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 easily 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...
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.

<table>
<thead>
<tr>
<th>English</th>
<th>Portuguese</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student</td>
<td>Estudante</td>
<td>Masculine/Feminine</td>
</tr>
<tr>
<td>Student</td>
<td>Aluno</td>
<td>Masculine</td>
</tr>
<tr>
<td>Student</td>
<td>Aluna</td>
<td>Feminine</td>
</tr>
</tbody>
</table>

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 answering 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 (Hos-sain 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
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.](https://cloud.google.com/natural-language)

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 “ter” 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,

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

The **Input question** is: *quantos alunos tem em Santa Rosa/RS na escola dom bosco no 9 ano do ensino fundamental*

**Spell checking** modifies “quants” to “quantos”.

**Google NLP** identifies entities: *Santa Rosa (location), RS (location), 9 (numeric)*

**Tokens (lemma):** *quanto, aluno, ter, em, Santa, 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.

by querying an auxiliary database. 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.

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

---

5This querying is carried out in an auxiliary database. 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 “ter” 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,
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.](image)

- **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, 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 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, 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

---

6Portuguese words and expression for School, Public School, and High School, respectively.
7Portuguese expression for Elementary School and Middle School together.
8Portuguese 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.

**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 technical 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.

In case where the question does not contain “where”, we first search for a valid school in the sentence, 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 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.

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 technical 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.

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 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.

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 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.
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.

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 are there in IFRS and Sagrada Familia located in Rolante/RS?". IFRS and Sagrada Familia correspond to schools in the Rolante city. However, the QA prototype searches for the city named Sagrada Familia, not the school as expected. As Sagrada Familia 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-
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\textsuperscript{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


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

\textsuperscript{13}Our QA prototype for querying Brazilian Educational Census is available in the following address: https://censoindex.duckdns.org.
Que a busca de dados educacionais brasileiros usando um abordagem NLP-híbrida

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?)