Querying Brazilian Educational Open Data using a Hybrid NLP-based Approach

Marco Antoni[1][a], Andrea Schwertner Charão[1][b] and Maria Helena Franciscatto[2][c]

[1]*Department of Languages and Computer Systems, Federal University of Santa Maria, Santa Maria, Brazil*

[2]*Department of Computer Science, Federal University of Paraná, Curitiba, Brazil*

Keywords: Question Answering, Open Data, Educational Information Retrieval, Natural Language Processing.

Abstract: The need for capturing information suitable to the user has favored the development of Question Answering (QA) systems, whose main goal is retrieving a precise answer to a question expressed in Natural Language. Thus, these systems have been adopted in many domains to make data accessible, including Open Data. Although there are many QA approaches that access Open Data sources, querying Brazilian Open Data is still a research gap, possibly motivated by the complexity that Portuguese language presents to Natural Language Processing (NLP) approaches. For this reason, this paper proposes a hybrid NLP-based approach for querying Open Data of Brazilian Educational Census. The proposed solution is based on a combination of linguistic and rule-based NLP approaches, that are applied in two main processing stages (Text Preprocessing and Question Mapping) to identify the meaning of an input question and optimize the querying process. Our approach was evaluated through a QA prototype developed as a Web interface and showed feasible results, since concise and accurate answers were presented to the user.

1 INTRODUCTION

With the large amount of data being produced and stored every day, there has been an increasing search for new ways to organize data in order to extract information that is useful in several application domains. Information Retrieval approaches (IR) are a common way to capture information suitable for the user's needs, by analyzing, structuring, and accessing information from large data sets (Sy et al., 2012).

However, they present some challenges, e.g., they often retrieve information that are not relevant or cannot be easilyF processed (Utomo et al., 2017; Ferrández and Peral, 2010).

Question Answering (QA): systems have been considered to be able to minimize these problems, since they aim at providing a precise answer that satisfies a specific question (Bouziane et al., 2015; Ferrández and Peral, 2010). These systems are considered a specialized subarea of IR, aiming to satisfy the users needs for specific information, by means of questions expressed in Natural Language.

Retrieving the information the user expects requires accessing structured and unstructured data, which lose significance if they are not presented clearly and meaningfully (Beniwal et al., 2018). QA systems can be very valuable as a way to make data accessible and exploitable, thus they have been investigated in many domains where transparent information is needed, such as *Open Data* (Wendt et al., 2012).

In the Open Data domain – especially Open Government Data – it is desirable that public information can be analyzed and used. In other words, with Open Data processing, many problems of the citizens can be solved through different services, projects and activities that improve the government's economy, efficiency and transparency (Mouromtsev et al., 2013). With the evolution of linked open data sources and the challenge of analyzing extremely large datasets, several QA researches are conducted in this domain, aiming to extract meaningful information (Marx et al., 2014; Yao and Van Durme, 2014; Charton et al., 2016; de Castro et al., 2019).

Although there are many QA approaches in the literature that access and process Open Data, mining *Brazilian Open Data* is still barely investigated (Oliveira et al., 2016). One possible reason

[a] https://orcid.org/0000-0002-7881-5715
[b] https://orcid.org/0000-0003-3695-8547
[c] https://orcid.org/0000-0002-5054-196X

Antoni, M., Charão, A. and Franciscatto, M.
Querying Brazilian Educational Open Data using a Hybrid NLP-based Approach.
DOI: 10.5220/0010486201200130
In *Proceedings of the 23rd International Conference on Enterprise Information Systems (ICEIS 2021) - Volume 2*, pages 120-130
ISBN: 978-989-758-509-8

for that is that Natural Language Processing (NLP) tools are limited or underperform when dealing with Brazilian Portuguese, due to its particular and complex structure (Rodrigues et al., 2018; Rocha and Lopes Cardoso, 2018). An example of this complexity is shown in Table 1: whereas in English the word *Student* is applied for all genders, the equivalent word in Portuguese language is inflected according to the gender. In addition, Portuguese has many synonyms, which are less frequent in other languages such as English.

Therefore, the investigation of NLP techniques applied to Portuguese is highly valuable, as it can leverage the mining of Brazilian Open Data and the availability of useful data.

Table 1: Differences between the word "Student" in English and Portuguese.

English	Portuguese	Gender
Student	Aluno	Masculine
	Aluna	Feminine
	Estudante	Masculine/Feminine

Motivated by this research gap, this paper presents a hybrid NLP-based solution for querying Brazilian Open Data, specifically the Brazilian Educational Census. The Educational Census is the main tool for surveying the country education, as it gathers data from different stages and modalities of basic and professional education in Brazil (INEP, 2020). For querying Brazilian Open Data, our solution is based on a combination of two NLP approaches (linguistic and rule-based) that were applied in Text Preprocessing and Question Mapping stages for identifying the meaning of an input question.

Considering the lack of QA systems for accessing Brazilian Open Data, we also implemented a QA prototype as a Web interface for evaluating our hybrid solution. The prototype accesses specific tables of the Brazilian Educational Census related to schools, students, and courses, obtaining satisfactory answers. To the best of our knowledge, there is not yet a study for retrieving information from these data sets using a hybrid NLP approach, thus the present work can potentially contribute to Brazilian Open Data mining.

The structure of the paper is presented as follows. In Section 2, the main concepts related to this research are presented. In Section 3 we report our NLP-based approach for querying Brazilian Open Data, and in Section 4 we discuss the results and limitations found. Lastly, in Section 5, we present the conclusions and final remarks.

2 BACKGROUND AND RELATED WORK

2.1 Natural Language Processing

Natural Language (NL) is the way that humans use to communicate, and has been increasingly investigated to mediate interactions between them and computers (Henderson et al., 2017). With respect to search engines, queries in the form of questions have grown considerably (including voice-based questions), thus leveraging the development of QA systems capable of answer NL questions (White et al., 2015).

The importance of understanding NL brought up the Natural Language Processing (NLP), a research domain that focuses on the development of computational models that allow human-computer interaction through NL (Ruiz et al., 2020). Specifically, the NLP domain comprises a set of computational techniques for inspecting and representing texts that occur naturally at one or more levels of linguistic analysis, in order to achieve human-like language processing for a variety of tasks or applications (Liddy, 2001).

The related literature discusses several approaches that aim to find the meaning of a question or information posed by the user, i.e., the *rule-based approach* and the *linguistic approach*.

The rule-based approach considers patterns used in the question formulation to identify the information need, i.e., a word or sequence of words contained in the question will define its core, meaning or what it refers to (Gupta and Gupta, 2012). For this task, rules are created using domain knowledge, then the information is extracted from the input question by means of matching with the rules (Lee et al., 2019). In its turn, the linguistic approach involves methods of lexical analysis, syntactic analysis, exploiting also semantic features (Nesi et al., 2015). We can mention several linguistic methods such as:

- Spell Checking: spelling and typing errors, slangs, regional terms, and word abbreviations are common in human language. Spell checking identifies these errors and replace them with the best possible combination of correct words, in order to avoid unexpected results in the processing (Kaur et al., 2014).
- Named-Entity Recognition (NER): this technique involves recognizing and classifying entities in a sentence according to the class it belongs to. This classification can be split into three main groups: entities (i.e., people, organizations, and locations), temporal expressions (dates and times), as

well as quantities (numeric values, monetary values, and percentages) (Goyal et al., 2018).

- Stemming e Lemmatization: both techniques are related to morphological analysis, aiming to simplify the analysis of a word by reducing its variants. Stemming is used to remove derivational suffixes and inflections of a word, obtaining a common root (Balakrishnan and Lloyd-Yemoh, 2014), whereas Lemmatization tries to remove inflectional endings, returning the word in its dictionary form (Singh and Gupta, 2017). These techniques are very important in languages of complex morphology, e.g., Portuguese.
- Stopwords Removal: Stopwords are words that add little or no value to the text analysis, i.e., articles, pronouns, adverbs, prepositions, conjunctions and vowels (Khosrow-Pour and Khosrow-pour, 2009). They can be removed from the sentence without prejudice to the processing.

Linguistic and rule-based approaches are commonly applied in QA systems, for recognizing the information that the user really wants in an input question. Thus, by performing this recognition, the QA system is able to provide right answers to the user, being effective in several application domains. Question Answering systems will be discussed in the next section.

2.2 Question Answering Systems

Question Answering systems aim to retrieve concise answers to the user for meeting a specific information need expressed in Natural Language (Utomo et al., 2017). These systems differ from common search engines, which deliver a set of results to the user rather than an accurate answer, in a way the user must find the most appropriate answer by himself. The main characteristic of QA, therefore, is the ability to return compact, comprehensible and correct answers, which may refer to a word, sentence, paragraph, or an entire document (Kolomiyets and Moens, 2011).

Question Answering domain has grown steadily, and in the present-day it is integrated in many personal assistants like Siri, Cortana, Alexa, and Google Assistant (Roy and Anand, 2020). Recent advances in NLP and Artificial Intelligence have powered these systems with more comprehensive interactions and effective analysis, driving QA-based proposals in many areas (Yin et al., 2019). E.g., the study in (Masseroli et al., 2014) presents an approach to support integrated search of distributed biomedical-molecular data, aimed at answering multi-topic complex biomedical questions. The QA system in (Mani et al., 2018) provides IT support to users' questions by exploring multimedia data, also allowing an automation to be attached to an answer. QA is also used in educational contexts, as an alternative to discussion forums for mediating online discussions (Srba et al., 2019).

In our work, we constructed a prototype of Question Answering system implemented as a Web interface, for querying Brazilian Open Data sets using both rule-based and linguistic NLP approaches (i.e, the computational techniques mentioned in Section 2.1). Further information about Open Data will be given as follows.

2.3 Open Data

Open Data is strategy that encourages mostly public organizations to release factual and nonperson-specific data (generated through the delivery of public services) to anyone, allowing the free use of these data without any copyright restrictions (Hossain et al., 2016; Murray-Rust et al., 2010). In addition, Open Data allow interdisciplinary scientific cooperation, helping researchers to make decisions based on data, integrate and analyze heterogeneous data sources, as well as solve complex problems in several areas (Sowe and Zettsu, 2015).

Sharing Open Data with the public has led to the need to maintain numerous databases for storing important information, making the manipulation and processing of this data a non-trivial task (Zhang and Yue, 2016). Considering the importance of the subject, many studies have investigated the manipulation of Open Data to extract relevant information from large datasets. The work described in (Sowe and Zettsu, 2015) presents an Open Data development model that addresses limitations in the extraction and integration of data from different sources. The model demonstrates possibilities of using and reusing data, inserting functionalities and transforming data for building a new generation of open data repositories. With respect to Open Data and Question Answering, the study in (Molina Gallego, 2018) applies NLP techniques in a Conversational Agent to recognize conversation topics, identify data sets within Open Data based on these topics, and present the information to the user.

Despite the existing proposals, mining Brazilian Open Data is still little investigated, possibly motivated by the format in which these data are made available, the incompleteness of information, as well as particularities of the Portuguese language for NLP processing (Oliveira et al., 2016; dos Santos Brito et al., 2015). Thus, in the next section we present our linguistic and rule-based approach for querying

Brazilian Open Data.

3 HYBRID NLP-BASED APPROACH FOR QUERYING BRAZILIAN OPEN DATA

3.1 Data Characterization

Our NLP-based solution for querying Brazilian Open Data focuses on the *Brazilian Educational Census*[1], which surveys education in Brazil by gathering data from different modalities of basic and professional education, covering data about teachers, schools, classes, and education administrators. For exploration purposes, we used a reduced set of data contained in the 2019 Census, thus applying a collection of evaluation questions on the data tables **Students, Schools** and **Courses** (details about the questions are given in subsection 3.2.2).

The **Schools** table has approximately 230 thousand records, and allows to access information related to identification, location, infrastructure and places offering in Brazilian schools. The **Students** table allows to access information on students identification and their educational context (e.g., if the student has special needs, which means of transportation she/he uses, and which modalities of education she/he is linked to). The data in this table are divided according to the region of the country, so we used the set referring to students from the state *Rio Grande do Sul*, which has approximately 7 million records. Finally, the **Courses** table is an auxiliary table[2] that contains information about all technical courses present in the National Catalog of Technical Courses (i.e., 240 records). The relation of this table with the Students table is optional, as not every student is enrolled in a technical course.

Although the data used in our approach are freely available, they do not follow all W3C (World Wide Web Consortium) recommended guidelines[3]: they are not linked nor are in RDF format, which makes it difficult to use a standard query language (Isotani and Bittencourt, 2015). Despite that, data are available in CSV and tabular format, thus they can be stored in relational databases and accessed by our QA-based prototype, by means of NL questions.

3.2 Design and Implementation

Many studies have focused on translating NL questions to SQL syntax (Giordani and Moschitti, 2012), however, approaches of this nature demand knowledge about the data dictionary and the metadata involved in the dataset. Thus, we apply a combination of *linguistic and rule-based* NLP approaches, since their main advantage is the provision of accurate answers within a specific application domain (Utomo et al., 2017).

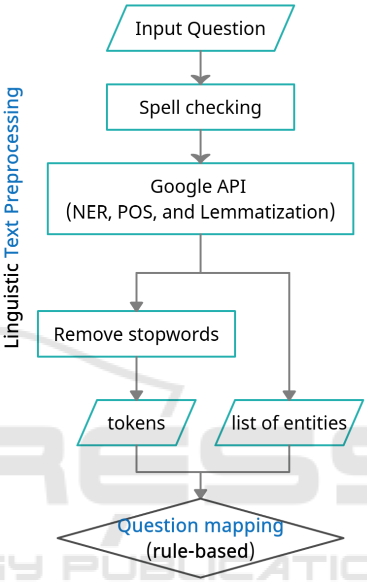


Figure 1: Overview of the hybrid NLP-based approach for querying Brazilian Open Data.

In Figure 1, we present an overview of our hybrid solution, covering two main stages of execution named *Text Preprocessing* and *Question Mapping*, where linguistic NLP and rule-based NLP are applied, respectively. The Figure shows that given an input sentence (i.e., a question) posed on our QA prototype, the *Text Preprocessing* stage executes linguistic methods such as Spell Checking, NER and POS (Part-Of-Speech) tasks, Lemmatization, and Stopwords removal, thus obtaining a list of entities and tokens that represent the meaningful part of the sentence. The entities and tokens are then given as input to the next processing stage, *Question Mapping*, that applies a set of rules for querying specific data tables and find the required information.

In Preprocessing stage, we choose Google Natural Language API[4], seeing that it offers a set of resources that are favorable to our approach. E.g., it supports multiple languages and allows to run NER, Part-Of-

[1] https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-escolar

[2] The *Courses* table is an auxiliary table since its data is not collected in Brazilian Census.

[3] https://www.w3.org/standards/

[4] https://cloud.google.com/natural-language

Speech and Lemmatization tasks, which are essential for our proposal. Also, the API does not demand any particular training, as it provides the same Deep Learning technology used by Google search engine for analyzing natural language texts (Google, 2021).

Although the chosen approaches are suitable for the nature of the data used, the Portuguese language involves a complex process of grammatical and mapping rules construction, due to its particularities. To face this complexity, we adopted relaxed conditions on the rules applied to the input questions, allowing that sentences written differently, but with the same semantics can obtain the same answer. For example, if the QA prototype receives the questions "*how many students are there at the Visconde de Cairu School in Santa Rosa?*" or "*how many students are there in Santa Rosa, at the Visconde de Cairu School?*", the answer presented should be the same, since they are equivalent questions.

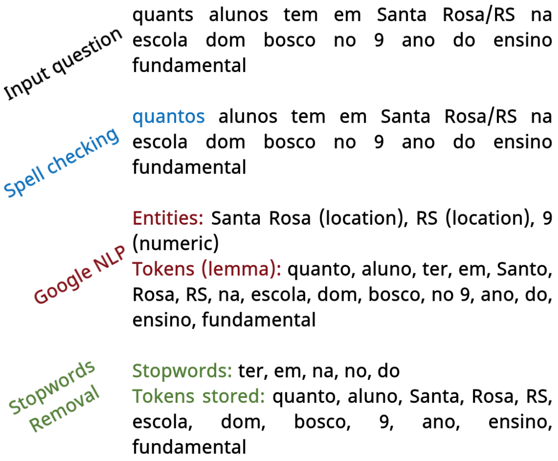
Given this overview about the design of our proposed solution, the next subsections address details of its two main stages: *Text Preprocessing* and *Question Mapping*.

3.2.1 Text Preprocessing

The *Text Preprocessing* stage aims to perform the first treatments on the input question, in order to obtain a cleaner sentence. Thus, given the user question posed on our QA prototype, a Spell Checking is firstly performed; then, the text is analyzed through Google NL API for lemmatizing the words, removing the stopwords, and identifying the named entities in the sentence.

The Figure 2 exemplifies this sequence of steps: the input question is firstly analyzed by the spell checking, so the word "quants" is modified to "quantos" (Portuguese word for *how many*), as highlighted in blue color. After this step, the Google API performs Named Entity Recognition and Part-of-Speech extraction, resulting in a list of tokens. Some tokens are lemmatized, e.g., the word "tem" is modified to "ter" in the sentence (Portuguese word for the verb *to have*). Finally, the stopwords are removed from the sentence, resulting in a list of tokens that are stored for further processing.

It is worth mentioning that extracting entities from a sentence is one of the most important tasks in our approach, since the entities referred to locations are present in all possible questions and directly interfere in answers accuracy. In the best use case, cities are informed along with the state abbreviation in capital letters (e.g., "Santa Rosa/RS"), thus we identify a Location entity. When this occurs, it is necessary to discover if this location is actually a Brazilian city, by querying an auxiliary database[5]. When the state is not informed by the user, ambiguities may occur, since there are cities with the same name em different states of the country. This will trigger an *entity classification error*. In these cases, we use the user location obtained through IP address to identify the interest region, thus improving the precision of the search.



quants alunos tem em Santa Rosa/RS na escola dom bosco no 9 ano do ensino fundamental

quantos alunos tem em Santa Rosa/RS na escola dom bosco no 9 ano do ensino fundamental

Entities: Santa Rosa (location), RS (location), 9 (numeric)
Tokens (lemma): quanto, aluno, ter, em, Santo, Rosa, RS, na, escola, dom, bosco, no 9, ano, do, ensino, fundamental

Stopwords: ter, em, na, no, do
Tokens stored: quanto, aluno, Santa, Rosa, RS, escola, dom, bosco, 9, ano, ensino, fundamental

Figure 2: Steps of Text Preprocessing applied to an input question.

As shown in Figure 2, Text Preprocessing outputs a list of tokens, which is given as input to the next step of our approach, i.e., *Question Mapping*. This step identifies keywords in the resulting list for knowing the information need of the user question, and will be discussed in the next subsection.

3.2.2 Question Mapping

To identify the meaning of the question, i.e., the information that the user really wants, it is necessary to extract keywords that give clues about this expected information. In other words, this step takes as input the resulting list of tokens from Text Preprocessing stage, and tries to identify (from the left to the right) keywords in this list that refer to Census data tables.

As mentioned in subsection 3.1, we have the tables **Students**, **Schools** and **Courses**, so valid questions should contain one of these keywords, or equivalent words. If one of these keywords is not identified, the processing will be stopped, since we infer that the question is not related to the application domain. Otherwise, when keywords are found, the rule-based NLP is triggered: each keyword found is linked

[5] This querying is carried out in an auxiliary table named *Cidades* (cities, in English), which is not part of the educational data. The table contains data collected in another government survey, thus the codes for cities, states and regions are the same as the ones used in Brazilian Census.

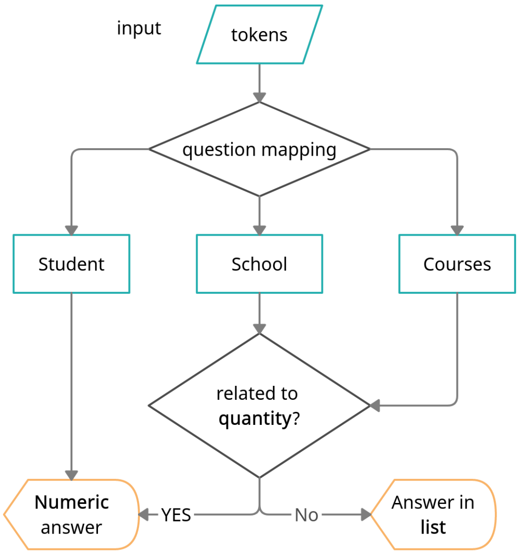


Figure 3: Question Mapping flowchart.

to a specific rule concerning different queries on the three data tables, as demonstrated in Figure 3.

The task of identifying the meaning of a question and providing the appropriate answer usually covers many steps that must be executed in sequence, which are called *pipelines* (Marx et al., 2014). Thus, for our rule-based mapping, we constructed pipelines for each type of keyword found (i.e., Student, School and Course). For exemplifying, in the resulting list of tokens from Figure 2, the second token “aluno” (Portuguese word for *student*) will trigger the processing pipeline for Student, which involves query attributes related to students. Similarly, if the keywords *School* or *Course* are found, their respective pipelines will be executed.

In the last processing stage of Figure 3, the information need regarding Schools and Courses may be related to quantity. We identify this relation when the token “quanto” (i.e., *how many*) is present in the list of tokens, thus we infer that the user is asking for values or quantities. If this token is not present, then the answer is given as a list.

Models of questions in Portuguese related to student, school, or course are presented in the Appendix of this paper, and are applied in our prototype of QA system developed as a Web page. In the next subsections, details of pipelines covered by each type of question will be minutely addressed.

Student-related Questions. These questions are limited to quantitative answers (see Figure 3) related to students information, i.e., number of students (1) enrolled in a given school, (2) linked to a specific educational modality/level, and (3) that use a specific mean of transportation. The Figure 4 shows the pipeline for answering student-related questions. The pipeline includes the following data:

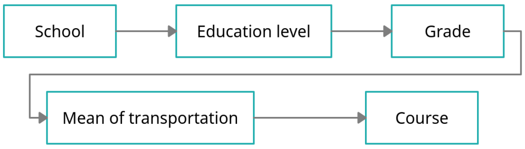


Figure 4: Student-related pipeline.

- **School.** The question may require information on a specific school. For identifying whether there is a school in the question, all tokens are examined and tested over the School table, considering a location. If any result is returned, we assume the presence of school in the question. In this step, it is necessary to ignore tokens that are not stopwords for Portuguese, but usually appear in the schools names. E.g., virtually all Brazilian schools have in their names the words *Escola, Estadual, Ensino Médio*[6], which are ignored since they are not relevant for the processing.
- **Educational Level.** It retrieves information on the educational level within a question. When the term *Ensino Fundamental*[7] is found, a filter is applied for searching for students aged 6 to 15 years.
- **Grade.** Elementary, Middle and High schools cover many grades. In case of the retrieved educational level is *Ensino Fundamental*, we consider grades within a range of 1 to 9. Otherwise, if the educational level is *Ensino Médio*[8], we consider grades within a range of 1 to 3.
- **Mean of Transportation.** It allows to retrieve information about students that use means of transportation to get to school. Keywords related to valid means of transportation are identified in the input question.
- **Course.** The question may require information on number of students enrolled in a given course, which can refer to a school or location/region. In this case, we examine all tokens from the sentence, testing them over the table Courses, similarly to the process performed for schools.

School-related Questions. The focus of these questions is discovering which schools are private or

[6] Portuguese words and expression for *School, Public School*, and *High School*, respectively.

[7] Portuguese expression for *Elementary School* and *Middle School together*.

[8] Portugese expression for *High School*.

public in a region (city or state). Besides that, it is possible to find public schools by informing their type of linkage with the State (e.g., informing if the school is municipal or national), and applying filters according to the educational level (childhood education, elementary school, high school, etc.) or modalities (professional education, e-learning). The Figure 5 shows the pipeline for answering school-related questions, which includes:

- **School Type.** The queries concerning School Type filter schools according to their linkage with the State. Thus, a school can be private or public. If the school is public, either the city, the state or the country governments can be responsible for its resources.
- **Educational Offer.** It allows to discover schools by applying filters related to which target audience is served, e.g., Elementary and Middle Schools offer education for students of the 1st to the 9th grade.
- **Ordination.** It allows to order decreasingly the answer in list. This resource is only possible for school-related questions, since the list of schools is much more extensive than the other answers lists. For this, we search for the words “desc” (i.e., *descending*) or “inverso” (i.e., *inverse*) informed in the question to perform ordination.

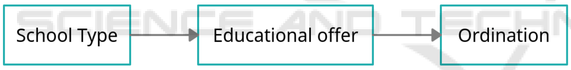


Figure 5: School-related pipeline.

Course-related Questions. These questions present a variation in their structure, since they require the discovery of three types of information: courses offered in a location, courses offered by a school, or schools that offer a given technical course. There is only one pipeline responsible for the decision, and the answer for this type of question can be related to courses or schools.

The Figure 6 presents the flowchart with the decisions taken for answering course-related questions: questions that begin with the interrogative adverb *Onde* (Portuguese word for *Where*) express the idea of requiring information about a *place*, thus we infer that this place is a school where a given course is offered. When “where” token is found, we identify the presence of a valid technical course in the sentence. We execute this task by examining all tokens and checking if one of them corresponds to a valid entry in the Courses table, according to an identifier. A `distinct` SQL query is performed on the Students table, adding the Course identifier as restriction, thus retrieving the school’s name.

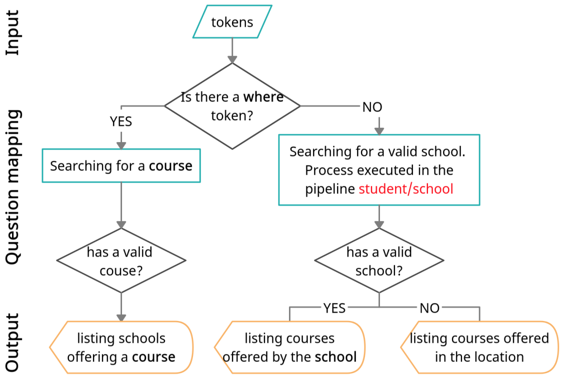

Input
Question mapping
Output
tokens
Is there a where token?
YES
NO
Searching for a course
Searching for a valid school. Process executed in the pipeline student/school
has a valid couse?
has a valid school?
YES
NO
listing schools offering a course
listing courses offered by the school
listing courses offered in the location

Figure 6: Course-related flowchart.

In case where the question does not contain “where”, we first search for a valid school in the sentence[9], then we perform a `distinct` SQL query in the Students table, adding the School identifier as restriction for retrieving the school’s name.

If no school is valid, it is necessary to search for all courses offered in that location. For this, we execute the same SQL query without adding the school identifier.

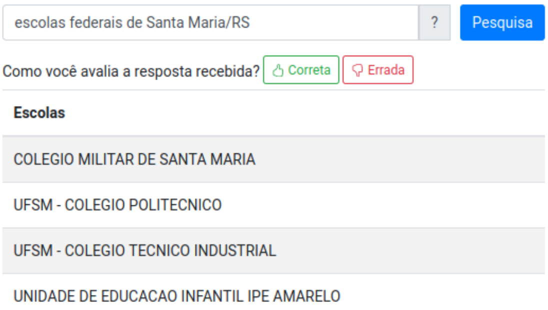
4 DISCUSSION

In this section we evaluate the success rate of answers obtained with our NLP-based approach, considering the three types of questions presented in subsection 3.2.2: student-related questions, school-related questions, and course-related questions. Models of questions in Portuguese accepted in our QA prototype are shown in the Appendix.

With respect to the school-related questions, the Figure 7 shows an example of input question, where it is informed the Portuguese sentence for *public schools in Santa Maria/RS*, where “Santa Maria” is a city, and RS means “Rio Grande do Sul”, a Brazilian state. The question aims to retrieve a list of schools maintained by the federal government, thus it applies the School Type filter in the school pipeline, as demonstrated in Figure 5. Concerning the results presentation, the lack of keywords indicating descending order means that the answer follows an ascending order, which is the ordination standard.

The Figure 8 presents an example of course-related question, whose objective is searching for schools that offer the Technical Nursing Course in

[9] This processing is sent to the Student/School pipeline, since the searching rules are the same.



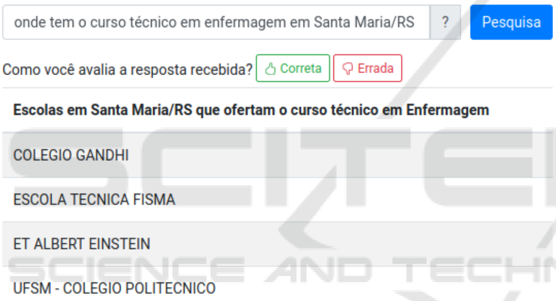
escolas federais de Santa Maria/RS

Como você avalia a resposta recebida?

Escolas
COLEGIO MILITAR DE SANTA MARIA
UFSM - COLEGIO POLITECNICO
UFSM - COLEGIO TECNICO INDUSTRIAL
UNIDADE DE EDUCACAO INFANTIL IPE AMARELO

Figure 7: Results obtained from a school-related question.

Santa Maria. We can observe that this question starts with the term “onde” (*where*), thus it requires information about a place, as shown in Figure 6. By following the course-related flowchart, the answer lists all schools located in Santa Maria that offer the Technical Nursing Course. The results were validated through searches performed in the websites of the institutions, which confirmed the veracity of answers.



onde tem o curso técnico em enfermagem em Santa Maria/RS

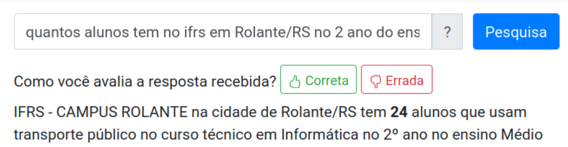
Como você avalia a resposta recebida?

Escolas em Santa Maria/RS que ofertam o curso técnico em Enfermagem
COLEGIO GANDHI
ESCOLA TECNICA FISMA
ET ALBERT EINSTEIN
UFSM - COLEGIO POLITECNICO

Figure 8: Results obtained from a course-related question.

Finally, the Figure 9 presents an example of student-related question that combines the other pipelines, i.e., school and course. The input question shown in the Figure aims to discover “*how many **students** attending the **2nd year** of the **Computing Technical Course**[10] are there in **IFRS**[11], located in **Rolante/RS**[12], and that use **public transportation**?*”. The answer obtained from this question highlights that 24 students match the conditions informed. This result shows that although the possible queries are not highly complex to execute, the proposed model allows to give complete answers by combining different pipelines, which increases the usefulness of information retrieved.

For improving our proposal and future evaluations, our QA prototype allows the user to give feedback about the retrieved answers: above each list of answers in figures 7, 8 and 9, there is the Portuguese sentence “*Como você avalia a resposta recebida?*”, which asks to the user *how he/she evaluates the answer received*. The sentence is accompanied by two buttons for the user to inform if the answer is *Correct* or *Incorrect*, allowing to assess the success rate of our approach. In addition, the user can leave comments in our QA prototype about the consistency of answers, contributing for the system improvement.



quantos alunos tem no ifrs em Rolante/RS no 2 ano do ens

Como você avalia a resposta recebida?

IFRS - CAMPUS ROLANTE na cidade de Rolante/RS tem **24** alunos que usam transporte público no curso técnico em Informática no 2º ano no ensino Médio

Figure 9: Results obtained from a student-related question that includes information from school and course pipelines.

4.1 Limitations

In the evaluation phase of our proposal, we identified limitations that affected the accuracy of some questions. First, in some cases, the Spell Checking erroneously changed a city name that was correct, producing an unexpected output. Errors also occurred in NER tasks, e.g., in the Portuguese question corresponding to “*how many students are there in IFRS and Sagrada Família located in Rolante/RS?*”, IFRS and Sagrada Família correspond to schools in the Rolante city. However, the QA prototype searches for the **city** named Sagrada Família, not the school as expected. As Sagrada Família city does not have any of the two schools mentioned, an unexpected output is produced.

A possible solution for facing this limitation could be adding a third module in our proposal, responsible for disambiguating the question meaning. This could be done by retrieving, e.g., names of valid locations within a sentence, and asking the user a confirmation about which location is correct. This approach has been applied in several NLP-based approaches, achieving satisfactory results (Mani et al., 2018; Damljanovic et al., 2011; Park et al., 2015).

5 CONCLUSIONS AND FUTURE WORK

In this work we presented a NLP-based approach for querying Open Data from Brazilian Educational Census, which is the most important educational statistical research in the country. Brazilian Open Data are available in Portuguese language, which is still a re-

[10] High School course.

[11] Name of public school.

[12] City and state in Brazil.

search gap in the NLP domain due to its complexity and particular structure. Also, there is a lack of QA approaches for accessing and processing Brazilian Open Data.

For facing these issues, we implemented a QA prototype developed as a Web interface for querying Brazilian Educational Census, proposing a hybrid solution that applies linguistic and rule-based NLP in two main stages of question processing: Text Preprocessing and Question Mapping. In Text Preprocessing, we applied techniques such as Spell Checking and stopwords removal, as well as Google NLP API for Lemmatization and entity recognition tasks. In Question Mapping, we discovered the meaning of the input question, recognizing keywords related to specific data tables from the Brazilian Census (schools, courses, and students), and following their respective pipelines (i.e., sets of rules) for retrieving the correct answer.

Our approach was validated through questions related to the data tables, and posed on our QA prototype[13]. The answers obtained with our hybrid NLP approach were satisfactory, since they presented precise information, with the possibility of combining information from more than a pipeline.

As future work, we aim to validate and promote the effective use of our approach, by including education professionals (and other users who are interested) in evaluation tasks, so that inconsistencies in the answers can be better handled. Particularly, we aim to include a disambiguation module in our hybrid approach, for dealing with both Spell Checking and NER limitations, in which human participation can be highly effective. We also aim to expand the number of questions accepted by our prototype, and investigate other NLP techniques for optimizing aspects of accuracy, response time and clarity of answers.

REFERENCES

Balakrishnan, V. and Lloyd-Yemoh, E. (2014). Stemming and lemmatization: a comparison of retrieval performances. *Lecture Notes on Software Engineering*, 2:262–267.

Beniwal, R., Gupta, V., Rawat, M., and Aggarwal, R. (2018). Data mining with linked data: Past, present, and future. In *2018 Second International Conference on Computing Methodologies and Communication (ICCMC)*, pages 1031–1035. IEEE.

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

Charton, E., Ghoula, N., and Meurs, M.-J. (2016). Open data for local search: Challenges and perspectives. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 641–644.

Damljanovic, D., Agatonovic, M., and Cunningham, H. (2011). Freya: An interactive way of querying linked data using natural language. In *Extended semantic web conference*, pages 125–138. Springer.

de Castro, B. P. C., Rodrigues, H. F., Lopes, G. R., and Campos, M. L. M. (2019). Semantic enrichment and exploration of open dataset tags. In *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web*, pages 417–424.

dos Santos Brito, K., da Silva Costa, M. A., Garcia, V. C., and de Lemos Meira, S. R. (2015). Is brazilian open government data actually open data?: An analysis of the current scenario. *International Journal of E-Planning Research (IJEPR)*, 4(2):57–73.

Ferrández, A. and Peral, J. (2010). The benefits of the interaction between data warehouses and question answering. In *Proceedings of the 2010 EDBT/ICDT Workshops*, pages 1–8.

Giordani, A. and Moschitti, A. (2012). Generating sql queries using natural language syntactic dependencies and metadata. volume 7337, pages 164–170.

Google, L. (2021). Cloud natural language. https://cloud.google.com/natural-language.

Goyal, A., Gupta, V., and Kumar, M. (2018). Recent named entity recognition and classification techniques: A systematic review. *Computer Science Review*, 29:21–43.

Gupta, P. and Gupta, V. (2012). A survey of text question answering techniques. *International Journal of Computer Applications*, 53:1–8.

Henderson, M., Al-Rfou, R., Strope, B., Sung, Y.-H., Lukács, L., Guo, R., Kumar, S., Miklos, B., and Kurzweil, R. (2017). Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

Hossain, M. A., Dwivedi, Y. K., and Rana, N. P. (2016). State-of-the-art in open data research: Insights from existing literature and a research agenda. *Journal of organizational computing and electronic commerce*, 26(1-2):14–40.

INEP (2020). Censo escolar. http://portal.inep.gov.br/web/guest/censo-escolar. Accessed: 2020-12-30.

Isotani, S. and Bittencourt, I. I. (2015). *Dados Abertos Conectados: Em busca da Web do Conhecimento*. Novatec Editora.

Kaur, A., Singh, P., and Rani, S. (2014). Spell checking and error correcting system for text paragraphs written in punjabi language using hybrid approach. *International Journal of Engineering and Computer Science*, 3(09).

Khosrow-Pour, M. and Khosrowpour, M. (2009). *Encyclopedia of information science and technology*, volume 1. IGI Global Snippet.

Kolomiyets, O. and Moens, M.-F. (2011). A survey on question answering technology from an infor-

[13] Our QA prototype for querying Brazilian Educational Census is available in the following address: https://censoindex.duckdns.org.

mation retrieval perspective. *Information Sciences*, 181(24):5412–5434.

Lee, J., Yi, J.-S., and Son, J. (2019). Development of automatic-extraction model of poisonous clauses in international construction contracts using rule-based nlp. *Journal of Computing in Civil Engineering*, 33(3):04019003.

Liddy, E. D. (2001). Natural language processing.

Mani, S., Gantayat, N., Aralikatte, R., Gupta, M., Dechu, S., Sankaran, A., Khare, S., Mitchell, B., Subramanian, H., and Venkatarangan, H. (2018). Hi, how can i help you?: Automating enterprise it support help desks. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Marx, E., Usbeck, R., Ngomo, A.-C. N., Höffner, K., Lehmann, J., and Auer, S. (2014). Towards an open question answering architecture. In *Proceedings of the 10th International Conference on Semantic Systems*, pages 57–60.

Masseroli, M., Picozzi, M., Ghisalberti, G., and Ceri, S. (2014). Explorative search of distributed bio-data to answer complex biomedical questions. *BMC bioinformatics*, 15(S1):S3.

Molina Gallego, G. (2018). Analysis, discovery and exploitation of open data for the creation of question-answering systems. *Undergraduate Thesis, Universidad de Alicante*.

Mouromtsev, D., Vlasov, V., Parkhimovich, O., Galkin, M., and Knyazev, V. (2013). Development of the st. petersburg's linked open data site using information workbench. In *Open Innovations Association (FRUCT), 2013 14th Conference of*, pages 77–82. IEEE.

Murray-Rust, P., Neylon, C., Pollock, R., and Wilbanks, J. (2010). Panton principles, principles for open data in science. *Panton Principles*.

Nesi, P., Pantaleo, G., and Sanesi, G. (2015). A hadoop based platform for natural language processing of web pages and documents. *Journal of Visual Languages & Computing*, 31:130–138.

Oliveira, M. I. S., de Oliveira, H. R., Oliveira, L. A., and Lóscio, B. F. (2016). Open government data portals analysis: the brazilian case. In *Proceedings of the 17th International Digital Government Research Conference on Digital Government Research*, pages 415–424.

Park, S., Kwon, S., Kim, B., Han, S., Shim, H., and Lee, G. G. (2015). Question answering system using multiple information source and open type answer merge. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 111–115.

Rocha, G. and Lopes Cardoso, H. (2018). Recognizing textual entailment: challenges in the portuguese language. *Information*, 9(4):76.

Rodrigues, R., Gonçalo Oliveira, H., and Gomes, P. (2018). Nlpport: a pipeline for portuguese nlp (short paper). In *7th Symposium on Languages, Applications and Technologies (SLATE 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Roy, R. S. and Anand, A. (2020). Question answering over curated and open web sources. *arXiv preprint arXiv:2004.11980*.

Ruiz, M., Román, C., Garrido, Á. L., and Mena, E. (2020). uais: An experience of increasing performance of nlp information extraction tasks from legal documents in an electronic document management system. In *ICEIS (1)*, pages 189–196.

Singh, J. and Gupta, V. (2017). A systematic review of text stemming techniques. *Artificial Intelligence Review*, 48(2):157–217.

Sowe, S. K. and Zettsu, K. (2015). Towards an open data development model for linking heterogeneous data sources. In *Knowledge and Systems Engineering (KSE), 2015 Seventh International Conference on*, pages 344–347. IEEE.

Srba, I., Savic, M., Bielikova, M., Ivanovic, M., and Pautasso, C. (2019). Employing community question answering for online discussions in university courses: Students' perspective. *Computers & Education*, 135:75–90.

Sy, M.-F., Ranwez, S., Montmain, J., Regnault, A., Crampes, M., and Ranwez, V. (2012). User centered and ontology based information retrieval system for life sciences. *BMC bioinformatics*, 13(1):1–12.

Utomo, F., Suryana, N., and Azmi, M. S. (2017). Question answering system: A review on question analysis, document processing, and answer extraction techniques. *Journal of Theoretical and Applied Information Technology*, 95:3158–3174.

Wendt, M., Gerlach, M., and Düwiger, H. (2012). Linguistic modeling of linked open data for question answering. In *Extended Semantic Web Conference*, pages 102–116. Springer.

White, R. W., Richardson, M., and Yih, W.-t. (2015). Questions vs. queries in informational search tasks. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, page 135–136, New York, NY, USA. Association for Computing Machinery.

Yao, X. and Van Durme, B. (2014). Information extraction over structured data: Question answering with freebase. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 956–966.

Yin, Z., Zhang, C., Goldberg, D. W., and Prasad, S. (2019). An nlp-based question answering framework for spatio-temporal analysis and visualization. In *Proceedings of the 2019 2nd International Conference on Geoinformatics and Data Analysis*, pages 61–65.

Zhang, C. and Yue, P. (2016). Spatial grid based open government data mining. In *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International*, pages 192–193. IEEE.

APPENDIX

Models of questions in Portuguese accepted by the QA prototype:

School-related Questions.

- Quais escolas federais tem em Santa Maria/RS? *(Which federal schools are there in Santa Maria/RS?)*
- Quais escolas particulares existem em Frederico Westphalen/RS? *(Which private schools are there in Frederico Westphalen/RS?)*
- Quais escolas públicas existem em Frederico Westphalen/RS? *(Which public schools are there in Frederico Westphalen/RS?)*

Student-related Questions.

- Quantos alunos tem na cidade de Rolante/RS? *(How many students are there in Rolante/RS city?)*
- Quantos alunos tem na escola Visconde de Cairu em Santa Rosa/RS? *(How many students are enrolled at Visconde de Cairu school in Santa Rosa/RS?)*
- Quantos alunos tem em Rolante/RS que usam transporte público? *(How many students in Rolante/RS use public transportation?)*
- Quantos alunos tem no IFRS Rolante/RS no curso técnico em informática? *(How many students are enrolled in the Computing Technical Course at IFRS located in Rolante/RS?)*
- Quantos alunos tem na pré-escola em Santa Rosa/RS? *(How many nursery school students are there in Santa Rosa/RS?)*
- Quantos alunos tem em Santa Rosa/RS na escola Dom Bosco no 9 ano do ensino fundamental? *(How many students at Dom Bosco school in Santa Rosa/RS are in the 9th grade of Elementary School?)*

Course-related Questions.

- Onde é ofertado o curso técnico em informática em Porto Alegre? *(Where is the Computing Technical Course offered in Porto Alegre?)*
- Quais cursos tem na cidade de Taquara/RS? *(Which courses are there in Taquara/RS city?)*
- Quais cursos tem no IFRS em Rolante/RS? *(Which courses are offered at IFRS in Rolante/RS?)*