than just a document or list of documents, the user receives "specific pieces of information as an answer" (KOLOMIYETS; MOENS, 2011, p. 5412). On the other hand, the same author describes both techniques as supplying

information needs “expressed as natural language statements or questions”.

On a different approach, CRISCUOLO (2017) proposes that Question answering “is a Natural Language Processing subject [...] [with similarities to] Information Retrieval” but distinguished by two specific aspects: (1) “receives as entry text segments” and (2) “presents as a result only one answer, instead of a document collection, as occurs in IR” (CRISCUOLO, 2017, p. 35) (the markings are ours).

This work does not propose, as stated by KOLOMIYETS; MOENS (2011), a technique where specific pieces of information are delivered to the user in a compositional manner. But rather, it uses CRISCUOLO (2017) definition as a means of classification. Since the corpus used is already composed by a set of questions and the strict, specialized answer to them, the final step mentioned by KOLOMIYETS; MOENS (2011) was considered as dismissive.

Regarding Question Answer Systems, there is already some consistent bibliography and researches on the matter. Surveys realized by KOLOMIYETS; MOENS (2011) and MISHRA; JAIN (2016) point out to an increase in researches, as well to the diversification of techniques applied. Also, BOUZIANE et al. (2015) makes a summary of current researches, focusing on their results and precision percentiles.

Related to Information Retrieval state of the art approaches, the *Text REtreival Conference* (TREC) is one of the best sources. Held annually by USA’s National Institute of Standards and Technology (NIST)[1], this competition is the source for many challenges and breakthroughs in Information Retrieval and, for a time, Question Answering[2]. Also, many papers are published anualy as part of the Association for Computational Linguists (ACL) conferences and journals[3].

Portuguese Language Question Answer System (QAS) literature, however, is quite limited. When we take into account Brazilian Portuguese variation only, this number falls even more. Two notable researches/systems in the area are COMUNICA, by WILKENS et al. (2010) and PRIBERAM by AMARAL; FIGUEIRA (2006), with the last, for European Portuguese, being the highest scoring QAS so far for any variation of the Portuguese language.

Down to definitions, KOLOMIYETS; MOENS (2011) proposes several important concepts in evaluating QAS’s. First of all, QAS’s are driven by *Questions*, which are “a natural language sentence, which usually starts with an interrogative word and expresses

[1] `https://trec.nist.gov/` - Last accessed May 27th, 2019.

[2] Since 2016 no *track* was provided for Question Answer - tracks are sets of challenges for specific problems, with an aim to collect as many distinct solutions as possible toward a previously specified issue. The last Question Answering track was summarized by AGICHTEIN et al. (2016) and had 14 participants total.

[3] `https://www.aclweb.org/portal/` - Last accessed May 27th, 2019.

some information need of the user". According to the authors, when classifying QAS's, two aspects must be taken into account: the *Information Sources*, which could be any type of structured or unstructured data collections (including not only texts, but also video, audio and other types of data); and the *Retrieval Model*, or the way which the QAS queries its Information Sources to obtain the answer (KOLOMIYETS; MOENS, 2011, p. 5414).

A set of distinct type of questions is also defined by KOLOMIYETS; MOENS (2011) to help classifying each System answering capabilities, namely: *factoids, list, definition, hypothetical, causal, relationship, procedural, and confirmation questions* (KOLOMIYETS; MOENS, 2011, p. 5414).

MISHRA; JAIN (2016) proposes a more detailed set of aspects, naming other important criteria for QAS classification. Eight aspects are proposed and are presented in Table 2.

Table 2 – Criteria for QAS classification based on (MISHRA; JAIN, 2016), a similar table with more details can be found in (MISHRA; JAIN, 2016, p. 349,350).

Criteria No.	**Criteria**	**Brief Explanation**	**Types**
1	Application Domain	The domain to which the QAS is developed	Restricted or Open Domain
2	Types of Questions	The Type of the information requested. When compared to (KOLOMIYETS; MOENS, 2011), the set of possible types is less granular.	Factoid, list, hypothetical, confirmation and causal questions
3	Types of analysis done to input text and documents	This criteria could be compared to pipeline stages in QAS's. They are not mutually exclusive, but additive in a way that a System can encompass one or more of them.	Morphological, Syntatical, Semantic, Pragmatic/Discourse, Expected Answer Type and Focus on Recognition of Questions

Continued on next page.

Source: Created by author (2019).

Table 2 – Criteria for QAS classification based on (MISHRA; JAIN, 2016), a similar table with more details can be found in (MISHRA; JAIN, 2016, p. 349,350).

Criteria No.	**Criteria**	**Brief Explanation**	**Types**
4	Type of Consulted Data	The type of data, which can also be cumulative.	Structured, Unstructured, Semi-Structured and Semantic Web
5	Data Source Characteristics	This criteria is not defined by a deterministic set of possibilities, but rather a continuous range, which makes up for many distinct possibilities. Perhaps this criterion includes the biggest variations among QASs today.	Size, Language, Heterogeneity, Genre, Media
6	Types of Representation and Matching Function	Criteria related to the functions used to rank or retrieve correct answers from dataset. Does not have to do with answer preparation.	Set Theoretic, Algebraic, Probability, Feature based and Conceptual Graph models

Continued on next page.

Source: Created by author (2019).

Table 2 – Criteria for QAS classification based on (MISHRA; JAIN, 2016), a similar table with more details can be found in (MISHRA; JAIN, 2016, p. 349,350).

Criteria No.	**Criteria**	**Brief Explanation**	**Types**
7	Types of Techniques used in QASs	The core techniques used for searching, matching and returning answers. It takes in account the whole picture, but the authors of the survey proposes that some of the other criteria match specifically with each of this criteria types. A comprehensive table with matching proposals can be found in (MISHRA; JAIN, 2016, p. 357)	Data Mining, Information Retrieval, NLU and Knowledge Retrieval Techniques
8	Form of Generated Answers	This criteria takes into account the way that the answer is provided to the user. The options are related to Text Retrieval and Natural Language Generation techniques.	Extracted Answer and Generated Answer

Source: Created by author (2019).

With either of the classification criteria, it is easy to point out that while many QASs have been developed since the earliest days of computer applications development, QAS specificity makes it so that hardly two systems share the exactly same traits. Also,

in the way that QASs are being developed today, hardly one system can be easily adapted to answer the questions proposed in a different research.

Measuring the efficiency of a QAS, however, depends on the type of question asked and how the answer is provided to the user. If the system is expected to behave similarly to an IR system where all information is to be extracted and presented as a list where document/answer order is not the most important factor, Precision, Recall and F-Score are probably the best bet to evaluate it (MANNING et al., 2008).

However, taking in account that the objective of the most QAS's is to provide the best, direct and most concise answers to a question and not a random list of related answers, other means of evaluation are needed. Some of them are presented in MANNING et al. (2008). Two measuring approaches that are going to be relevant to this study are *Mean Average Precision* (MAP) and *Precision at k* (P@k).

MAP is an evaluation technique that is popular among TREC community and is based on the idea of Average Precision. Average Precision, on itself can be defined as "the average of the precision value obtained for the set of top k documents existing after each relevant document is retrieved, and this value is then averaged over information needs"(MANNING et al., 2008, p. 159, 160). Considering a set of correct answers for a question $q_j \in Q$ defined as $\{a_1, ...a_m\}$ and a set of retrieved answers R_jk, the mathematical representation of MAP is presented in equation 2.1.

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} Precision(R_{jk}) \tag{2.1}$$

Where Precision is defined according to equation2.2.

$$Precision = \frac{\#(correct\ answers\ retrieved)}{\#(total\ answers\ retrieved)} \tag{2.2}$$

If the interest is in a strict 1:1 match between a question and an answer, in other words, $|Q| = 1$ and $m_j = 1$, the formulae could be summarized as in equation 2.3.

$$MAP(Q) = Precision(R_{jk}) \tag{2.3}$$

Or, the same as the Precision of the system (the number of hits over the overall response).

As another possibility, the *Precision@k* (P@K) evaluation computes the precision value at the first k results (CRISCUOLO, 2017, p. 55), as in the equation 2.4.

$$P@k = \frac{\#(correctanswersinfirstkvalues}{k} \tag{2.4}$$

This measurement could be averaged for all the questions as in equation 2.5.

$$P@K_{mean} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} P@K(q_i) \tag{2.5}$$

As mentioned by CRISCUOLO (2017), one interesting particularity happens to P@k when taking k=1, which could be indicated as P@1, considering only the correct responses at the first position, or strict answers (CRISCUOLO, 2017, p. 55).

2.1.1 Question Answering Systems State of the Art

The current State of the Art for QAS according to the ACL is the one described in MADABUSHI et al. (2018)[4]. This QAS, tested over TRECQA (the dataset prepared by TREC for testing QASs) scores a MAP of 86.2%. The used dataset and the system are of English language, which stacks up the better NLP modules so far. The authors point out that the increase in efficiency from other attempts occurs due to the use of enhanced Question Classification techniques (MADABUSHI et al., 2018, p. 3285), where using a wide set of predefined models for distinguishing question types allow for the addition of features that help up in the distinction between candidate answers.

The work in (MADABUSHI et al., 2018) also relies on the use of Wikipedia to help with Named Entity Recognition through a process called Wikification (MADABUSHI et al., 2018, p. 3287), where detected entities in a question are confirmed by the use of Wikipedia titles. Interestingly, while being based on a Convolutional Neural Network that uses a Bag Of Words as input, the system developed relies heavily in hand modeling to extract features that are essential to reach the achieved results, showing that these types of techniques are getting ever more attention nowadays due to their clearly visible benefits.

Another relevant research is Priberam, which is presented in (AMARAL; FIGUEIRA, 2006). It was, for a good while, the best ranking QAS for any Portuguese Language. While Priberam was created for European Portuguese, a shallow comparison of NLP in this language can be made with Brazilian Portuguese, since they share several common structural characteristics, vocabulary and commonly defined rules.

Priberam was built on top of a vast Lexicon obtained from years of work with a Portuguese Dictionary. With this lexicon as a base, the researchers behind *Priberam* built a system scoring a MAP of 0.645 on top of the Cross-Language Evaluation Forum (CLEF) Question Answer Track (QA@CLEF)[5]. *Priberam* system is based on five major tasks, which are: (1) indexing, (2) question analysis, (3) document retrieval, (4) sentence

[4] For the updated list of State of the Art for QAS, a wiki page is kept at `https://aclweb.org/aclwiki/Question_Answering_(State_of_the_art)` - The current presented results were obtained May 28th, 2019.

[5] More information can be found in `http://www.clef-initiative.eu/track/qaclef`.

extraction, and (5) answer selection (AMARAL; FIGUEIRA, 2006, p. 414). As seen, Priberam shares a lot of common traits with IR systems.

One interesting fact about *Priberam* is that it also relies heavily on question categorization to achieve its results (AMARAL; FIGUEIRA, 2006, p. 412), showing that this is a proven tendency in QASs and an important element to take into account to improve any work. Besides, one interesting step taken by the system is expanding the question queries through the means of a thesaurus of synonyms - this idea is similar to what was implemented in this work, but only considering the lexical relations between words.

The research more closely related to this work was done by CRISCUOLO (2017), which proposed *Slimrank*, a domain specific QAS for Brazilian Portuguese focused on dairy farming. In fact, (CRISCUOLO, 2017) research was done on top of the originally targeted corpus and, together with relevant concepts provided for the field, can serve as a baseline for result comparison.

Slimrank is based on a Convolutional Neural Network (CNN) for feature extraction and model classification. It uses Word Embeddings as feature types and represents text using textual graphs (CRISCUOLO, 2017, p. 82) through techniques such as TextRank and LexRank. However, the use of graphs was solely devised as feature representation for the used model, since some CNN kernels are able to process graph inputs - different from this work proposal, the used graphs are not a knowledge network structure, but rather a statistical structure recovered from the analysis of the input text.

Using *MilkQA*, the dataset developed from the corpus, *SlimRank* scored a MAP of 0.702 using Word Embeddigs generated from random content and a CNN-LDC model. In P@1 measures, the best results lies at 0.58 for the same configuration.

As a basis for further comparison with the current work, the criteria provided by MISHRA; JAIN (2016) can be used to characterize *SlimRank* as follows:

1. Application Domain: restricted domain - Dairy Farming;
2. Types of Question: most questions could be tagged Factoid or Confirmation[6];
3. Types of Analysis: the analysis is mainly morphological, with a bit of inferred semantic analysis;

[6] It is important to point out that CRISCUOLO (2017) defines a new type of questions in his research, namely, "Customer Questions", which could be characterized as a Factoid or Confirmation questions followed by stories and information. However, to simplify comparison, it is going to be focused only on the core of the questions, while not ignoring the fact that the context provided in such questions can offer hints that can help matching the correct answer.

Table 3 – State of The Art Results for QASs, including the best results for English, Brazilian Portuguese and European Portuguese.

Model Name	**Reference**	**MAP / Precision**	**P@1**	**Language**	**Year**
Question Classification + Pairwise-Rank + Multi-Perspective CNN	(MADABUSHI et al., 2018)	0.862	N/A	English	2018
SlimRank	(CRISCUOLO, 2017)	0.702	0.580	Brazilian Portuguese	2017
Priberam	(AMARAL; FIGUEIRA, 2006)	0.645	N/A	European Portuguese	2006

Source: Created by author (2019).

4. Type of Consulted Data: the data consulted is unstructured, since the corpus is textual;
5. Data Source Characteristics:
   a) Size: The question sizes are above the average.
   b) Language: Brazilian Portuguese.
   c) Heterogeneity: Not heterogeneous, the data is all in a single location (the set of questions received by Embrapa and the specialist provided answers).
   d) Genre: Informal, written in a descriptive manner.
   e) Media: Text media extracted from customer service e-mails.
6. Types of Representation and Matching Function: The matching functions are feature based.
7. Types of Techniques used: The technique used is that of Information Retrieval.
8. Generated Answer Forms: The answers provided are neither extracted or generated, but rather a full match (pointing towards Information Retrieval as classified by some authors).

With the evaluation methods described in the presentation of this section, a summary of these three related proposals can be presented. Table 3 shows the three models with their respective results for MAP or P@1 (if that has been calculated).

As can be perceived in Table 3 and presentations, while reasonable results can be found for English, Portuguese language QASs are still incipient, proving that further research is needed considering the language specificity.

As stated by MISHRA; JAIN (2016), Question Answering Systems can be classified according to the Types of Representation used to retrieve the correct answers. These refer to how the knowledge is represented and matched in the system. In order to provide a basis for the approach presentation, the next section discuss about ways to represent knowledge as network models. Why these have been adopted in the proposed approach is presented during the discussions in Chapter 3.

2.2 KNOWLEDGE REPRESENTATION MODELS

Today's most well known network model for representing knowledge are Ontologies. For this reason, no other knowledge network model can be explained without referencing them. While this work does not intend to develop a full fledged Ontology, it was the basis for the research and the modeled structure, another reason why this theme can't be ignored.

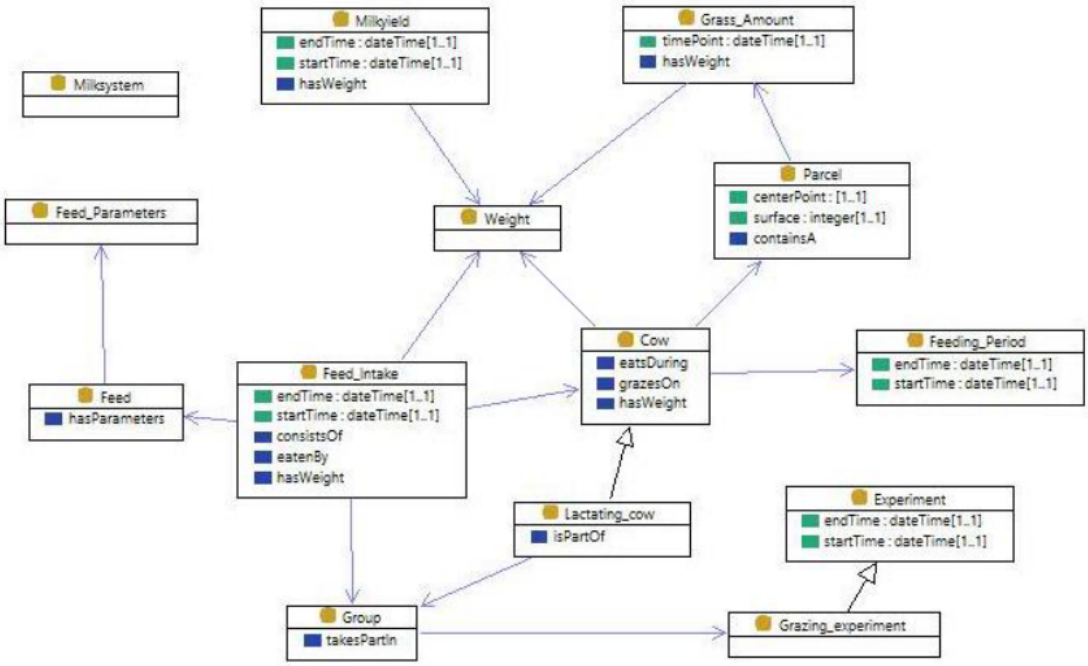
In Computer Science, Ontologies could be classified as a "formal and explicit specification of a domain terminology and their relations to a shared conceptualization" ((GRUBER, 1993) apud (MATOS, 2008, p. 23)). In other words, Ontologies are a formalization of domain knowledge and concepts to store and retrieve knowledge in a structured manner.

As seen in the previous chapter, Ontologies can be considered *Structured Data Sources* and are the focus of many of today's researches in the area of Artificial Intelligence. They are the core of the so-called Web 3.0, or *Semantic Web*, as defined by BERNERS-LEE et al. (2001) in the famous Scientific American article that pointed to the aurora of a new way of navigating the internet.

Empirically, Ontologies can be represented in the way of a Graph Structure, mainly composed by three elements (MATOS, 2008, p. 23-24):

- Classes: the elements that represent domain concepts - these can be organized in a hierarchical or taxonomic manner (In a Graph, these would be nodes).
- Relations: the associations between classes (in a Graph, these would be edges).
- Instances: individual *materialization* of the Classes (in a Graph, these would be

Figure 3 – An excerpt of some concepts proposed in the Common Dairy Ontology (VERHOOSEL; SPEK, 2016).



Source: (VERHOOSEL; SPEK, 2016, p. 6)

Nodes that pointed to Class Nodes, representing the general concepts that the instances materialize).

Ontologies are not solely related to Web, Information Retrieval or Natural Language Understanding tasks, but are rather deeper structures that are being used for intelligent decisions and data reasoning. For example, MAGALDI et al. (2018) described an Ontology in Dairy Farming domain used for the integration of heterogeneous databases.

Ontologies, however, rarely encompasses a domain entirely. The previously mentioned ontology on Dairy Farming only takes into account a subset of Animal Nutrition concepts and their relations, as a mean to tackle a database integration problem for animal feeding. It is not unusual to have several small Ontologies for a domain, each one designed for a specific and granular purpose.

Still in the Dairy Farming domain, VERHOOSEL; SPEK (2016) propose the creation of the *Common Dairy Ontology* (CDO), a more generalistic Ontology to channel Big Data flows into a single framework. This is probably the further developed, most generalistic Ontology on dairy farming so far, but since its creation has been sponsored by Dutch dairy industry, it is not yet available publicly. Figure 3 is a representation of a part of CDO, as proposed in (VERHOOSEL; SPEK, 2016).

Things get fuzzy when trying to distinguish Ontologies and other related technologies, such as Knowledge Graphs (KG) and Knowledge Bases (KB). EHRLINGER; WÖSS (2016) discuss on the matter, showing that these terms have been used indiscriminately in the last years. However, the authors bring an interesting definition to mind, pointing out that "An ontology is as a formal, explicit specification of a shared conceptualization that is characterized by high semantic expressiveness required for increased complexity" and that "an ontology does not differ from a knowledge base" (EHRLINGER; WÖSS, 2016, p. 3).

However, since Google's original publication on its blog[7] about the new underlying technology supporting its engine, it is very hard to define a single taxonomy for knowledge oriented, network-like systems. EHRLINGER; WÖSS (2016) try to advocate on the need for a distinction by pointing to two specific characteristics that can define a KG in distinction to Ontologies: size and extended requirements (EHRLINGER; WÖSS, 2016, p. 3). While acknowledging that these two aspects boundaries are yet to be defined, no further development on the matter has been proposed so far.

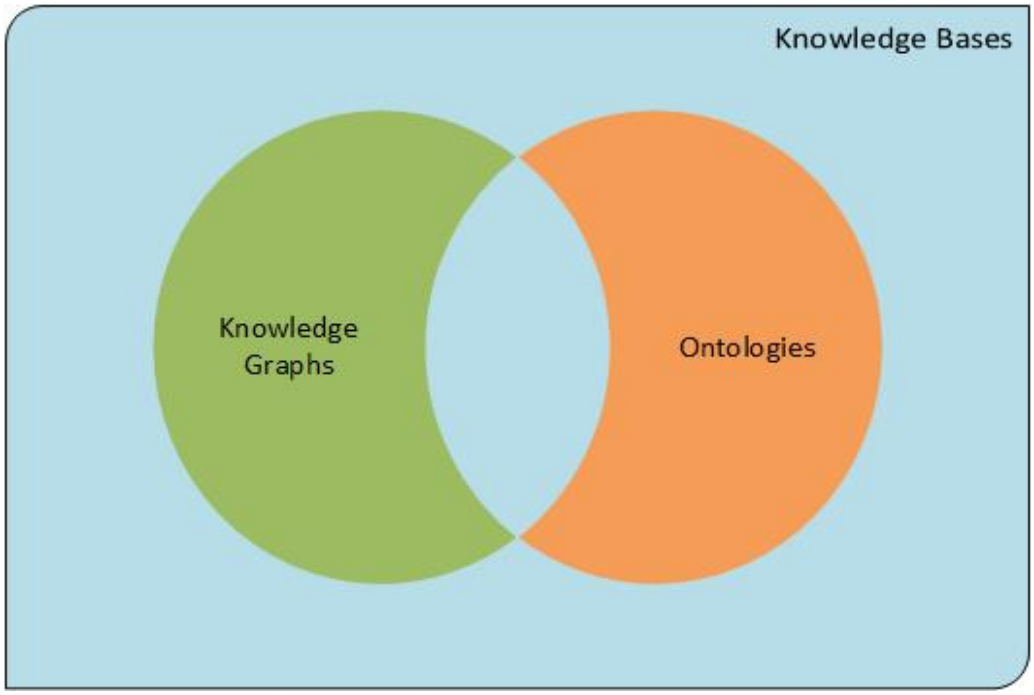
CIMIANO; PAULHEIM (2016) made a study on the recent development of self-titled Knowledge Graphs, pointing out that, while Google has a position on the term spotlight, it is not the only such network out there. The authors once again point to the nonexistence of a common definition for Knowledge Graphs. Instead, they point out to 4 characteristics that compose the minimum set of attributes that a KG must have to be considered such: (1) describes real world entities and their relations in the form of a graph; (2) defines the possible classes and relations of entities in a schema; (3) allows for arbitrary interrelation between entities; (4) covers various topical domains (CIMIANO; PAULHEIM, 2016, p. 2, 3). The authors also point out that, taking the first two characteristics only, an Ontology without instances could be considered a Knowledge Graph, thus making a set relation of "intersection" between the two models (or even a "belongs to" relation).

Knowledge Bases, on the other hand, are more generically referred. However, its definition is probably one of the oldest when related to Computer Science and can be found referring back to 1989, when JARKE et al. (1989) made the following definition: "[a knowledge base is] A representation of heuristic and factual information, often in the form of facts, assertions and deduction rules"(JARKE et al., 1989, p. 384)(the markings are ours). Using a more colloquial definition, Knowledge Bases would be any collection of structured or unstructured information that can be used by a computer system. Therefore, a Knowledge Base could be considered a superset of both Ontologies and Knowledge Graphs. Figure 4 attempts to summarize these distinctions

As will be presented further down, the proposed methodology to address the problems that are the focus of this work relies on a structure where knowledge is stored

[7] The blog entry can be read at: `https://www.blog.google/products/search/introducing-knowledge-graph-things-not/` - Last Access May 28th, 2019.

Figure 4 – An Euler Diagram summarizing the discussions on the distinction between Ontologies, Knowledge Graphs and Knowledge Bases.



Source: Created by author (2019).

and accessed in a connected way similar to a Graph. It is advocated that the structure used can be considered a Knowledge Graph as proposed by CIMIANO; PAULHEIM (2016), since it describes real world entities and their relations, defines classes and relations in a schema-like structure, allows for arbitrary relations between entities and, while it's focused on the Dairy Farming domain, this domain is much wider than just a similar set of ideas and labels, therefore being considerable generalist. However, since it lacks formalization and instantiation capabilities, which are proposed to be core characteristics of Ontologies, it cannot be defined as such.

In order to translate between the Knowledge Representation Model and an Entry in Natural Language, it is necessary to apply techniques held in the scope of Natural Language Processing and Natural Language Understanding. The next section presents some discussions to give a basis for the techniques adopted in the proposed approach.

2.3 NATURAL LANGUAGE PROCESSING AND CONNECTIONISM

Perhaps one of the best introductions to Natural Language Processing and how it developed in the course of the last seventy years is the one presented in (JURAFSKY; MARTIN, 2014). As pointed out by the authors, the first ideas towards having a machine process and understand natural language were born together with the notion of "Artificial

Intelligence" as classically proposed by Alan Turing. In the famous paper "Computing Machinery and Intelligence", the author proposes as one of the first attempts of AI that a machine could be taught "to understand and speak English" through the means of the "best sense organs that money can buy, and then teach[ing]" (the markings are provided by the author)(TURING, 1959, p. 460).

JURAFSKY; MARTIN (2014) then describe six generations of Natural Language Processing development, starting from Chomsky's formal language theories, passing through the rise of Empiricism and finishing the narration with the establishment of Machine Learning and probabilistic theories at the core of NLP from 2007 inwards (JURAFSKY; MARTIN, 2014, p. 9-13). It is around the 1970's that, according to the authors, Natural Language Understanding (NLU) steps out as a field. Referring to a series of works done in the period, the authors relate NLU as an area "focused on human conceptual knowledge such as scripts, plans and goals, and human memory organization" (JURAFSKY; MARTIN, 2014, p. 11).

This history is supported by the review done in 1994 by JONES (1994), who described the development of NLP from 1940's to the date of the article publication. There, the same distinction between a descriptional approach and a probabilistic one is posed, with the prominence of the last being established in the last period. In this work, however, there's a reference to a rise of connectionist approaches during the 1980's related to probabilistic networks, establishing a link between descriptional methods and probabilistic methods (JONES, 1994, p. 10).

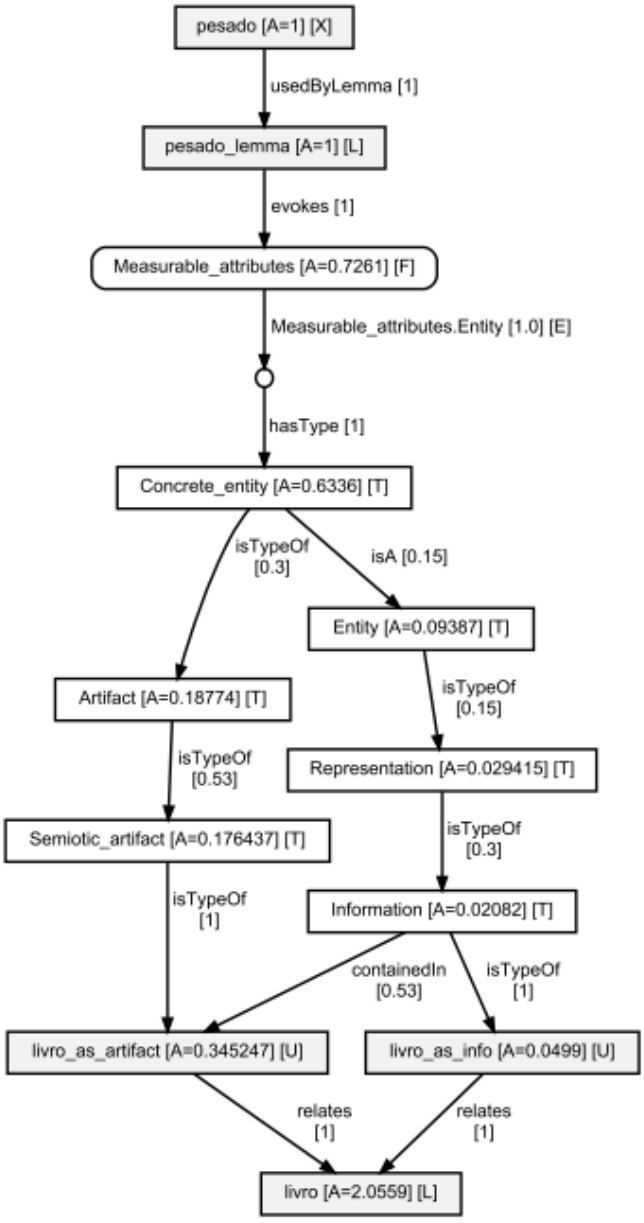
While both probabilistic and descriptional methods are distinct regarding the way the model is structured, both can be represented in a connected way such as an Ontology or an Artificial Neural Network. These are structured on top of a set of representation approaches called Connectionism. This is discussed in the following subsection.

2.3.1 **Connectionism**

The Connectionist approaches are based on the ideas proposed by MCCULLOCH; PITTS (1943) in 1943. In the famous paper which gave the basis to Artificial Neural Networks research, the authors analyze the neural activity in a formalized and mathematical way, creating the bases of the Artificial Neuron. Following Neurological researches, the work describes the neural activity as a network composed of units (the neurons) whose connections are built during the process of learning, inherent to the cognitive process (MCCULLOCH; PITTS, 1943, p. 119).

The ideas of MCCULLOCH; PITTS (1943) were then developed into what is now called Connectionism. In this approach to the study of human cognition, mental process related tasks are disposed in network-like structures made of interconnected processing units. Similar to what was discovered to happen to neurons, in connectionist approaches,

Figure 5 – The result of the reasoning activity in the LUDI Framework for "*livro pesado*" (heavy book).



Source: (MATOS, 2014, p. 113).

the connection between processing units have a "threshold" value which is regulated by an "activation function", causing interaction between units only to happen when a certain amount of energy is given to the connection.

While Connectionism is nowadays more commonly related to Neural Networks and statistical machine learning, earlier applications of the theory were based on hand

tailored models that attempted to capture representations of the world in a connectionist fashion. Mainly regarded as wasteful due to the amount of manual labor needed to build the models, these approaches have been nearly forgotten until recent researches in NLU pointed to the need of a logical formalization in order to enable computers to simulate human thought.

Under this setting, a few researches come to mind when relating connectionist model based approaches to NLP and NLU activities. One, directly related to this work proposal and reasoning is described in (MATOS, 2014). Following the researches on FrameNet, especially the Brazilian Portuguese FrameNet, the work proposes a Framework based on a connectionist network like structure to do disambiguation tasks. Shortly, this Framework uses Spread Activation over a Composite Ontology to try to obtain the most probable meaning of an ambiguous word. A lot of expertise obtained in this framework was used to orient the current work. Picture 5[8] shows the product of the Framework after being activated by one test input - note how the reasoning process occurs through connections between elements of the system.

Another connectionist research based on a network like structure is presented in (SINHA, 2008). In this work, the author address the problem of answering questions about complex events, following a non conventional path in QA research. To do it, the author uses Ontologies in the form of the so-called X-nets, a type of network capable of dynamically modelling events through the use of two main types of nodes: “Places”, which holds resources and knowledge; and “Transitions”, which can actively create, destroy and test resources stored in “Places” (SINHA, 2008, p. 28). Although this is a QA research, it is limited to addressing one type of complex questions, which are the Event questions, not applying the proposed solution to more common types of questions, such as factoid and list questions.

The work presented in (OU et al., 2008) also proposes an Ontology based QA system, this time for questions about the touristic domain. In this work, the developed Ontology serves the purpose of addressing a set of entities that can be referenced by a class described in Natural Language. The input is then processed in the NLP pipeline fashion and then parsed using WordNet, finally using the elucidated words to retrieve information from the underlying Ontology.

While not particularly a QAS, another relevant Natural Language, network based connectionist research is the one developed by FrameNet Brasil[9] for 2014’s World Cup. As a result of this research, a trilingual dictionary was developed using FrameNet as an

[8] There are also other metaphorical interpretations for “heavy book” such as “dense book” and “book full of violence”, however, in the depicted analysis these were not retrieved in the disambiguation process.

[9] `https://www.ufjf.br/framenetbr/publicacoes/` - Last Accessed May 29th, 2019.

underlying reasoning mechanism, allowing for interconnection of related terms (called lexical units in FrameNet)[10].

While several other NLP/NLU works could be found under the connectionist umbrella, none of them share the same domain, corpus characteristics and purpose of use of connectionist structures of this work, making it unique among current researches.

[10] The resulting product can be accessed in `http://www.dicionariodacopa.com.br/` - Last Accessed May 29th, 2019.

3 A GRAPH BASED APPROACH FOR QUESTION ANSWERING

3.1 INTRODUCTION

As stated on the previous chapters, current NLP techniques and researches rely heavily on statistical machine learning approaches to solve problems. As pointed out, these techniques usually require large amount of data, either previously treated by annotation or using unsupervised machine learning methods to acquire statistically recoverable features. QA and IR are not different in this matter, since most systems are fueled by feedback stacks or structured on gold-standards, which provide the main features and their distributions among classification individuals (being it document classes or questions).

The data used for the development of this research is not massive, instead, there's a single answer as a gold-standard for each question. Developing a system that can answer each of the questions proposed by the book is a matter of matching every question (and possibly newly formed, similarly built questions) to the exact answer provided by the specialists who wrote the book. This amount of information for each class would hardly enable any statistical technique to weight properly the features. This problem is reported and explained in the following section.

3.1.1 **An Evaluation of Traditional Methods**

In a statistical approach, the whole corpus would be considered, then all words would be extracted, weighted and, as is usually done, the most informative ones would be selected through any of many techniques (such as TF-IDF and Page-Rank)(MANNING et al., 2008, p.109-133). These selected words would compose the features for the multiclass classification task, which then would either be grouped in a bag-of-words vector or an word embedding vector (MIKOLOV et al., 2013).

The training process would be the next step, using already classified samples, the training set, to weight the selected feature vector according to what is relevant to each answer class. There, many machine learning techniques can be used (examples are Naive Bayes, Decision Trees and Support Vector Machines). Finally, the results are compared with the testing set (another set of data), to see how the features and the technique parameters fare in the task. Another very common approach is to use cosine similarity to approach the question vertex to all answer vertices, trying to find which is more closely related (MOHAMED et al., 2012, p. 4).

Before jumping straight into the proposed approach, some attempts of providing a solution using statistical techniques were made. The first of these attempts was done using traditional Machine Learning techniques and a simple feature extraction mechanism. As such, features were extracted using the TF-IDF technique. These were then trained into a

MultiNomial Naive Bayes (MNB). Feeding all the questions and classifying them in 500 classes resulted in a P@1 of about 0.42. Still, it is clear that these are not reliable results, since no real test set was built - results were obtained by re-runs of the data through the model, possibly causing it to overfit.

In the second attempt, the problem was approached under a topic classification tone. In this case, a TF-IDF dictionary was built as a basis for feature extraction and then the questions were fed through a set of machine four learning algorithms, namely: Linear Support Vector Machine, Logistic Regression Classifier, MultiNomial Naive Bayes and Random Forest Classifier. The idea was to first identify the topic and then identify the correct answer. However, since the sections are not "sampled" equally, further process would be needed for better accuracy, resulting in successive complex task assignments. In face of the already identified problems with a question classification approach, the results obtained at this step did not seem to justify further development in this direction. The results are displayed in Table 4.

Table 4 – Topic classification results using four Machine Learning Techniques one the questions.

Model	**Mean Accuracy**
Linear Support Vector Machine	0.753801
Logistic Regression Classifier	0.611513
MultiNomial Naive Bayes	0.627780
Random Forest Classifier	0.332334

Source: Created by author (2019).

After the above mentioned tests, it became clear that a statistical approach was not appropriate to the current corpus number of samples. This directed the research to a model oriented task. Focusing on the model means to understand the context and the corpus more clearly, which then allows the creation of a feasible computational model to solve the original problem.

When attempting to understand the corpus, one clear detail was noted: the questions and the answers of the corpus came each from a distinct background. This caused that many questions seemed unrelated to their answers when addressed by someone from outside the dairy farming context.

The corpus analysis made it clear that the way that questions and answers were written did not favor a matching using traditional NLP methods. This is probably related to the fact that the book was curated to be a resource for quick references.

These characteristics led to a deeper study on the structure of the questions and

the answers, especially related to the seeming gap between them. The next title addresses this phenomenon in specific.

3.1.2 The Question Answering Gap

The question answer gap could be defined as a phenomenon where an answer, deprived of its motivating question, while grammatically and lexically well formed, cannot be considered semantically complete.

The natural language representation through which a question and an answer are expressed does not provide all the information needed to establish a bridge between them. This problem, very common to current statistical approaches, is pointed out by CAMBRIA; WHITE (2014), who recognize that

> most of the existing approaches are still based on the syntactic representation of text, a method that relies mainly on word co-occurrence frequencies. Such algorithms are limited by the fact that they can process only the information that they can "see" (CAMBRIA; WHITE, 2014, p. 48).

SINHA (2008) also describes a similar situation when analyzing State-of-the-Art QASs. As pointed by the author, in these systems, "selecting answer passages relies on a quantitative measure that evaluates the degree to which a passage shares the words in the expanded query keyword set" (SINHA, 2008, p. 18).

To understand this problem better, a statistical analysis was done based on the corpus. After an analysis of the corpus data, it was found that out of keywords extracted from questions and keywords extracted from answers, for all 500 questions, there's a median of 8% intersection between them.

At the same time, the median difference in size of characters between question and answer is that questions are about 10.2 times shorter than answers. Taking into account only the keywords that are intersecting and the keywords in the questions, the average is that 60% of question keywords are intersecting with answer keywords[1]. In integer numbers, it was found that 3 keywords intersect as average in the used corpus.

Figure 6 presents four histograms with data distribution. Note that the size relation between questions and answers follows a power law, which is in accordance with the Zipf's Law (ZIPF, 1949). Also, while the number of intersecting keywords follow some kind of normal distribution with an average of 3, the number of question keywords which intersect

[1] It is important to take into account that, due to the lack of good Named Entity Recognizers for Brazilian Portuguese, especially for dairy domain, some entities were not accounted in these statistics, which could cause an even bigger difference in these numbers. The problem with the entities was solved during the development of the final proposed solution where identified entities were directly inserted into the model.

Figure 6 – Histograms about keyword and size distribution. 7(a) presents the relation between intersecting keywords (appear both in question and answer) and total question keywords. 7(b) presents the number of intersecting keywords over total answer keywords. 7(c) presents the distribution of intersecting keywords and 7(d) presents the distribution of size relation between questions and answers in characters.

Source: Created by author (2019).

is scattered all over. However, the number of intersecting keywords over total keywords also follows a power law distribution, depicting the fact that questions have much less meaningful words than the complete answer.

Nonetheless, it is clear that not all question/answer pairs suffer from the Question Answer Gap as proposed. In fact, only 6 of the whole set of 500 questions didn't share any keyword with their answers. But to distinct degrees, most questions offered few informative keywords to match them to their answers.

Therefore, it can be seen that gap makes it difficult to align a question with the appropriate answer for most cases by using a traditional word frequency approach. One way to increase the overlap of concepts (and consequently the statistical similarity between two texts) would be by using query expansion mechanisms based on synonyms. Some

ways to do it would be by using resources such as WordNet or other types of thesaurus (SINHA, 2008, p. 18). Regardless, this approach would still keep the gap open, since expanding the query to synonyms doesn't grant that these new keywords appear in the answers. The knowledge behind the keywords is still absent.

To illustrate it, consider the following question extracted from the corpus used in this research:

> Quais os microelementos essenciais ao gado de leite? (*Which are the essential micro elements to dairy cattle?*).

The official answer found in the book is

> São cobre, ferro, zinco, cobalto, iodo, manganês, selênio, molibdênio, cromo, flúor, vanádio e silício. (*They are copper, iron, zinc, cobalt, iodine, manganese, selenium, molybdenum, chromium, fluorine, vanadium and silicon.*).

A reader can see straight up that if the question and the answer are treated as sets of keywords, the intersection between the two sets would be null. This null (or small) intersection between question and answer wording is what is being called "gap". Although the answer is grammatically correct, no knowledge can be extracted by treating it in isolation. It can be argued that the answer *depends on* the question to provide knowledge.

Figure 7 – An example of a simple "Match the Column" activity. Note that without previous knowledge of what Cow, Milk or Pasture actually are, no matching could be made aside from random guessing.

Match the two columns.

• (A) Cow	• ( ) A nutrient rich, white liquid produced by the mammary glands of mammals.
• (B) Milk	• ( ) An area where farmers keep livestock for grazing
• (C) Pasture	• ( ) A large quadrupedal mammal usually domesticated and raised to produce milk or meat.

Source: Created by author (2019).

On another hand, its not just because a question and its answer do not share the same wording that they cannot be matched together. Take, for example, a common game used in learning activities and tests, where two columns are presented to the person being tested and the person has to match them. Usually, in order to make the test challenging,

the two columns does not share the same concepts or wording. How, then, can the person makes the linking between them? The answer is in the underlying knowledge structure that the person should have acquired previously to the test in order to perform well - this could have been done through studying or observation. This knowledge structure functions as a bridge to cover the gap. A simple example is presented in Figure 7.

With these examples in mind, it is clear that question-answering requires much more than statistical guessing to be successful. There's a need for one underlying knowledge structure which can encompass the subject domain knowledge and be queried providing links between what is asked and what has to be responded.

This research proposes to use connectionist models as the supporting structure to solve the question answer gap. The following sections will describe the development of the proposed approach, as well as analyze the efficiency of the technique in solving the question answer gap.

3.2 A GRAPH BASED APPROACH FOR QUESTION ANSWERING: THE PROPOSAL

3.2.1 **Development steps**

To better understand the steps that are going to be presented over Section 3.3, this section presents the reasoning logic in its distribution.

As previously presented, it was decided to attempt a model based solution. To elaborate a computer model that represents the context in a way that it can be used in a question answering system, the first step is to retrieve knowledge about the context. This can be done in several ways. Section 3.3.1 describes some attempts and the final process used: manual annotation along with NLP techniques.

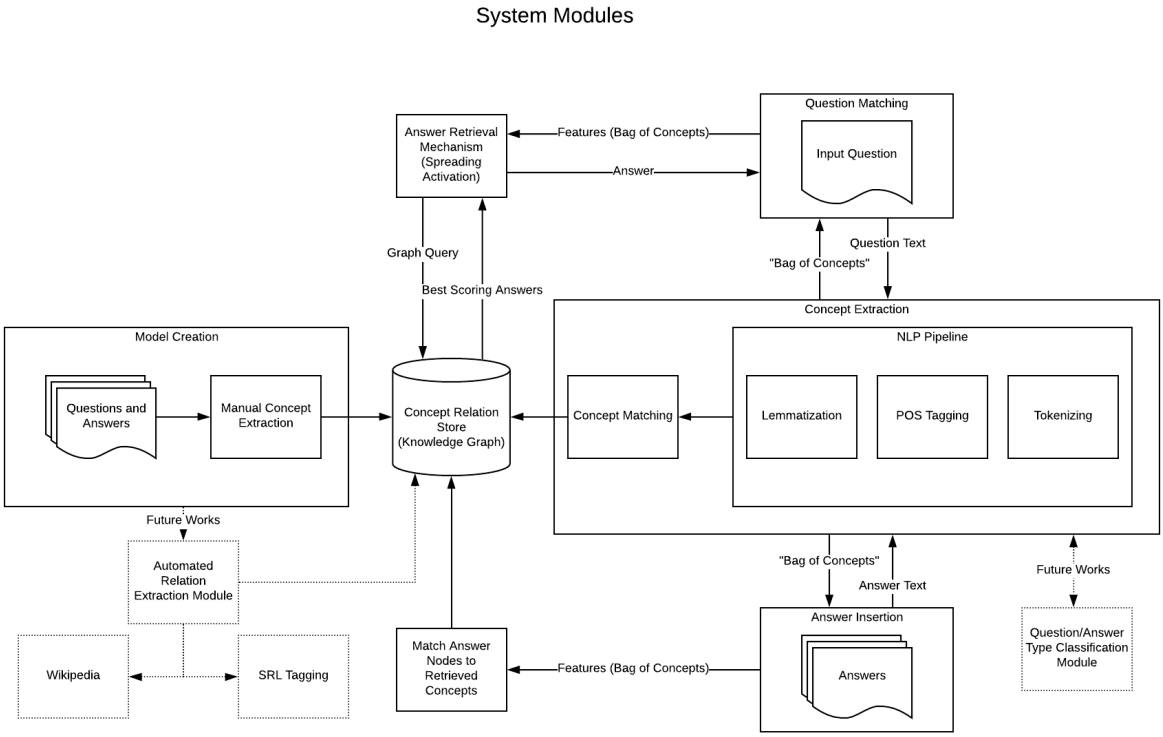
The following activity involves getting the extracted knowledge together in a computational structure. This was done using a Knowledge Graph, whose construction is presented in Section 3.3.2.

Finally, to retrieve the answers from the Graph, some technique has to be used. It was decided to use the "Spreading Activation" technique. This network technique is presented in Section 3.3.3.

3.2.2 **Approach Architecture**

For a better understanding of the developed work, Figure 8 presents a diagram representing the organization of a QAS architecture according to the proposed approach. Also, the Figure displays some proposals for future work and show how they fit in the schema - these are going to be discussed in Chapter 4.

Figure 8 – The developed system architecture, with a focus on the modules.



Source: Created by author (2019).

Using the criteria proposed by (MISHRA; JAIN, 2016) can help summarize the type of QAS that can be developed using the proposed approach. This can be compared to the one presented in Section 2.1 for the system proposed by (CRISCUOLO, 2017). The classification of the approach resulting from this research could then be described as follows:

1. Application Domain: restricted domain - Dairy Farming.
2. Types of Question: although it was not a matter of deep analysis, factoid, confirmation and list questions were identified.
3. Types of Analysis: Morphological, Syntactical (NLP pipeline), Semantic and Pragmatic/Discourse (this last analysis is the core of the Graph-Based approach, even though it is not formally described as such).
4. Type of Consulted Data: Unstructured, since in a IR point of view a Book in Natural Language is considered an unstructured data source. Also, attempts were done in the use of Semi-Structured Data Sources through the use of Wikipedia but were note used in the final solution.
5. Data Source Characteristics:

a) Size: the data source could be considered very small, since it does not provide multiple individuals for each class.

b) Language: Brazilian Portuguese.

c) Heterogeneity: Not heterogeneous, the data is all in a single location (the selected book).

d) Genre: While a formal book, it does contain a selection of customer questions which are mostly presented in informal language. Therefore, it can be both considered formal (questions) and informal (technical answers).

e) Media: Text media extracted from a book.

6. Types of Representation and Matching Function: Conceptual Graph Model.

7. Types of Techniques used: Based on Knowledge Retrieval and Discovery.

8. Generated Answer Forms: No answer was generated, but rather extracted from an answer set.

3.3 APPROACH DEVELOPMENT

3.3.1 **Extracting Knowledge from Corpus**

To build a computational model, it is necessary to obtain knowledge from the domain in some format that can be represented and comprehended by computers. This process of extracting knowledge can be either made manually or automatically.

In this research, the handmade model substitutes the automated probabilistic one commonly used together with Machine Learning technique. This model is used to provide the underlying knowledge structure necessary to fill the gap between questions and answers, allowing for the unwrapping of condensed knowledge offered by the corpus.

The first step in building a knowledge model is to establish what are the knowledge units that are going to be the basis of the model. This varies depending on the structure the work is based on - for example, in Ontological models, more than just classes, it has to be looked for instances of classes. In this work, the knowledge units used were concepts.

“Concepts” can be defined as “the sorts of things that get grasped, possessed, or understood in coming to have beliefs (and ultimately knowledge) about the world”(WASKAN, a). They can be represented by words, images or even abstract thoughts. In the case of this research, the concepts are represented by the words in the text (such as “cow”, “pain” and “to milk”) and multi-word expressions (one example would be “bovine parasitic disease”).

The first attempt to extract concepts was done using a corpus collected from various online sources about dairy farming. Since the original idea was to gather the most

Table 5 – Syntax patterns used to extract knowledge from an earlier corpus.

Syntax Pattern	Outcome Example	Extracted Knowledge
Noun Preposition Noun	Farelo de Soja (Soybean Meal)	Farelo (Meal) can be of type Soja (Soybean)
Noun (for) Verb	Água para Irrigar (Water for Irrigation)	Água (Water) can be used for Irrigação (Irrigation)
Verb Preposition Noun	Irrigar com Estrume (Irrigate with Manure)	The act of Irrigar (to Irrigate) uses Estrume (Manure)

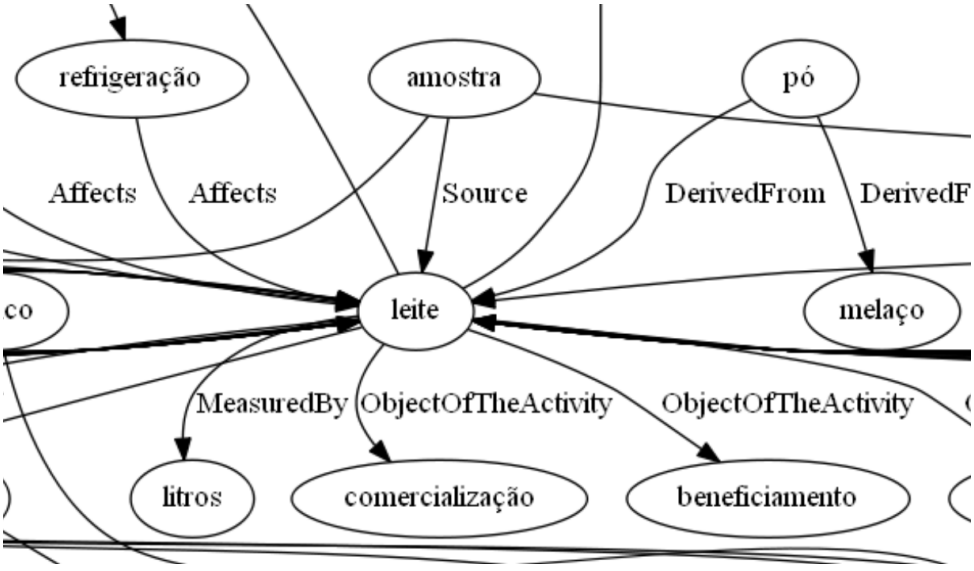
Source: Created by author (2019).

informative concepts in the domain to use in a statistical based approach, this was done without much worry with corpus uniformity. At the end of this first attempt, about 23,000 lines of text were collected.

The subsequent attempt to extract concepts involved providing a Text Analysis Framework some syntax patterns which could be used to extract concepts and the relation between them. Table 5 shows some of the patterns used, an example and the knowledge that can be extracted from the pattern.

This syntax matching process resulted in over 4,400 extracted relations. However, due to what seemed as problems in the Framework NLP processing pipeline, it was noticed that these results had to be handpicked. The resulting set of usable relations were 834, including two-way relations (in the case of relations where there's a reciprocal, such as "uses" and "used by"). This handpicking process was very slow and tedious, especially when done by a single person. Figure 9 presents a small piece of the graph resulting from this modeling attempt - in total, the graph had 532 unique nodes.

Figure 9 – Part of the graph developed in the first attempt to model domain knowledge.



Source: Created by author (2019).

This first developed model, however, proved to be ineffective. Many important relations were missing, since they were not recovered by the syntax matching process.

Also, with the failure in obtaining the original corpus, a new model proved necessary due to the absence of many of the new corpus concepts in this first structure.

The new attempt was completely focused on the corpus. The text from both questions and answers was the basis to extract the model concepts. This new attempt was driven towards a more manual process, where an annotator went on reading the text, picking the concepts and inserting their relations in the model.

Since the corpus is composed of 500 questions and answers about several distinct subjects, reading and interpreting all of them would be a research by itself. In order to try to encompass the whole book, 27 out of the 500 questions and answers were chosen to be manually processed and tested. This number of questions and answers was determined by the fact that it would be necessary to have at least one representative for each Section (resulting in a total of 11 questions) combined to an attempt to measure the solution efficiency inside a single question domain, which corresponds to another 14 extra questions about Animal Health. Finally, two extra questions were selected due to special features, one due to small answer length and the other because it used concepts more common to another section.

To model these questions efficiently and then be able to recover the information from the model programmatically, some NLP techniques had to be used. Also, the use of some of these techniques helped in the process of cleaning and preparing the data for processing. This step was not present in the first attempt, due to the fact that a closed box framework was used instead.

Considering that the book used as corpus was originally available as a downloadable PDF, the questions and answers had to be extracted and correctly correlated to their numbers in order to provide a means for testing afterwards. Therefore, a set of techniques that combines information extraction and pattern matching had to be used in order to process the file and provide the data in an organized manner. One thing to consider is that in some answers, more than just text, provides tables and formulas. This work did not take any of these into account, however it can be pointed out that in the modeled structure, they would fit naturally with some minor tweaks[2].

After cleaning and preparing the data, it was time to start extracting information. Even though word sequence, tense and variation of gender and degree bear important information[3], it was opted to create the model without delving deep into these variations. While this information offer a more fine-grained control over the modeled data, it also

[2] For example, these knowledge bearing objects could be encapsulated in a natural language description and then referenced by related concepts.

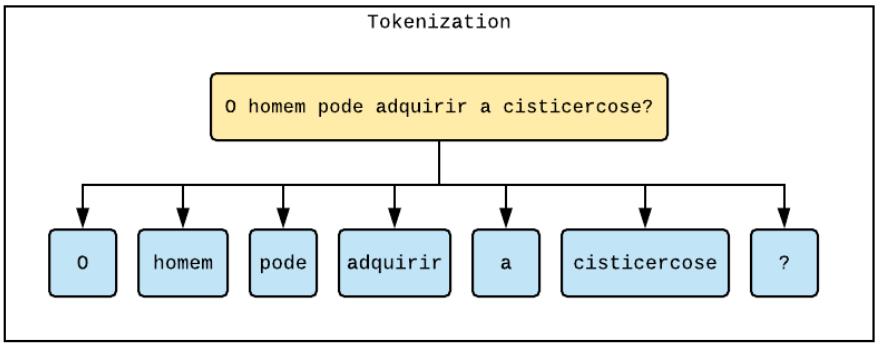
[3] Portuguese is one of world languages where these aspects are presented through a variation of the words. Also, a rather large list of verbs follow an irregular pattern, such as happens in English. Due to these reasons, a word that carries a core meaning can be represented in many distinct forms.

makes the process of hand modeling more complex, tiresome and prone to errors.

To extract the information, a pipeline was built composed by three main steps: Word Tokenization, Part of Speech (POS) Tagging and Lemmatization.

Tokenization is the act of chopping a Natural Language entry (in this case, a string) into its more granular sub-parts. The task includes breaking a paragraph into sentences and sentences into words (or tokens). This process can be rule-based or statistical (MANNING et al., 2008, p. 21, 22) and the result is a set of smaller strings where each one is a token. Figure 10 provides a visual exemplification of the process.

Figure 10 – A simple depiction of the Tokenization process.



Source: Created by author (2019).

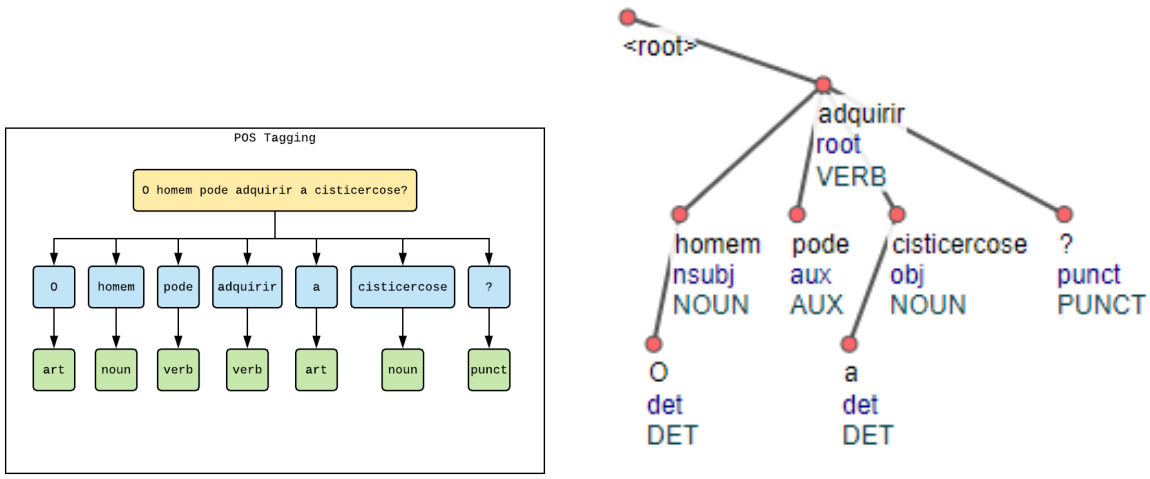
POS Tagging is the next step in the pipeline and can be resumed as attributing the Part of Speech[4] for each word. This process is language dependant and is usually based on pretrained machine learning models currently available widespread for many languages. For this work, the model used was *nlpnet*, a model provided by NILC (*Núcleo Interinstitucional de Linguística Computacional*/Interinstitucional Center for Computer Linguistics), a lab hosted by the University of São Paulo (USP). This model, based on a Brazilian Portuguese News Corpus is currently the most accurate publicly available tagger for this language, with about 93,3% accuracy for a corpus with out-of-vocabulary words[5]. This tagger is described in (FONSECA et al., 2013).

Figure 11 provides two examples of POS tagging. The figure at the right is a

[4] Parts of speech are categories that words are assigned according to their syntatical function in a phrase. Some examples of Parts of Speech are Verbs, Nouns and Adjectives.

[5] POS Tagging models are usually tightly coupled to the vocabulary they use, since statistical distribution composes their core. Words outside the vocabulary have their tags guessed according to their surroundings or characteristics, depending on the model. Since the used corpus has several domain specific words that do not commonly appear in a News text or have a different meaning, many out-of-vocabulary tagging are supposed to occur. For more information about *nlpnet* and implementation details, visit `http://nilc.icmc.usp.br/nlpnet/index.html` - Last Accessed May 10th, 2019.

Figure 11 – Two figures representing the Part of Speech Tagging process.



Source: Created by author (2019).

dependency tree obtained through the use of UDPipe[6], a Dependency Parser Pipeline based on the Universal Dependencies project[7]. This project is based on a technique known as Dependency Parse, which is based on Machine Learning algorithms and is considered to have almost solved the POS tagging problem[8].

The above mentioned phases of the NLP pipeline do not actually cause any modification to the processed text, but rather prepare the text for a better comprehension on the computer side and extracts and aggregates semantic information constrained in a syntactical and lexical form. So far, it can't be said that any information was lost. The next step in the adopted pipeline simplifies the entry, exchanging information for better computer representability - it's the Lemmatization process.

Lemmatization means reducing any lexical unit (or word) to its lemma. A lemma is a word base form, stripped of any temporal, gender or degree variation. This means that verbs are presented in infinitive form (if the language allows so), nouns and adjectives are always in the unmarked gender and singular and so on. This process usually helps reducing the amount of input words to be treated and maintained into a model and fits well with concept representation. However, as it has been mentioned, this can have a deep impact into semantic analysis through the removal of relevant traits. For this work, it was

[6] An online service can be accessed in `http://lindat.mff.cuni.cz/services/udpipe/` - Last Access June 5th, 2019.

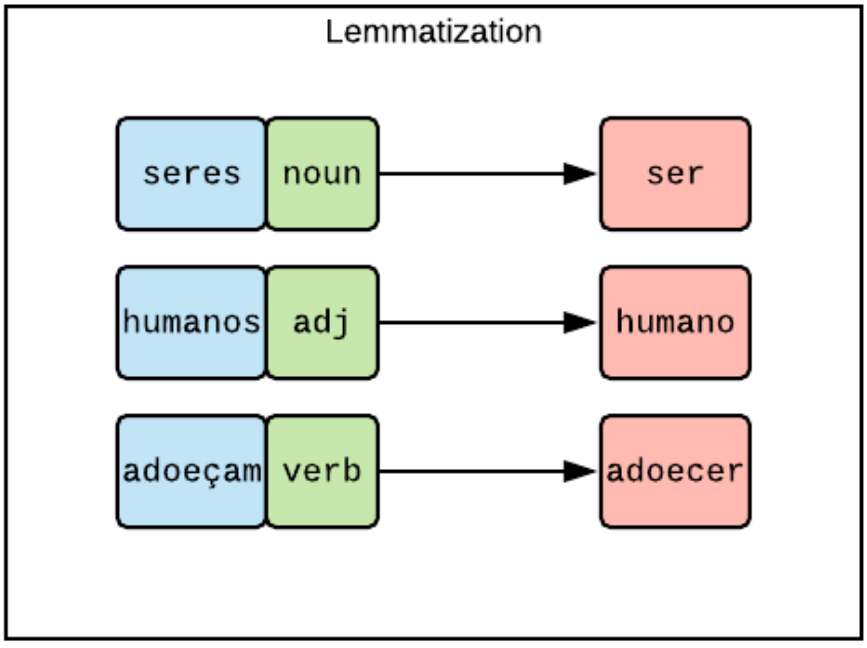
[7] Universal Dependencies is a Framework developed for cross-linguistical grammatic annotation. It is available for European Portuguese. For more information, see `https://universaldependencies.org/` - Last Access June 5th, 2019.

[8] For some more information about the actual State-of-the-art in POS tagging, visit `https://aclweb.org/aclwiki/POS_Tagging_(State_of_the_art)` - Last Access June 5th, 2019.

opted to make this exchange so as to simplify the model, since the focus are in the core concepts and their relations. The activity is depicted in Figure 12.

Lemmatizers rely on rules and extensive word annotations. So far, few good Brazilian Portuguese lemmatizers are available and accessible. At one point, it was opted to develop a tool that used a thesaurus to do word look up and replace for its lemma - but the available open Portuguese dictionaries are not extensive enough to cover the whole vocabulary, and they do not provide POS based disambiguation for lemmatizing. With this, due to the simplicity of implementation, it was opted to use LEMPort, part of NLPPort, a set of tools developed for the (European) Portuguese pipeline. The adoption of this solution required some tweaks to be made, since some of the rules used for gender neutralization changed important concepts for the domain, such as lemmatizing "cow" into "bull". NLPPort development, research and proposals are described in (RODRIGUES et al., 2018).

Figure 12 – A simple depiction of the Lemmatizing process.



Source: Created by author (2019).

Having each token reduced to its lemmas, a final step was used to select the relevant ones to compose the model. This filtering process involved using some heuristics and outer resources to decide which words are to be treated as concepts by the system. This process, named "Concept Modeling", could be defined as "the process of formulating and collecting conceptual knowledge about a Universe of Discourse" (MATOS, 2008, p. 18).

In order to grasp the main concepts of the domain, all nouns followed by adjectives

were selected and inserted into the system with an "attribute" relation between them. Afterwards, using a Semantic Role Labeling (SRL) model provided by the same authors as the POS Tagger, it was attempted to capture simple semantic relations, such as subject-action-object relations. However, with a precision score of about 66% for news domain semantic roles, most tagged relations were of little use for the modeling process.

As an attempt to automate some of the relation extraction, one last preprocessing step was taken which included the use of semi-structured data available publicly in the internet. The resource used was the Portuguese version of Wikipedia[9]. In this processing phase, all extracted nouns and verbs were queried to Wikipedia. Since Wikipedia uses hyperlinks to point to related pages, the related pages list is retrieved and compared to the set of terms extracted from text - those matching were then selected to be inserted in the model with a generic "Wikipedia" relation in between. A description of a related use of English Wikipedia is made in (GOUWS et al., 2010).

One important parenthesis must be made on this relating of this approach weakness and how it could be mitigated. Since it's based on collaborative content, Wikipedia is not a specialized source of knowledge and have articles ranging from simple descriptions of a subject to specialized, catered technical sources. This difference is especially noted for languages outside English - while the anglo-saxonic language had, in 2014, 1.7 billion words, Portuguese and Spanish together had about 500 million words (the second language being twice as big in number of words in relation to the first one)[10], which can be easily seen as reflecting in the quality of the articles. Therefore, while some of the concept pages were very precise and provided meaningful relations to the context, others didn't fare as good. Also, the lack of a precise disambiguation technique also affected the results, since some words represent several concepts, each of which have different relations to other concepts in the domain - addressing this problem could lead to better results, but would compose a matter of research itself. Another problem with the Wikipedia data extraction approach is related to the "generic" assignment of relations, since a crucial part of the knowledge that each relation encompass, its type, is ignored. This could be addressed by a set of heuristics and castings, but as said for disambiguation, this would compose a full research by itself.

The usefulness of Wikipedia in concept extraction could be described as follows: Assuming A as the set of words extracted from a text, W as the set of Wikipedia pages and C as the set of domain concepts referenced by the text. Taking two words $a, b \in A$ if there's a relation $a \rightarrow b \in W$, represented by a hyperlink between a page referred by a and a page referred by b, it is highly probable that $a, b \in C$. Figure 13 depicts a few relations

[9] `https://pt.wikipedia.org/` - Last Acessed May 28th, 2019.

[10] Data was collected from the statistics page held by Wikimedia Foundation, available at `https://stats.wikimedia.org/EN/Sitemap.htm#comparisons` - Accessed in May 13th, 2019.

Figure 13 – Concepts retrieved using Wikipedia. Note that while not qualified, all the described relations are plausible.

Source: Created by author (2019).

obtained using this method - note that all concepts referenced are somehow related within the context.

Even though automating the extraction of concepts is useful and labor saving, the above mentioned restrictions proved it to be unfeasible at this point of the research. Besides, in order to have better control over the model and focus on the QA task effectiveness, it was opted to develop the model manually based on the corpus text. This manual process and the resulting structures are going to be described in the following section.

3.3.2 **Representing Knowledge Through Graphs**

According to connectionist theories, Knowledge is conformed not by isolated fragments, but is better understood as an interconnected structure where every unit is

"plugged" to a number of other units to which they share some type of relation. This proposal is based on researches about the neuron nature, which concluded that neurons are organized in a network manner.

These researches found that when one of the brain terminations receive an input, an electrochemical pulse traverses through the neuron network which then activates related reactions, feelings, memories and thoughts. One situation that vividly depicts this idea is that of autobiographical memories evoked by odors, or in other words, when feeling a certain specific smell, a person brings to mind nostalgic memories or feelings that were composed when first smelling that odor. Studies propose that the same thing happens to words and concepts (ARSHAMIAN et al., 2013).

In a similar way, a Knowledge Graph attempts to make a computational representation of domain knowledge through the building of a network model. In this model, every node is a knowledge unit and every relation represents how two units are linked together. These relations can be, or not, weighted and qualified, depending on the modeling process. Also, in such a Graph, not all units need to be of the same type, allowing for distinct types of knowledge to be kept together and referenced by a single structure. As presented in Chapter 2, KGs share properties with Ontologies, as will be clearly evidenced by the network construction process described in the following lines.

As pointed out in the previous section, the modeling process used in this work was done manually by a human annotator. This annotator read the selected questions and answers and then highlighted the main words which represented the text concepts. These concepts were then related to each other using a set of predefined "Qualia Relations" as a base for comparison.

Qualia relations are based on "qualia structures", proposed in (PUSTEJOVSKY, 1991). They are the representations of how two entities relate to each other. In general, philosophical researches divide qualia in four types: (1) formal quales, representing is-a relations between entities; (2) constitutive quales, representing part-hole relations between entities; (3) telic quales which represent the purpose of use of an entity in relation to another; and (4) agentive quales, which encompass agent-patient roles in a cause-consequence manner (KAZEMINEJAD et al., 2018, p. 2645).

However, it has to be pointed out that while there is some consensus when defining qualia categories, there's no universal set of qualia relations and rules for qualia granularity (KAZEMINEJAD et al., 2018, p. 2645). Usually, the set of qualia relations in a knowledge model is defined by the domain knowledge in question and by the granularity needed by the model.

To define a new set of relations in a more structured manner, this research used the SIMPLE Qualia Structure as presented in (LENCI et al., 2000). The SIMPLE

Qualia Structure, created to allow multilingual lexicon building, is built on three types of formal entities: **Semantic Units**, which encodes words senses (ways to represent a concept through words); **Semantic Types**, which delimits the relation types between two Semantic Units; **Templates** which proposes a schematic structure that acts as a guide for lexicographers (LENCI et al., 2000, p.6). This definition and the base set of relations proposed in (LENCI et al., 2000) guided the extraction of conceptual relations from the corpus, resulting in the final graph structure used in this research. Table 6 is provided to better contextualize the SIMPLE Semantic Types with the corpus context.

Table 6 – Relation types according to the SIMPLE Ontology project. (LENCI et al., 2000) and applied to this research context.

ID	Semantic Type	Description	Template	Reciprocal	Type
1	activityOf	This relation denotes a lexical relation.	**infectar.v** *activityOf* **infecção.n**		Telic
2	affects	A relation where the first entity provoques any effect over the second entity by affecting it.	**coma.n** *affects* **animal.n**	affectedBy	Constitutive
3	atLocation	A relation that describes the position of the first entity in relation to the second.	**bezerro.n** *atLocation* **pasto.n**		Constitutive
4	attribute	The first entity has the seccond entity as an attribute or characteristic. Usually an adjectival relation. Most of them can be automatically extracted using POS tagging.	**ambiente.n** *attribute* **úmido.adj**		Constitutive
5	causes	The first entity is te reason why the second starts exhisting.	**verme.n** *causes* **verminose.n**	causedBy	Constitutive

7	composes	The second entity doesn't exist without the first, from which it is composed.	**enxofre.n** *composes* **sulfa.n**	composedBy	Constitutive
8	consumes	The second entity ceases existing by being used by the first in a consumption relationship.	**bezerro.n** *consumes* **colostro.n**	consumedBy	Telic
9	hasA	A relationship where entity 1 is directly related to the existence of entity 2.	**aborto.n** *hasA* **causa.n**		Constitutive
10	hasAgent	First entity is a situation or action realized or created by the second entity.	**intervenção.n** *hasAgent* **médico.n**	agentOf	Agentive
11	hasAsPart	A compositional relation where entity 1 is composed, but not dependant for existance on, entity 2.	**árvore.n** *hasAsPart* **copa.n**	partOf	Constitutive
12	hasAsProperty	Defines a property relation between a material entity and a categorical entity. More generic than the attribute relation.	**urina.n** *hasAsProperty* **cor.n**	propertyOf	Constitutive
13	instance	Entity 2 is a materialization of the categorial Entity 1.	**cor.n** *instance* **vermelho.n**	isInstanceOf	Formal
14	isA	Entity 1 is a more granular category in relation to Entity 2.	**intestino.n** *isA* **órgão.n**		Formal

16	measures	Entity 1 is a simbolic unit used to measure Entity 2.	**dose.n** *measures* **remedio.n**	mesuredBy	Constitutive
17	objectOfTheActivity	Entity 2 is the typical use if Entity 1, which is a material object.	**alimento.n** *objectOfTheActivity* **ingestão.n**		Telic
18	opposite	Entity 1 is antonymic to Entity 2 and vice versa.	**rehidratar.v** *opposite* **desidratar.v**		
19	produces	Activity innerent of Entity 1 results in the creation of Entity 2	**vaca.n** *produces* **leite.n**	producedBy	Constitutive
22	relatedTo	Entity 1 is generically related to Entity 2. Used to resemble lexical relations.	**infectado.adj** *relatedTo* **infecção.n**		Constitutive
23	requires	Entity 1 causes a demand for Entity 2.	**doença.n** *requires* **intervenção.n**		Agentive
24	resultOf	Entity 1 is the result of the ocurrence of Entity 2, which usually is a verb.	**morte.n** *resultOf* **morrer.v**	resultsIn	Constitutive
25	sameAs	Used for synonymy.	**remédio.n** *sameAs* **medicamento.n**		Formal
26	type	Entity 2 is a more granular instance of Entiy 1, which is an instance of some category.	mosca.n type bicheira.n	typeOf	Constitutive

27	typicalOf	Entity 1 is an action (usually a verb) typical of Entity 2 (usually a noun).	**urinar.v** *typicalOf* **animal.n**		Constitutive
28	usedAgainst	Entity 1 purpose is to affect negatively Entity 2.	**armadilha.n** *usedAgainst* **mosca.n**		Telic
29	usedBy	Entity 1 is used, by choice or as by nature, by Entity 2.	**brinco.n** *usedBy* **animal.n**	uses	Telic
30	usedFor	The use of Entity 1 results in Entity 2. Can represent immaterial entity relations.	**vacina.n** *usedFor* **prevenção.n**		Telic
31	usedIn	Entity 1 is used by Entity 2 passively, or without a choice. Also points to locational use.	**repelente.n** *usedIn* **animal.n**		Telic

Source: Created by author (2019).

At the end of the annotating process, a total of 2,965 relations between concepts were modeled, being 2,205 of them "attribute" relations extracted by the use of an automated pattern looking for "noun-adjective" syntactic relations. The other 760 relations were manually extracted through the described process. Between these relations, a total of 311 concepts were manually modeled from the 27 selected questions/answers.

The use of this network like structure aims to provide a query expansion mechanism in order to solve the question-answer gap described in Section 3.1.2. While modeling the concepts, extra care was taken in order not to leave question concepts isolated in the network, as if functioning merely as a bag-of-words bucket.

As an illustration of the proposed method, take, for instance, the question numbered 405: "O Homem pode adquirir cisticercose?" (*Can man acquire cysticercosis?* - self provided translation). The official answer is:

Sim. Mas não pela ingestão de carne contaminada. A instalação da doença

> ocorre pela ingestão acidental de ovos do cestoide, eliminados nas fezes de uma pessoa que apresente a tênia adulta em seu intestino. Essa situação pode ocorrer em ambientes sem higiene ou a partir de atos promíscuos. Por isso, é importante o estabelecimento de programas de educação sanitária.[11]

As it can be seen, a search for keywords in the question "Homem" (*Man*) and "Cisticercose" (*Cysticercosis*) would find no match in the answer. On another hand, just to mention just a few, the answer to questions 404, 411, 413 and 414 contains the word *Man*, but within different contexts. A method that only took into account the keywords in the answer as an isolated form would probably score any of those as better fit than the official answer. However, at going further than keyword matching and linking represented concepts to distinct word forms as in connecting *Man* to *Person*, *Cysticercosis* to *Cestode* and *Taenia* and so on, the query can be expanded using these new words, establishing a path that gives chances for the correct answer to be selected.

One last stop in building the Network was the "training" phase. While playing with the words, this activity is not related to machine learning techniques as usually associated with the term "training". Instead, this phase included adding the answers to the network so that they could be retrieved. In order to do that, all answers had their text processed following the NLP pipeline steps mentioned in the previous section. With the answer keywords extracted and lemmatized, each of their representative concept nodes were pointed to that specific answer node through an "answer" relation.

After adding the answer to the network and linking it to the related concepts, it can then be retrieved through these relations using the concepts extracted from the questions. For this to work, a network specific technique has to be used. This technique is going to be described in the following section.

3.3.3 **Spreading Activation**

The Spreading Activation technique is a common technique used in Connectionist Models. Just as an example, it has been used over the last few years in activities such as Word Sense Disambiguation (TSAOIR, 2014) with the use of WordNet (MATOS, 2014, p. 78) and FrameNet (MATOS, 2014), two well known proposals for NLP/NLU network based models.

This technique works on top of network like structures (Spreading Activation Networks - SANs) and, simulating a neurological pulse, "spreads" an input over the

[11] Translation: *Yes. But not through the ingestion of contaminated meat. The installation of the disease occurs by the accidental ingestion of the cestode eggs, expelled by the feces of a person who presents adult taenia in its intestine. This situation can occur in environments that lack hygiene or from promiscuous acts. Therefore, it is important the establishment of sanitary education programs.*

network and retrieves the activated nodes as an output. Usually, there's one or more entry points (nodes) from which the activation is "spread" over the neighboring nodes. During this spreading phase, which occurs in "pulses" (GOUWS et al., 2010, p. 47), if a node is activated by more than one relation, the activation values stack up, highlighting nodes that are more central to the network.

At every pulse, the activation energy (or value) is reduced according to a predefined activation function (one example would be a variation of the logistic function, as presented in (MATOS, 2014, p. 175)). Although it was experimented with the other functions, such as logistic and ReLU, the final results were obtained by the use of a simple decay multiplication function.

The stopping condition can be given by a threshold constraint imposed on the energy decay or through a distance constraint (GOUWS et al., 2010, p. 48). The idea is to avoid activation to spread for too long, which would increase the technique complexity and reduce accuracy. The importance of this type of constraint was noticed during the test sessions, where increasing the depth of spreading from 2 to 3 caused a drop in accuracy from an average of 40% down to 7%.

As a type of network traversal technique, there are different customizations that can be used to achieve the results. For example, the traversal technique can be implemented either as Depth-first (GOUWS et al., 2010, p. 50) or as Breadth-first (MATOS, 2014, p. 78). While Depth-first traversal can be more efficient and less memory heavy (which might be a concern in very large networks) with the use of constraints, it does not allow for an easy implementation of activation functions that channel incoming activation pulses to an increased output value due to its linear nature. The impact of this was verified during earlier testing done with the first 11 annotated questions, where breadth-first scored an average of 46% accuracy (with top accuracy of 53%) while depth-first resulted in an average of 33% (with top accuracy of 40%).

GOUWS et al. (2010) point to several other constraints and measurements that can be used in a SAN and which provides improvements to the results by limiting spread. One of them is qualifying (through weights) the relations between nodes. Even though the distinct relationships presented in the previous sections were not weighted differently during the course of this research, this is a possible improvement that could increase the accuracy of the system.

Nevertheless, relationship weighting constraint was used in answer type relations to avoid that longer answers get better results (because there are more related keywords). This means that, for every answer, each of incoming relationships have a weight represented as a fraction of 1. The weighting scheme for answer relationships is given as follow:

Taken an answer A, its set of retrieved concepts represented by word lemmas and

their POS $Cc_1, c_2, ..., c_n$, the number of total concepts in A as $|A|$, the weight $w_{c \to A}$ of a relation between a concept c and the answer A is given by the number of occurrences of c over the total number of concepts $|A|$. This is presented in Equation 3.1.

$$w_{c \to A} = \frac{|c|}{|A|} \tag{3.1}$$

Another possible weighing scheme, as pointed by GOUWS et al. (2010), would be to implement something similar to Term Frequency-Inverse Document Frequency (TF-IDF) for every answer relation.

Through the Spreading Activation Approach, retrieving the correct answer for a question would then mean selecting the question concepts through NLP processes, spreading over them in the network and finally retrieving the most well activated answers. This simple process could be improved by several other techniques, such as by adding question and answer types as nodes in the graph, providing "shortcuts" to disambiguate answers that share the same concepts but with different purposes (such as questions 404 and 405, which are similar in matter but different in intent).

Overall the use of the presented model and the spread activation technique proved to be able to target the selected problem. However, as will be pointed out in the following two sections, the proposed model is far from being sufficient, requiring several future improvements in order to answer questions in equal number to traditional approaches.

3.4 RESULTS AND COMPARISONS

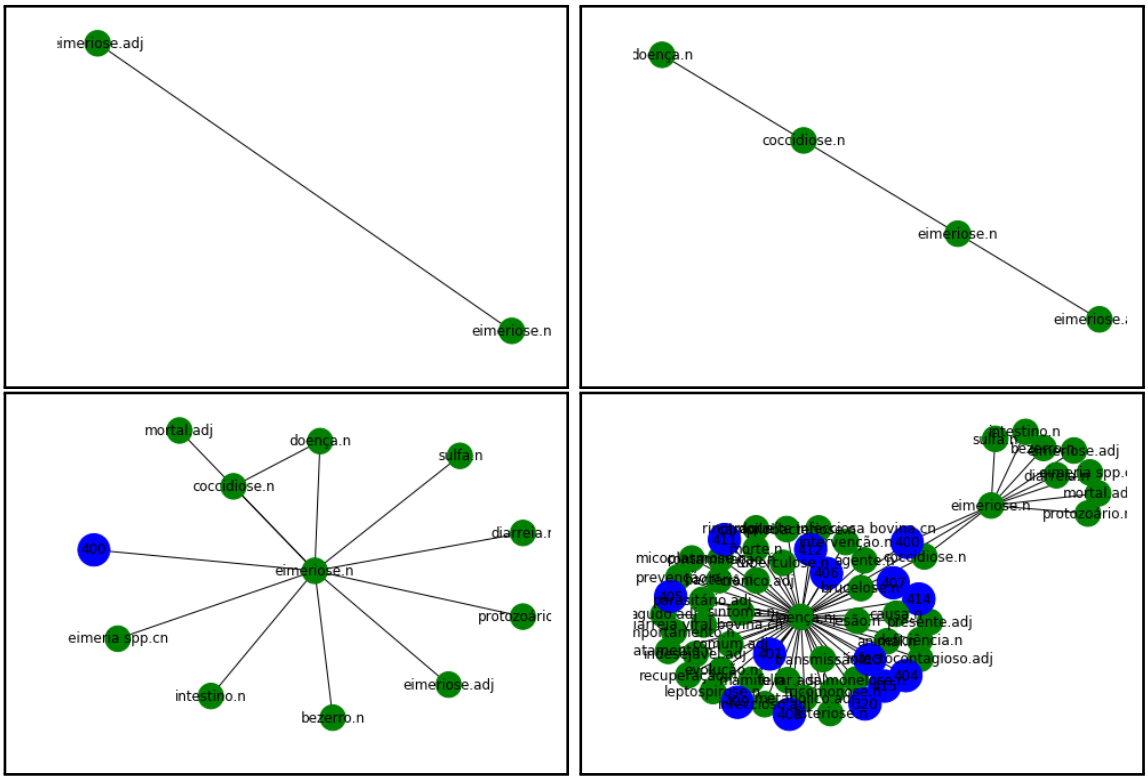
Before treating numerical results, it is interesting to discuss about some advantages that were perceived by the proposed method in comparison to other techniques. When compared to probabilistic methods, or even to ontology based methods, some advantages are to be pointed out:

First, the proposed approach is easy to debug and visualize, since its result retrieving process can be printed graphically and easily comprehended - this, in turn, can make it easy for annotators to correct the underlying relation structure if some unexpected leak is visualized. As an example, Figure 14 shows a picture of the resulting activated graph after each pulse on evaluating Question 400.

Second, since it is a text based conceptual model rather than a formal ontological model, with little research a nonspecialist annotator can do the annotating process - the model used, for example, was created by a single computer scientist with little knowledge on dairy farming, aside from the one obtained through the corpus reading itself.

Third, while changes in the model structure can impact the results and the spreading paths, as a network structure, this only happens locally and does not require a complete

Figure 14 – A visualization of the spreading process over Question 400. The order is Left-Right, Top-Down. Numbered dots are answer nodes.



Source: Created by author (2019).

recreation of the network at each modification. Therefore, additions and modifications can be done to the graph without risk of unbalancing the entire model - this does happen, for example, when changing Machine Learning Models, where any modification requires a complete re-training of the structure.

While it has to be acknowledged that the building process of a manually tailored graph structure is much more time consuming than training a machine learning weight vector, the graph structure is easily scalable and could compensate over time. This was proven during testing, when adding new answers to the system did not change the outcome of previous answer matching.

Finally, the mainly expected result was achieved: the gap between question and answer was mostly dealt through the use of the spreading activation technique, allowing for questions to match related answers even if they do not share the same vocabulary. This can be seen in Table 7, where almost all questions where the QA Gap could be identified found an answer either at the most ranked, or among the 3 most ranking answers.

In conclusion, the proposed approach allows to relate questions and answers with the proposed corpus without promoting overfitting. Also, the model proposed in the approach

Table 7 – Result of the questions selected for testing after using the best parameters found (2 depth, 0.3 decay and 0.05 Threshold). Questions with * are questions that would clearly fit the description of the "Question Answer Gap" - note that only one of them (408) no answer could be found even among the top scoring ones.

Question	**Number of Intersetcting Concepts**	**Precise Hit**	**Answer in top 3**
11	2	No	No
64	4	No	No
93	2	Yes	-
142	2	No	No
299	4	Yes	Yes
320	2	No	Yes
361	2	No	No
400	3	Yes	-
401	3	Yes	-
402	3	Yes	-
403	1	Yes	-
404	1	No	Yes
405	0*	No	Yes
406	0*	No	Yes
407	2	No	Yes
408	0*	No	No
409	5	Yes	-
410	2	No	No
411	0*	No	Yes
412	0*	Yes	-
413	2	Yes	-
414	2	No	No
415	1	Yes	-
457	2	No	No
466	3	No	No
476	2	No	No
483	2	Yes	-

Source: Created by author (2019).

allows for a computational way to solve the question answering gap, partly automating the question-answering process. Therefore, the questions posed in the introduction were answered and confirmed.

It has to be recognized that the proposed solution at most reached equal performance to results when compared to traditional approaches for Brazilian Portuguese. Using P@1 (this evaluation method was described in 2.1) for the evaluation of the results, the best result was similar to that in (CRISCUOLO, 2017), obtained using Deep Learning techniques

Table 8 – Results with some of the used parameters. Decay and Threshold parameters were fixed because they displayed better results at earlier testing and did not interfere with the depth chosen. Best results are marked in bold.

Model	**Depth**	**Decay**	**Threshold**	**P@1**	**P@2**
11 Questions	1	0.3	0.05	0,4666	
11 Questions	2	0.3	0.05	**0,5333**	
27 Questions	1	0.3	0.05	0,4444	0,5185
27 Questions	2	0.3	0.05	0,4074	**0,6666**

Source: Created by author (2019).

in a similar corpus (in fact, the corpus used in this research is a compilation of some of the questions present in the corpus used by the author) - this result was obtained when matching 11 distinct questions and answers in a more enclosed context. After moving to the final set of 27 questions, the P@1 measure decreased to 40% clearly due to modeling mistakes (no previously correct answer was changed).

The developed prototype allows for three parameters to be tweaked, as mentioned in the previous section. The parameters are the activation value decay at each jump (or just decay), the activation threshold and the maximum depth. The best results for P@1 were obtained at 1 depth, 0.3 decay and 0.05 Threshold (this means that basically the activation only spreads to the level next to the original words - decay and threshold are almost unnecessary in this setup, only playing a role in skipping very weak answer relations). Table 8 is presented as a summary of the results achieved with the approach.

While finding the precise answer did present worse results than achieved by traditional methods, it was noticed that the precision of correct answers obtained with the approach jumped up to 66% when considering the first two most ranked answers. It was observed that for some questions the best (yet incorrect) answer had the same (or very close) activation value as the correct answer. This measure was called P@2 in the comparisons presented in Table 8.

Since precision values change with the parameters and correct answers can be found among the best raking ones, it can be argued that while the numeric results are not optimal, the approach can be improved to achieve better precision values. This can be done in several ways, which are going to be presented in chapter 4.

3.4.1 **Restrictions**

Another important information to look at are the restrictions presented with the approach. While the results can be improved by minor tweaks, some of the restrictions have to be dealt through deeper research. Before detailing Future works in the next chapter, it is necessary to point out and analyze these restrictions, trying to balance the

proposal benefits against its weaknesses.

One of the main problems of the proposed approach is related to an old paradigm in model based approaches: the need of human annotators to interpret the knowledge available as natural language in order to tailor a "computer understandable" structure. This time consuming task is doomed by most computer scientists today, who, in an immense sea of data, try to find in statistics the solution for this problem. The small model used for the 27 questions, for example, took two weeks of modeling to achieve the current results. On another hand the first modeling attempt mentioned in 3.3.1 took two months of a single annotator work to model about 4.000 relations.

Another restriction of the system is related to how it represents the concepts - as a pair of lemma and its POS. This representation causes a serious problem when dealing with polysemy. Polysemy happens when one word have several meanings (or represent distinct concepts). In the proposed model, since each concept is represented by a single (or composite) word, it cannot naturally represent ambiguous words. While this provides a natural transition between the NLP pipeline results and the Knowledge Graph, in a context where some words are polysemous depending on the context, a single node in the network could provide too many generic relations, causing unpredicted short-circuits in the spreading activation phase. In the modeled set of concepts, however, no word used was clearly polysemous inside the context.

As briefly mentioned in the previous section, one weak point of the proposed solution is also related to its time complexity when compared to Machine Learning Algorithms. While ML and especially Deep Learning techniques have a heavy time complexity during the training phase, the prediction (solution) phase is way less complex (sometimes just a linear solution). The proposed method, on another hand, as a graph search algorithm, can be more complex in the solution phase[12], depending on the input and the parameters used. Its also important to mention the increased space complexity inherent to graph traversal techniques.

One last shortcoming to be mentioned is the one related to the possibility of the algorithm not returning any answer for a given input. The reasons can be related to the absence of question concepts in the graph (to act as starting point of the spread-propagation pulses). This can also happen when the distance from the question concepts to the "closest" answer in the graph is further away than the maximum search depth or is constrained by the Decay and Threshold parameters. This was experienced during test phases and required further modeling and parameter tweaking to be circumvented, but more formalization is needed to understand how the concept modeling process has to be driven in order to prevent such situations.

[12] Breadth-first search, for example, in the worse case scenario can be of quadratic complexity.

4 FINAL REMARKS AND FUTURE WORKS

This chapter presents the final remarks for this research. It presents a summary of the advancements and solutions provided, as well as the analysis of future works intended to refine the proposed approach and circumvent its restrictions.

The main product of this work is the knowledge produced from the process of modeling a book into a knowledge representation graph for Question Answering Systems. More than a solution to a specific problem, the proposed approach is a step forward in the reuse of human structured knowledge presented in natural language (such as books and manuals) to build knowledge bases that are computer understandable. Even though much of the work was done through manual annotation, it was clearly perceived that a big part of this time consuming process can be simplified through heuristics and the use of some state-of-the-art NLU mechanisms and proposals such as FrameNet.

The positive results prove that the technique and structure described in the approach can be used for Question Answering Systems, as well as to tackle the question answer gap problem. In comparison to the solution published in (CRISCUOLO, 2017), the developed solution is less complex and easier to manipulate due to the nature of the underlying reasoning structure. Therefore, with some improvements it can become the core of user applications such as Chatbots and answering online education questions.

Another important contribution is the formal analysis of the Question Answer gap. While this problem was clearly perceived and attempted to be circumvented by earlier works, none of them formally described the gap or attempted to analyze its reasons or effects. While the formalization proposed is far from being considered a research on cognitive linguistics theory, it can offer a starting baseline for future research on the subject.

Finally, the Knowledge Graph and the algorithms developed can be used as a basis for a publicly available Dairy Farming Chatbot that can help dairy farmers in Brazil to find the answers for their questions. However, the current accuracy achieved by the technique is far from being considered a publishable solution, reason why some possible future works have to be done in order to improve its quality.

Most future works related to this research are improvements meant to achieve better results. One starting point would be to review the entire graph structure using graph visualization tools available in most programming languages to follow closely every pulse done for each question that wasn't correctly matched to its response. With this, missing relations and concepts could be spotted and corrected, increasing the solution accuracy. This process would also provide data for a better formalization of the modeling process. Another way to improve the model is by imposing restrictions to the annotating process, forcing more formalized models - a study on some of such techniques can be found

in (CHANG et al., 2015).

Another future implementation is to change relation weight based on relation type. This would cause some relations to either increase or decrease propagation strength. One starting step would be to give *SameAs* relationship a weight that kept the pulse strength, connecting all synonyms in a single pulse and helping with the relation distribution between them. This may require further studies on the impact over the selection system, since the graph would change as a whole. One interesting way to tweak on that would be to use optimization techniques to adjust the relationship weights, going a step further in the biomimeticism of that drives the proposed approach.

A third idea is to implement the fan-out constraints proposed in (GOUWS et al., 2010), which more than limiting the depth of the activation pulse, would provide ways to prevent pulses to spread equally to any and all neighboring nodes. This would be an alternative to the relationship type weight mentioned above, but would require more data or the use of external sources to be successful. Nevertheless, the use of the Inverse Link-Frequency constraint would be a natural evolution of the distributed weight for answer concepts and could improve the correct answer accuracy by pinpointing the most important concepts in an answer.

Following the line of automating the concept extraction process, further effort should be applied on the use of semi-structured publicly available data sources, such as Wikipedia. Several already well developed NLU techniques could be applied, helping in the detection of related concepts found in a concept's Wikipedia page. One example would be to use FrameNet to attempt to figure out Frame Elements among the page links and see if they relate to each other in according to some known "frame". Also, further development of Semantic Role Labeling would be of great benefit to this process.

Regarding the model restrictions, current state of model based approaches mostly rely on hand annotation. However, at automating part of the concept relationship retrieval process, as mentioned above, the effort spent on this activity can decrease heavily. Also, the development of an user interface especially designed for this kind of annotation could make the process less prone to error and more easy to visualize.

To treat the polysemy question, using a more complex network structure would be a resilient approach. In fact, a set of works on disambiguation are network based, such as the one proposed in (MATOS, 2014). The aforementioned work also relies on spreading activation for disambiguation, which makes it a natural refinement step to implement over the presented work. Adding an extra layer of semantic information could further enrich the system while keeping the answer selection mechanism simple.

Considering the problem of time and space complexity, this is still as a complete open question. Since the model was not used over a very large set of concepts, this

did not affect execution times significantly. However, as a technique that proposes to solve problems found in technologies that are every day more common, the effect of this complexity has to be reviewed in order to verify its applicability to large scale commercial products.

Finally, to address one last restriction of the proposed approach, further studies have to be done in order to establish metrics, best practices and formalism to Knowledge Graph building processes. These could help assess if the proposed graph is sufficient for the domain and application in question. Issues related to this process have been open for a while and are still far from being solved. Interestingly, some approaches to solve similar problems with Ontologies rely on Spreading Activation approaches and therefore could be a starting point for future researches (TSAOIR, 2014).

REFERENCES

ABEND, O.; RAPPOPORT, A. The State of the Art in Semantic Representation. In: **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, 2017. ISBN 9781482203516. ISSN 10417915.

AGICHTEIN, E.; CARMEL, D.; PELLEG, D.; PINTER, Y.; HARMAN, D. **Overview of the TREC 2016 LiveQA Track**, 2016. Available in: <https://trec.nist.gov/pubs/trec25/papers/Overview-QA.pdf>. Date accessed: 08.05.2019.

AMARAL, C.; FIGUEIRA, H. **Priberam's Question Answering System for Portuguese A Workbench for NLP**, 2006. 410–419 p.

ARSHAMIAN, A.; IANNILLI, E.; GERBER, J. C.; WILLANDER, J.; PERSSON, J.; SEO, H.-S.; HUMMEL, T.; LARSSON, M. The functional neuroanatomy of odor evoked autobiographical memories cued by odors and words. **Neuropsychologia**, v. 51, n. 1, p. 123–131, jan 2013. ISSN 00283932. Available in: <http://www.ncbi.nlm.nih.gov/pubmed/23147501 https://linkinghub.elsevier.com/retrieve/pii/S0028393212004617>. Date accessed: 08.05.2019.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. **Scientific American**, 2001.

BOUZIANE, A.; BOUCHIHA, D.; DOUMI, N.; MALKI, M. Question Answering Systems: Survey and Trends. **Procedia Computer Science**, Elsevier, v. 73, p. 366–375, jan 2015. ISSN 1877-0509. Available in: <https://www.sciencedirect.com/science/article/pii/S1877050915034663>. Date accessed: 08.05.2019.

CAMBRIA, E.; WHITE, B. Jumping NLP Curves: A Review of Natural Language Processing Research. **IEEE Computational Intelligence Magazine**, n. May, 2014. Available in: <http://www.krchowdhary.com/ai/ai14/lects/nlp-research-com-intlg-ieee.pdf>. Date accessed: 08.05.2019.

CHANG, N.; PARITOSH, P.; HUYNH, D.; BAKER, C. Scaling Semantic Frame Annotation. p. 1–10, 2015.

CIMIANO, P.; PAULHEIM, H. **Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods**, 2016. v. 0, 1–1 p. Available in: <http://www.geonames.org/>. Date accessed: 08.05.2019.

CRISCUOLO, M. **SlimRank: um modelo de seleção de respostas para perguntas de consumidores**. 143 p. Tese (Doutorado em Ciências de Computação e Matemática Computacional) — Universidade de São Paulo, São Carlos, 2017. Available in: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-31012018-140412/pt-br.php>. Date accessed: 08.05.2019.

CRUZ, J. C.; MAGALHãES, P. C.; FILHO, I. A. P.; MOREIRA, J. A. A. **Gado de leite: O produtor pergunta, a Embrapa responde.** 3. ed., 2011. 311 p. ISSN 1098-6596. ISBN 9788578110796.

DOZAT, T.; MANNING, C. D. **DEEP BIAFFINE ATTENTION FOR NEURAL DEPENDENCY PARSING**, 2017. Available in: <https://arxiv.org/pdf/1611.01734.pdf>. Date accessed: 08.05.2019.

EHRLINGER, L.; WÖSS, W. Towards a Definition of Knowledge Graphs. In: **SEMANTICS**, 2016. Available in: <http://www.semantic-web-journal.net/content/>. Date accessed: 08.05.2019.

FONSECA, E. R.; LUÍS, J.; ROSA, G. Mac-Morpho Revisited: Towards Robust Part-of-Speech Tagging. In: **Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology**, 2013. p. 98–107. Available in: <http://opennlp.apache.org/>. Date accessed: 08.05.2019.

FREITAS, C.; ROCHA, P.; BICK, E. **Um mundo novo na Floresta Sintá(c)tica-o treebank do Português**, 2008. v. 6, n. 3, 142–148 p. Available in: <http://visl.sdu.dk/>. Date accessed: 08.05.2019.

GOUWS, S.; ROOYEN, G.-J. van; ENGELBRECHT, H. A. Measuring Conceptual Similarity by Spreading Activation over Wikipedia's Hyperlink Structure. **Proceedings of the 2nd Workshop on Collaboratively Constructed Semantic Resources, Coling 2010**, n. August, p. 46–54, 2010. Available in: <https://www.aclweb.org/anthology/W10-3506>. Date accessed: 08.05.2019.

GRUBER, T. R. A Translation Approach to Portable Ontology Specifications. **Appeared in Knowledge Acquisition**, v. 5, n. 2, p. 199–220, 1993.

JARKE, M.; NEUMANN, B.; VASSILIOU, Y.; WAHLSTER, W. KBMS Requirements of Knowledge-Based Systems. In: SCHMIDT, J. W.; THANOS, C. (Ed.). **Foundations of Knowledge Base Management**, 1989. cap. 17, p. 381–394. Available in: <http://www.springerlink.com/index/10.1007/978-3-642-83397-7_17>. Date accessed: 08.05.2019.

JONES, K. S. **Natural Language Processing: A Historical Review**, 1994. Available in: <https://pdfs.semanticscholar.org/7445/69e1ff6377ba0b7e3a8e2bcd88ac94d9a02d.pdf>. Date accessed: 08.05.2019.

JURAFSKY, D.; MARTIN, J. **Speech and Language Processing**, 2014. 441–458 p. ISBN 0130950696.

KAZEMINEJAD, G.; BONIAL, C.; BROWN, S. W.; PALMER, M. **Automatically Extracting Qualia Relations for the Rich Event Ontology**, 2018. 2644–2652 p. Available in: <https://www.aclweb.org/anthology/C18-1224>. Date accessed: 08.05.2019.

KOLOMIYETS, O.; MOENS, M.-F. A survey on question answering technology from an information retrieval perspective. **Information Sciences**, Elsevier, v. 181, n. 24, p. 5412–5434, dec 2011. ISSN 0020-0255. Available in: <https://www.sciencedirect.com/science/article/pii/S0020025511003860>. Date accessed: 08.05.2019.

LENCI, A.; BEL, N.; BUSA, F.; CALZOLARI, N.; GOLA, E. SIMPLE : A general framework for the development of multilingual Lexicons . SIMPLE : A General Framework for the Development of Multilingual Lexicons. **International Journal of Lexicography**, v. 13, n. May 2014, p. 249–263, 2000.

MADABUSHI, H. T.; LEE, M.; BARNDEN, J. Integrating Question Classification and Deep Learning for improved Answer Selection. In: **Proceedings of the 27th International Conference on Computational Linguistics**, 2018. p. 3283–3294. Available in: <www.harishmadabushi.com/research/questionclassification/>. Date accessed: 08.05.2019.

MAGALDI, H.; BRAGA, R.; ARBEX, W.; CAMPOS, M. M.; BORGES, C. C. H.; DAVID, J. M. N.; CAMPOS, F.; STROELE, V. Análise de eficiência alimentar de gado leiteiro a partir da integração de bases heterogêneas e ontologias. In: **Anais do XII Brazilian e-Science Workshop**, 2018. Available in: <http://portaldeconteudo.sbc.org.br/index.php/bresci/article/view/3267>. Date accessed: 08.05.2019.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to Information Retrieval**, 2008. 569 p. ISBN 0521865719. Available in: <https://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf>. Date accessed: 08.05.2019.

MATOS, E. E. d. S. **CelOWS: Um Framework Baseado em Ontologias como Serviços Web para Modelagem Conceitual em Biologia Sistêmica**. 134 p. Dissertao (Mestrado em Modelagem Computacional) — Universidade Federal de Juiz de Fora, 2008. Available in:

<http://bdtd.ibict.br/vufind/Record/UFJF_f02737763b39a08becffb545beb87f4e#details>. Date accessed: 08.05.2019.

MATOS, E. E. d. S. **LUDI: Um framework para desambiguação lexical com base no enriquecimento da Semântica de Frames**. 202 p. Tese (Doutorado em Linguística) — Universidade Federal de Juiz de Fora, 2014. Available in: <https://repositorio.ufjf.br/jspui/handle/ufjf/695>. Date accessed: 08.05.2019.

MCCULLOCH, W. S.; PITTS, W. H. A Logical Calculus of the Ideas Immanent in Nervous Activity. **Bulletin of Mathematical Biophysics**, v. 5, p. 115–133, 1943. Available in: <http://www.cse.chalmers.se/ coquand/AUTOMATA/mcp.pdf>. Date accessed: 08.05.2019.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Distributed Representations of Words and Phrases and their Compositionality. In: **Proceedings of the 26th International Conference on Neural Information Processing Systems**, 2013. Available in: <https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf>. Date accessed: 08.05.2019.

MISHRA, A.; JAIN, S. K. A survey on question answering systems with classification. **Journal of King Saud University - Computer and Information Sciences**, Elsevier, v. 28, n. 3, p. 345–361, jul 2016. ISSN 1319-1578. Available in: <https://www.sciencedirect.com/science/article/pii/S1319157815000890?via%3Dihub>. Date accessed: 08.05.2019.

MOHAMED, A.; ALLAM, N.; HAGGAG, M. H. The Question Answering Systems : A Survey. **International Journal of Research and Reviews in Information Sciences (IJRRIS)**, v. 2, n. 3, p. 2046–6439, 2012.

OU, S. Y.; PEKAR, V.; ORASAN, C.; SPURK, C.; NEGRI, M.; European Language Resources, A. Development and Alignment of a Domain-Specific Ontology for Question Answering. **Sixth International Conference on Language Resources and Evaluation, Lrec 2008**, p. 2221–2228, 2008.

PUSTEJOVSKY, J. The generative lexicon. **Computational Linguistics**, MIT Press, Cambridge, MA, USA, v. 17, n. 4, p. 409–441, dec. 1991. ISSN 0891-2017. Available in: <http://dl.acm.org/citation.cfm?id=176321.176324>. Date accessed: 08.05.2019.

RODRIGUES, R.; OLIVEIRA, H. G.; GOMES, P. NLPPort: A Pipeline for Portuguese NLP. In: **7th Symposium on Languages, Applications and Technologies (SLATE 2018)**, 2018. p. 18:1–18:9. Available in: <http://opennlp.apache.org/>. Date accessed: 08.05.2019.

SINHA, S. **Answering Questions about Complex Events**, 2008. 207 p. Available in: <http://www.eecs.berkeley.edu/Pubs/TechRpts/2008/EECS-2008-175.html>. Date accessed: 08.05.2019.

STRUBELL, E.; VERGA, P.; BELANGER, D.; MCCALLUM, A. **Fast and Accurate Entity Recognition with Iterated Dilated Convolutions**, 2017. Available in: <https://github.com/iesl/>. Date accessed: 08.05.2019.

THANAKI, J. **Python Natural Language Processing**, 2017. 476 p. ISBN 9781787121423. Available in: <http://nemertes.lis.upatras.gr/jspui/bitstream/10889/5243/1/Σταυλιτη.pdf>. Date accessed: 08.05.2019.

TSAOIR, R. M. an. Using Spreading Activation to Evaluate and Improve Ontologies. In: **Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers**, 2014. p. 2237–2248. ISBN 9781941643266. Available in: <http://bioportal.bioontology.org/ontologies/SNOMEDCT>. Date accessed: 08.05.2019.

TURING, A. M. Computing Machinery and Intelligence. **Mind**, p. 433–460, 1959. Available in: <https://linkinghub.elsevier.com/retrieve/pii/B978012386980750023X>. Date accessed: 08.05.2019.

VERHOOSEL, J. P.; SPEK, J. Applying ontologies in the dairy farming domain for big data analysis. In: **CEUR Workshop Proceedings**, 2016. ISSN 16130073.

VINCENT, J. F. V.; BOGATYREVA, O. A.; BOGATYREV, N. R.; BOWYER, A.; PAHL, A.-K. Biomimetics: its practice and theory. **Journal of the Royal Society, Interface**, The Royal Society, v. 3, n. 9, p. 471–82, aug 2006. ISSN 1742-5689. Available in: <http://www.ncbi.nlm.nih.gov/pubmed/16849244 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1664643>. Date accessed: 08.05.2019.

WASKAN, J. Concepts. In: FIESER, J.; DOWDEN, B. H. (Ed.). **The internet encyclopedia of philosophy**. ISBN 2161-0002. Available in: <https://www.iep.utm.edu/concepts/>. Date accessed: 08.05.2019.

WASKAN, J. Connectionism. In: FIESER, J.; DOWDEN, B. H. (Ed.). **The internet encyclopedia of philosophy**. ISBN 2161-0002. Available in: <https://www.iep.utm.edu/connect/>. Date accessed: 08.05.2019.

WILKENS, R.; VILLAVICENCIO, A.; MULLER, D.; WIVES, L.; SILVA, F. da; LOH, S. **COMUNICA-A Question Answering System for Brazilian Portuguese**, 2010. 21–24 p. Available in: <http://www.ibge.gov.br/ibgeteen/>. Date accessed: 08.05.2019.

YOUNG, T.; HAZARIKA, D.; PORIA, S.; CAMBRIA, E. **Recent Trends in Deep Learning Based Natural Language Processing**, 2018. Available in: <http://veredshwartz.blogspot.sg.>. Date accessed: 08.05.2019.

ZIPF, G. K. **Human Behaviour and the Principle of Least Effort**, 1949.

APPENDIX A – Questions and Answers Used

Below are the questions and answers chosen to be modeled in the system created to test the proposed approach. They were not translated for the sake of simplicity.

11. Qual a vantagem de descornar os bezerros ainda jovens?

A descorna do animal ainda jovem é mais fácil e segura de ser feita, facilita o manejo do bezerro e dá maior segurança no trato com os animais adultos. É uma prática relativamente fácil de ser realizada, e evita acidentes decorrentes de brigas entre animais na fase adulta.

64. Há vantagem em se adicionar água ao volumoso e ao concentrado para vacas leiteiras?

Não. Ainda não foi identificada nenhuma vantagem na mistura de água com o alimento sólido, seja concentrado ou volumoso. É preciso considerar que isso pode aumentar o custo com a mão de obra, ou mesmo, a perda do alimento concentrado. Assim, o uso do popular "sopão" não é recomendado.

93. Até que idade um reprodutor pode ser utilizado?

Não existe uma idade limite, desde que o reprodutor se mostre sadio, com libido e boa produção de espermatozoides.

142. O que são plantas estoloníferas?

São plantas com hábito de crescimento rasteiro (decumbente ou prostrada), que se multiplicam por meio de estolões, ou caules (ramas), e se fixam ao solo pelas raízes que se formam em seus nós. São plantas que proporcionam boa cobertura do solo, ao contrário das de crescimento ereto (cespitosas), que formam touceiras. Os capins estrela-africana e angola são exemplos de gramíneas forrageiras estoloníferas.

299. O que é exame andrológico? O que ele indica?

Exame andrológico é a avaliação da capacidade do touro de cobrir e emprenhar uma vaca. Esse exame indica o potencial do touro de emprenhar, no momento do exame, e não por sua vida inteira, pois a fertilidade é fortemente dependente das condições em que o animal vive, e que podem mudar de uma propriedade para outra, e de mês a mês.

320. Quais as principais causas de abortos?

Abortos podem ser provocados por diversos fatores, como estresse calórico e de qualquer outro tipo, transporte, ingestão de plantas tóxicas, aplicações de hormônios, tumores, defeitos genéticos, etc. Contudo, as causas infecciosas de abortos são as mais importantes. Entre as doenças infecciosas mais comuns estão a brucelose, leptospirose, campilobacteriose, tricomonose, diarreia viral bovina (BVD) e rinotraqueíte infecciosa

bovina (IBR). Doenças como tuberculose, salmonelose, listeriose e micoplasmose, e outras não específicas, como mamites, que provoquem um processo febril grave, também podem causar abortos.

361. O grau de sangue, ou composição genética, tem influência na fertilidade do animal?

Sim. Sob condições desfavoráveis de manejo e de temperatura, animais mais azebuados apresentam melhor desempenho reprodutivo quando comparados àqueles com maior percentagem de raças europeias. Mas em boas condições de manejo, alimentação e clima, animais mestiços com predominância de raças europeias apresentam índices reprodutivos que podem superar os mestiços com predominância de raças zebuínas em sua composição. Esse aspecto está relacionado à contribuição dada pelos zebuínos à rusticidade ou adaptação dos animais mestiços aos ambientes adversos.

400. O que é eimeriose (ou coccidiose)?

É uma doença causada por um protozoário denominado *Eimeria spp.*, que acomete o intestino dos bezerros. Seu principal sintoma é diarreia, que pode ter sangue. Apesar de ser uma doença de animais jovens, também pode atingir adultos. A higienização bem feita das instalações e a redução da aglomeração de animais são as principais maneiras de se reduzir a doença, tendo sido de grande ajuda nesse aspecto as casinhas móveis para bezerros. Os piquetes de acesso aos bezerros lactentes devem ser formados com pastagens apropriadas, de folhas finas, e mantidas baixas, para que os raios de sol ajudem a mantê-las menos propícias à contaminação com Eimeria spp., além de outros parasitas que acometem os bezerros. O tratamento da eimeriose, ou coccidiose, é feito com produtos à base de sulfas. E, considerando que a doença é aguda, podendo ter mortalidade elevada, e que seus sintomas podem ser confundidos com os de outras doenças, recomenda-se a orientação de um médico veterinário.

401. Quais as principais moscas causadoras de prejuízos econômicos no meio rural?

São as moscas do berne, da bicheira, doméstica, dos estábulos e dos chifres. Os prejuízos são determinados, dependendo da espécie, pela retirada de sangue e estresse dos animais em virtude de picada, transmissão de agentes causadores de doenças e depreciação dos couros.

402. Como realizar o controle das moscas do meio rural?

Higiene é a palavra-chave quando o assunto é controle de moscas, sobretudo em relação à limpeza das instalações e à destinação adequada dos dejetos de fezes dos animais. O tratamento dos animais com mosquicidas deve ser realizado preventivamente no início da época das chuvas, uma vez que ambientes quentes e úmidos são propícios à proliferação de moscas das mais diversas espécies. A aplicação de brincos impregnados com

substâncias mosquicidas também é uma boa opção, mas devem ser retirados de acordo com o período recomendado pela bula, a fim de se evitar a proliferação de moscas resistentes, em consequência do contato com o veneno enfraquecido pelo tempo. Existem alguns tipos de armadilhas que capturam e eliminam moscas adultas. Para implementação de tais armadilhas, recomenda-se que sejam buscadas orientações no órgão estadual de assistência técnica e extensão rural mais próximo. Para obter êxito, é importante que o controle seja realizado de forma adequada e, ao mesmo tempo, na maior quantidade possível de propriedades da região, o que pode ser facilitado pela estimulação da população por meio de campanhas de combate às moscas.

403. Qual a diferença entre bicheira e berne?

A bicheira, ou miíase, é caracterizada pelo desenvolvimento de larvas de mosca da espécie Cochliomyia hominivorax em diversos tecidos do organismo animal. Para que a mosca adulta ponha os ovos e instale a bicheira, é necessário que haja uma "porta de entrada", que pode ser um ferimento ou umbigo de animal recém-nascido. Por isso, é importante a aplicação de medicamentos cicatrizantes e repelentes nesses locais. Em cada local de instalação, desenvolvem-se centenas de larvas, com alta capacidade de penetrar pelos tecidos (principalmente, músculos e cartilagens) durante 7 a 10 dias. O berne, outro tipo de miíase, é a larva da mosca Dermatobia hominis. Em cada nódulo, há apenas uma larva, que se desenvolve no tecido subcutâneo do animal por aproximadamente 40 dias. Não é necessário lesão prévia: as larvas penetram pelo tecido íntegro. Uma particularidade interessante consiste no fato de que não é a mosca do berne que vai ao animal para fazer a postura. Após a cópula, a mosca do berne captura outro inseto (geralmente uma mosca de outra espécie) e o utiliza como vetor, depositando os ovos em seu abdômen. Após o desenvolvimento dos ovos, quando o inseto vetor pousa em um bovino, a temperatura corporal do animal provoca a eclosão das larvas, que penetram ativamente pelo couro. Justamente por envolver a participação de outras espécies, o controle do berne é complexo e deve ser direcionado também ao combate de outras espécies de moscas da região, para que se obtenha êxito.

404. O que é cisticercose bovina?

É uma doença parasitária dos bovinos, causada pela fase larval do cestoide Taenia saginata, chamada Cysticercus bovis. Taenia saginata é um parasita do homem, que acomete bovinos quando ingerem pastagem contaminada com fezes humanas, contendo ovos do cestoide. Os bovinos, então, são considerados hospedeiros intermediários do parasita. O cisto se aloja nos músculos dos bovinos, tendo preferência por coração, língua e diafragma. A ingestão de carne contaminada leva ao desenvolvimento do verme adulto na espécie humana, conhecido popularmente como "solitária". Em casos de animais confinados, recomenda-se que os empregados sejam examinados e tratados periodicamente, para prevenir a contaminação de humanos e animais. Sugere-se, também, adotar normas

de higiene e ter dependências sanitárias adequadas para os empregados da fazenda.

405. O homem pode adquirir a cisticercose?

Sim. Mas não pela ingestão de carne contaminada. A instalação da doença ocorre pela ingestão acidental de ovos do cestoide, eliminados nas fezes de uma pessoa que apresente a tênia adulta em seu intestino. Essa situação pode ocorrer em ambientes sem higiene ou a partir de atos promíscuos. Por isso, é importante o estabelecimento de programas de educação sanitária.

406. O que é salmonelose? Como preveni-la?

Também conhecida como paratifo dos bezerros, é uma doença infecciosa causada por bactérias do gênero Salmonella. Bezerros até os 3 meses de idade são mais suscetíveis, mas animais em outras faixas etárias também podem ser acometidos, principalmente, quando se encontram debilitados. A transmissão ocorre pela ingestão de água ou alimentos contaminados ou pelo contato com fezes de animais doentes ou portadores do agente da doença. Os principais sintomas são febre alta, diarreia aquosa e intensa, dor no abdômen, prostração e morte. Para a prevenção da doença, devem-se manter as instalações sempre limpas, secas e desinfetadas, isolar os animais doentes e evitar o acesso dos sadios a pastos contaminados.

407. O que é colibacilose? Quais as medidas para tratamento e prevenção?

É uma doença infecciosa, causada pela bactéria Escherichia coli, afeta bezerros jovens, mas é rara em adultos. A não ingestão do colostro, a aglomeração e a manutenção dos animais sem a adequada higiene são fatores que favorecem o estabelecimento da infecção. Dependendo do local de instalação da bactéria, os sintomas podem variar, os mais frequentes são febre, falta de apetite, fraqueza e diarreia. Em casos mais graves, o animal pode entrar em coma, apresentando temperaturas baixas e mucosas pálidas. O tratamento é feito com antibiótico e soroterapia. Mais importante que o tratamento é a prevenção, que consiste em alimentar bem os animais, evitar aglomerações, utilizar instalações adequadas, limpas e secas, impedindo animais de idades diferentes no mesmo lote.

408. O que é pneumoenterite? Quais os sintomas, como preveni-la e tratá-la?

É uma infecção causada inicialmente por vírus, normalmente acompanhada por invasão bacteriana. Geralmente, ataca bezerros até os 2 meses de idade, atingindo os aparelhos respiratório e digestivo. É mais frequente em animais criados em bezerreiros úmidos e sem higiene. O animal doente apresenta febre alta, respiração acelerada e diarreia. O tratamento deve ser feito rapidamente para que a doença não se torne crônica. Evita-se a infecção mantendo os animais sempre bem alimentados, em instalações secas e limpas, e

evitando aglomerações. O tratamento deve ser feito com antibióticos, sempre prescritos por um médico veterinário.

409. Qual a causa da vaca urinar sangue? Qual o tratamento?

Quando a vaca está urinando sangue, suspeita-se inicialmente de três causas: 1) ingestão de planta tóxica, por exemplo, samambaia; 2) tristeza parasitária bovina; 3) Braquiária Tanner Grass. Quando o animal apresenta os sintomas e no pasto em que ele se encontra existe a samambaia (cujo nome científico é Pteridium aquilinum), a primeira suspeita é essa patologia. Por efeito da samambaia, desenvolve-se inicialmente uma irritação na mucosa da bexiga e logo há o desenvolvimento de neoplasia (câncer), mas não há tratamento eficaz. Existem vários princípios ativos na samambaia que afetam os animais. Para os bovinos, são substâncias cancerígenas (uma das principais é o norsesquiterpeno ptaquilosido), que produzem efeitos semelhantes à radiação no organismo animal. O princípio tóxico da samambaia vai se acumulando no organismo do animal até chegar a ponto de causar a doença. Por isso, em uma propriedade, existem animais com problemas e outros que não apresentam os sintomas, todos no mesmo pasto. Em geral, animais nascidos em fazenda que tem samambaia não a ingerem, por um aprendizado ainda desconhecido. Assim, é mais comum ter problema com animais oriundos de outras propriedades, comprados ou transferidos de outra fazenda que não tinha samambaia. Também é muito comum a propriedade do vizinho ter samambaia e não apresentar o problema, uma vez que os animais podem não estar pastando a samambaia porque "aprenderam" que é tóxica ou porque a pastagem está boa, com maior oferta de volumoso de qualidade. Então, o produtor se pergunta: porque só ocorre na minha fazenda? Para responder a essa pergunta é preciso analisar as condições encontradas, verificar o que está ocorrendo, a disponibilidade e a qualidade do pasto, a origem do gado. Isso mostra que a assistência técnica de um médico veterinário é fundamental nesses casos. Não há tratamento terapêutico eficaz para bovinos. Pode-se tentar transfusão de sangue e antibioticoterapia, visando conter as infecções secundárias. Uma boa medida pode ser o descarte do animal para corte. Outro ponto importante, se o problema for a samambaia, é adotar práticas agrícolas para eliminar essa planta, que ocorre mais em solos ácidos. O mais indicado é fazer a análise de solo e depois a calagem. E, na época própria, fazer o plantio de uma lavoura (milho, feijão) na área, por uns 2 anos seguidos, para eliminar a samambaia. Depois, pode-se formar pasto novamente. Também, deve-se evitar que os bovinos tenham acesso ao terreno infestado com samambaia, providenciando-se uma cerca. Se a propriedade não tem samambaia, a suspeita pode ser Tristeza Parasitária Bovina (TPB) ou pastagem de brachiaria Tanner grass. ATPB é uma das duas doenças cujos agentes causadores (Babesia spp., Anaplasma marginale) podem ser transmitidos por carrapato. Na babesiose, a urina pode tomar cor que varia desde vermelho até marrom-escuro. ATPB tem que ser tratada sob pena de morte do animal. Quando o tratamento for realizado em tempo hábil, a recuperação é relativamente rápida. Ainda

existe a possibilidade da causa ser as pastagens de braquiária. Quando o animal está pastando braquiária da espécie Tanner Grass, pode ocorrer a eliminação de urina com sangue, e, nesse caso, é só retirar o animal daquele pasto e tudo volta ao normal. No entanto, qualquer outra afecção que estiver instalada nas vias urinárias pode levar a uma hematúria, ou urina avermelhada. Faz-se, então, necessária a presença de um médico veterinário para que o diagnóstico, e a indicação do tratamento e da dosagem sejam realizados.

410. O que fazer em caso de bezerro com diarreia?

A diarreia pode ser causada por diversos fatores, como verminose, infecção por bactéria ou protozoário, alteração na alimentação, estresse por mudança de ambiente ou excesso de animais, entre outros. A consequência mais grave é a morte do animal por desidratação. Por essa razão, um bovino com diarreia deve ser imediatamente transferido para um ambiente limpo, seco e arejado, e receber soro. Se o processo estiver no início, pode ser administrado soro caseiro: 5 Lde água de boa qualidade, 250 g de açúcar, 45 g de sal e uma colher de sopa de bicarbonato de sódio. Um bezerro precisa receber de 5 L a 7 L de soro, por dia, distribuídos em 5 a 10 administrações por via oral. Esse procedimento reidrata, mas é imprescindível a intervenção de um médico veterinário, que prescreverá o tratamento.

411. O que é brucelose?

É uma doença infectocontagiosa, causada por bactéria do gênero Brucella e caracterizada por distúrbios de fertilidade nos machos e fêmeas. O diagnóstico deve ser feito por exame laboratorial específico, realizado pelo menos uma vez ao ano. Para a prevenção, devem ser vacinadas e marcadas as bezerras, entre o 3º e o 8º mês de idade, com a vacina B-19. Deve-se adquirir somente animais com resultado negativo para o teste, mantê-los isolados em quarentena antes de sua incorporação ao rebanho, e realizar novo teste após 30 dias. A ingestão de leite cru, proveniente de animal doente, e o contato com suas secreções corporais podem levar à instalação da doença no homem.

412. O que é manqueira?

O carbúnculo sintomático, também conhecido como manqueira, é uma doença provocada por bactéria do gênero Clostridium, mais frequente em animais jovens, principalmente aqueles com maior escore corporal. O agente causador encontra-se no solo e, ao ser ingerido, instala-se no organismo animal, determinando febre, falta de apetite, desânimo e manqueira. A manqueira só ocorre se a lesão atingir grandes massas musculares, como espádua, quartos e pescoço. O tratamento, mesmo intensivo, não surte efeito, e a doença, geralmente, é fatal. A vacinação dos animais jovens é o melhor meio para a prevenção da doença. Os bezerros devem ser vacinados aos 4meses de idade e receber uma dose de reforço após 30 dias. Deve-se revacinar a cada 6 meses, até os animais atingirem 24 meses

de idade.

413. Quais as medidas para a prevenção da raiva?

Deve-se vacinar os bezerros por volta do 4º mês de idade e repetir a dose 30 dias depois. Não se pode esquecer de revacinar anualmente todos os animais da fazenda. Uma pasta vampiricida deve ser aplicada na ferida deixada pelos morcegos e deve-se combater os hematófagos (que se alimentam de sangue). Para isso, é importante a atuação de técnicos especializados, que irão identificar as espécies de morcegos a serem controladas, evitando atingir espécies benéficas. As medidas de prevenção devem ser extensivas a outras espécies de animais domésticos, em virtude do caráter altamente contagioso da doença. O homem também pode ser atingido, devendo-se, portanto, evitar contato com secreções de animais supostamente doentes.

414. Os bovinos podem ser acometidos por tuberculose? Caso positivo, como evitar essa doença?

Sim. Normalmente, a doença é adquirida pelo contato direto ou indireto com secreções de animais infectados, mas, em alguns casos, o homem também pode ser a fonte de infecção. A evolução da doença é crônica e os sintomas são variados, devendo-se, portanto, realizar o teste de tuberculinização. A realização desse teste antes da compra, a aquisição de animais comprovadamente negativos para o agente da doença e sua manutenção em isolamento por 60dias, para realização de outro teste antes da incorporação ao rebanho, são as principais medidas de prevenção da doença. Uma observação importante: em todos os casos de doenças de animais, sugere-se consultar um médico veterinário da região. Ele deverá examinar o animal doente, fazer o diagnóstico, prescrever o tratamento, indicar a dosagem e o modo de usar os medicamentos. Ao usar qualquer medicamento, é muito importante ler atentamente a bula, o modo de aplicação, as indicações do fabricante, etc.

415. O que é mastite?

Mastite, ou mamite, é a inflamação da glândula mamária, desencadeada pela agressão da glândula por diferentes tipos de agentes, como microrganismos, irritantes químicos e traumas físicos. Na vaca leiteira, a mastite é quase sempre causada por bactérias que invadem o úbere, multiplicam-se, produzem toxinas e outras substâncias irritantes, que provocam a resposta inflamatória. É a doença mais comum e a que mais causa prejuízos aos rebanhos leiteiros.

457. O que é leite "verde"?

Entende-se por leite "verde", o leite produzido a pasto. Nesse caso, a alimentação básica dos animais é a pastagem, sem qualquer imposição normativa relacionada ao manejo dessas pastagens, à alimentação e ao uso de produtos químicos nos processos de produção e controle sanitário do rebanho, à semelhança das que são exigidas para a produção orgânica de leite.

466. Quais os itens mais onerosos no custo de produção de leite?

Os itens mais onerosos são geralmente relacionados à alimentação do rebanho, principalmente, a aquisição de concentrados que, em alguns casos, superam 50% do preço bruto do leite. Os gastos com a mão de obra, normalmente, são o segundo item de maior impor- tância econômica, e deveriam ser de aproximadamente 20%.

476. Qual o melhor tipo de sombra para bovinos?

Na maioria dos casos, a sombra mais eficiente para aliviar o estresse térmico dos bovinos provocado pelo calor é o sombreamento natural, com utilização de árvores. Durante o dia, ocorre o resfriamento do ambiente abaixo da copa da árvore pela interceptação da radiação solar direta, feita pela espessa massa de folhas da copa, e pelo resfriamento benéfico do ar, provocado pela evaporação da umidade das folhas (energia latente). Durante a noite, pelo metabolismo, há liberação de calor.

483. Qual a importância do tratamento e aproveitamento dos dejetos de bovinos nas propriedades?

Os prejuízos ambientais causados pela falta de tratamento e pelo manejo inadequado dos resíduos da produção animal são incalculáveis. Em muitos países, os efluentes oriundos da produção animal já são a principal fonte de poluição dos recursos hídricos, superando os índices das indústrias, consideradas, até então, as grandes causadoras da degradação ambiental. Esses resíduos orgânicos, ou dejetos animais, constituídos pelas fezes e urina, adequadamente manejados e reciclados no solo, deixam de ser poluentes e passam a constituir valiosos insumos para a produção agrícola sustentável. Produzir de forma sustentável implica reduzir ou evitar o consumo de recursos ou insumos externos. Dessa forma, deve-se trabalhar com o objetivo de que a importação de recursos seja equilibrada pela exportação. Uma das formas seria evitar desperdício de energia e de matéria-prima, aumentando a produtividade, e a competitividade do capital e da mão de obra. Tecnologias eficientes de tratamento e reciclagem de efluentes gerados pelas atividades agrícolas, por exemplo, os resíduos da produção animal, constituem importante ferramenta para aperfeiçoar a relação custo/benefício dos sistemas de produção.