Vol.2016-NL-226 No.15
Vol.2016-SLP-111 No.15
2016/5/17

An Over-generator of Questions: Towards Automatic Question Generation in Portuguese Language

Josimar Hermínio LOPES[1,a)] Yoshinori TAKEI[1,b)]

Abstract: The internet has become an indispensable tool for reference in the educational system. Often times, students are given articles or texts as study materials. However, such materials do not provide means of assessing the understanding of the students, thus making it a challenge. This work addresses such challenge, in the Portuguese language, by automating the generation of questions from a given text. We focus our study on generating questions from factual texts in Portuguese. We approach the question generation challenge by over-generating questions; Our system follows a rule-based approach, a set of rules that we constructed for the Portuguese language which enables the selection of answer phrases in declarative sentences, in order to generate questions. In the statistical analysis of the over-generated questions, we noticed that the number of questions is 1.77 times the number of input sentences. This results are promising for, our future work, the development of a system that will rank the over-generated questions.

Keywords: question generation (QG), rule-based system, portuguese, natural language processing (NLP).

1. Introduction

It seems completely normal and effortless for humans to generate questions, whereas machines struggle a lot to understand natural language (D. Jurafsky, 2009) [1] since they can not on their own. In Portuguese, question generation, question answering and related systems are still open domain problems. In recent years, question generation studies have been receiving significant attention due to promising applications in learning technologies such as educational systems, intelligent tutoring systems, and dialogue or conversation systems.

A question generation system function is to generate questions from a text, by simplifying the text into declarative sentences and then selecting potential answer phrases and thus generating questions, and possibly presenting such questions to the student for educational assessments and practice. For example, (J. Brown, 2005) [2] describes a system that automatically evaluates an individual's reading levels by assessing the user's vocabulary knowledge.

In this paper, we focus on generating questions for assessing the understanding of the student from a given text. Our efforts are directed towards educational materials in which knowledge extraction is the key point of this assessment. The assessment may be useful to educators or private readers as it presents a text, to the students or readers, and then questions related to the text that they read in order to assess their understanding. The scope of this work is simply aimed at (providing an implementation of a system by only) analyzing educational materials that contain facts (such as scientific articles) or information that can be extracted and used as answers to the questions being generated. Mostly, the system prioritizes the generation of questions (*Q-questions*) in which the answer phrases are contained in named entities and places at last (*Y/N-questions* or) questions which are easy to construct.

This paper is organized as follows: In §1 we detail the background of our QG system and discuss related works. In §2 we discuss preliminaries regarding concepts, conventions and main system tools. In section §3, we describe our implementation of the QG system for Portuguese. In §4 we make some experiments and discuss the results. In §5 we discuss further research and development. And finally in §6 we make concluding remarks.

1.1 Background

In this section, we describe the structure used to generate question from a declarative sentence of a given text. In our system, the process of generating a single question can be described in two main stages shown in **Fig. 1**. The figure emphasizes a similar strategy of "overgenerate-and-rank" that (Walker et al., 2001[3], I. Langkild., 1998[4], and Heilman and Smith, 2009[20]) employed in their works.

Though, raw text undergoes a set of NLP transformations (such as summarization, sentence compression, sentence splitting, sentence fusion, paraphrasing, textual entailment, lexical semantics for word substituition) for the extraction of declarative sentences, in this paper we do not perform these transformations as we assume that enough sets of declarative sentences, which do not require any further transformation, are fed into the system, and thus leaving the simplification stage aside.

The following is an example of simplification of a text, (an article on *filosofia* <philosophy>, from Wikipedia, 2016[5]):

"Filosofia é o estudo de problemas fundamentais relaciona-

[1] Nagaoka University of Technology, 1603-1 Kamitomiokamachi, Nagaoka-shi, 940-2188 Japan
[a)] s145062@nagaokaut.ac.jp
[b)] takei@nagaokaut.ac.jp

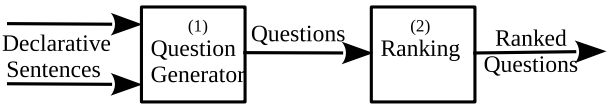

Fig. 1 Stages of the automatic question generation system

dos à existência, ao conhecimento, à verdade, aos valores morais e estéticos, à mente e à linguagem. Ao abordar esses problemas, a filosofia se distingue da mitologia e da religião por sua ênfase em argumentos racionais; por outro lado, diferencia-se das pesquisas científicas por geralmente não recorrer a procedimentos empíricos em suas investigações."
<Philosophy is the study of fundamental problems related to existence, knowledge, truth, moral and ethic values, mind and language. When addressing these problems, philosophy distinguishes itself from mythology and religion due to its stress in rational arguments; on the other hand, it distinguishes itself from scientific researches by generally not recurring to empirical procedures in its investigations.>
can be simplified into following declarative sentences:

- *Filosofia é o estudo de problemas fundamentais relacionados à existência, ao conhecimento, à verdade, aos valores morais e estéticos, à mente e à linguagem.* <Philosophy is the study fundamental problems related to existence, knowledge, truth, moral and ethic values, mind and language.>
- *A filosofia se distingue da mitologia e da religião por sua ênfase em argumentos racionais.* <Philosophy distinguishes itself from mythology and religion due to its stress in rational arguments.>
- *Filosofia diferencia-se das pesquisas científicas por geralmente não recorrer a procedimentos empíricos em suas investigações.* <It distinguishes itself from scientific researches by generally not recurring to empirical procedures in its investigations.>

Under the assumption that this simplification is done independently, we leave this issue aside for studies focused on summarization, sentence simplification and related areas.

In stage 1, a declarative sentence is provided as an input to the "question generator" which generates questions by applying syntactic transformations (such as Q-movement, subject-verb inversion, etc.), i.e. questions are generated for a specific declarative sentence.

In stage 2, the questions generated from the QG system (stage 1) will be scored and then ranked. The ranking process will be performed with the features collected from the input sentence, the questions, and the transformations applied during generation. However, in this work, we only discuss about the *Question generator (1)* stage and leave the *ranking (2)* stage for future works.

1.2 Related Works

Several Question Generation systems have been proposed over the years, researchers have turned attention and have attempted different solutions to it. The quest for responsive systems dates back to early years in which primitive language processing programs, such as ELIZA (J. Weizenbaum, 1966)[6] , provided responses by processing the input text.

One of proposed approaches to QG deal with transforming answers to questions by utilizing the question generation process as an intermediate stage in the question answering process (Echihabi and Marcu, 2003[7]; Hickl et al., 2005[8]). Other works approached this challenge by performing syntactic transformations for question generation (Heilman and Smith, 2009[20]; Wyse and Piwek, 2009[9]) having its applications in educational domains.

There are also works based on templates (Mostow and Chen, 2009[10]; Chen, Aist, and Mostow, 2009 [11]) and semantic transformations (Schwartz, Aikawa, and Pahud, 2004[12]; Sag and Flickinger, 2008[13]; Yao, 2010[14]). Most of these systems deduce the question phrases with the help of well-known named entity recognizer outputs (PERSON, LOCATION and ORGANIZATION) or template definition.

In portuguese language, QG systems are still scarce, (D. Diéguez, R. Rodrigues and P. Gomes, 2011[15]) discuss an approach that combines a case-base reasoning system and a module for question generation. The QG module uses manually built rules that are fed to the case-based reasoning engine for selecting which ones should be used.

This paper solves the QG challenge from Heilman and Smith approach. Our contributions in this paper are as follows:

- We construct a set of rules for identifying answer phrases in Portuguese declarative sentences.
- We also construct a set of regular expressions that perform the transformations of the declative sentence into a question (for example, answer to question phrase replacement, subject-verb inversion and question generation).
- We implement a system that also accepts texts without subjects (due to many verbal inflections in the Portuguese language) and which may include the presence of reflexive clitics; This may not be supported by other works, for example in 'Heilman and Smith, 2009's[20] QG for English language.
- We develop a QG system for the Portuguese language using a rule based approach, which have may not been implemented yet.

2. Preliminaries

We start by point out some important concepts and formalities that will help us understand the subsequent phases of our system.

2.1 Definitions

Throughout the paper we will use terms like "declarative sentence", "answer phrase", and "question phrase". A "declarative sentence" is a sentence from which questions will be generated (e.g., *A capital de Moçambique é Maputo.* <The capital of Mozambique is Maputo.>), i.e. a "declarative sentence" is given as input to the question generator. The term "answer phrase" refers to phrases in declarative sentences which can undergo Q-movement (i.e. phrases that can be replaced with interrogative pronouns, in Portuguese, in order to formulate a question) and therefore can be considered as answers to generated questions (e.g., *A capital de Moçambique* <The capital of Mozambique >). A "question phrase" refers to the phrase that replaces the "answer phrase" with a question word (e.g., "O que" in *O que é Maputo?* <What is Maputo? >).

In the construction of the QG system, we make use of Penn

Treebank (B. Santorini, 1990)[16] style phrase structure trees containing part-of-speech (POS) tags, and we interfaced the Stanford Parser (D. Chen and C. D. Manning, 2014)[17] with the LX-Parser (J. Silva, A. Branco, S. Castro and R. Reis, 2010)[18] in order to support Portuguese sentences (i.e., the outcome is Portuguese Pen Treebanks).

2.2 Tools: Tregex and Tsurgeon

The process of transforming declaratives sentences into questions is possible with a set of rules that search and manipulate specific portions of the trees, for this reason we resort to (R. Levy and G. Andrew, 2006)[19] tools, respectively *Tregex* (a tree query language) and *Tsurgeon* (a tree manipulation language). In previous QG papers like (M. Heilman and N. A. Smith, 2009)[20] and (D. M. Gates, 2008)[21] have made use of these tools.

The *Tregex* (i.e., "tree regular expression") utility provides a wide range of options for searching and matching patterns in trees, based on tree relationships and regular expression matches on nodes. It involves relations such as dominance, immediate dominance (denoted as "<<", "<"), and contraints on headship nodes. *Tregex* also allows regular expressions to be embbeded in the syntax. On top of *Tregex* ability of matching, *Tsurgeon* adds the ability to *manipulate* trees with operations like relabelling, deleting, moving and inserting nodes.

The *Tregex* and *Tsurgeon* work closely together in that patterns matched can be labelled and the then modified with the *Tsurgen* operations. Let us consider the *Tregex* expression: "$VP|V' < +(VP|V')\ (V\ \$\ (NP = unmv\ <\ CL))$", this would mean finding noun phrases ("*NP*"), which *immediately dominate* a clitic ("*CL*"), that have a verb ("*V*") as a *sister* (denoted as "$") which are *immediately dominated* by chain of verb phrases or verb composites ("+(*VP*|*V*′)") as head node. This expression matches specific nodes in sentences such as "Ele ofereceu-***me*** um livro" <He offered me a book>, '***me***' (**Fig. 2** shows an example of the tree format that *Tregex* accepts). The label (variable) *unmv* assigned to such matches can be used by *Tsurgeon* in order to perform operations.

In our implementation, we make an extensive use of *Tregex* and *Tsurgeon* to commit the constraints of our rule set (we will discuss "rule set" in the next section). Though some limitations prevail, as Heilman and Smith, 2009[20] pointed out in their paper, *Tsurgeon* cannot perform operations one at a time (all matching nodes must be transformed simultaneously), back-reference relabelled nodes, or include reserved words (e.g., "*insert*") in new node labels. Despite these limitations, the inclusion of these tools in our system produce the desired results (when marking unmovable phrases and selecting answer phrases, we will discuss these in the next section).

2.3 Portuguese Language

Portuguese is a Romance language and part of the Indo-European language family. It is closely related to Spanish. It is spoken by about 260 million people world-wide, principally in Brazil and Portugal (Wikipedia, 2016)[22]. The Portuguese spoken in Europe (EP) and the Portuguese spoken in Brazil (BP) are further apart in terms of pronunciation, spelling and vocabulary

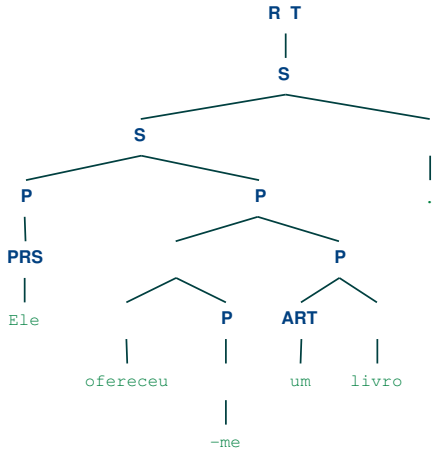

Fig. 2 Tree in Penn TreeBank form, which is accepted by Tregex and Tsurgeon.

than the English spoken in England and the English spoken in the USA.

A notable aspect of the grammar of Portuguese is the verb. Morphologically, more verbal inflections from classical Latin have been preserved by Portuguese than by any other major Romance language. There are many differences between EP and BP, whether in vocabulary or grammar and speaking or writing and so on. In this paper we focus on European Portuguese (EP) to generate questions.

The set of language resources and tools for the English language is vast and extensive given that it's widely used worldwide, whereas for Portuguese language such resources and tools are still emerging. So, it is a challenge to develop a question generation system, specially for the Portuguese language.

In this paper, we develop a question generation system considering limitations and try to minimize erroneous questions generated by the system.

3. QG System for Portuguese

In this section, we will clarify our QG system as we analyze QG stage along with the operations involved. In short, in this stage, the system accepts a declarative sentence and generates questions from it.

3.1 Question Generator (1)

Questions that are semantically correct is the most important goal of every question generation system. Though, realistically, that may not always be the case because some sentences do not provide facts which may be used as answer phrases. To reduce the generation of semantically incorrect questions, we understood that it is feasible to supply the system with fact-based sentences; Otherwise, generated questions tend to be vague and less meaningful. Furthermore, not every phrase on the declarative sentence can be used as *answer phrase*, to address this issue we constructed a rule-set of constraints of phrases that can not be considered as valid answer phrases. We will discuss such instances in 3.1.1.

The *question generator* takes as input a declarative sentence and produces as output a set of possible questions.[*1] The *question generator* is responsible to identify "answer phrases" that can be converted into "question phrases". The system considers

Table 1 Tregex expressions for identifying phrases that cannot be considered as *"answer phrases"*

Contraints that lead to unmovable phrases	**Expression**
The prepositional phrases in subjects are marked as unmovable, since the question phrases cannot occur within the subject (e.g., **O que localiza-se na república de África?* <What is located in the Republic of Africa?> from *"A República de Moçambique localiza-se em África"* <The Republic of Mozambique is located in Africa>).	NP $ VP <<PP=unmv
The subordinate clauses which are not children of verb phrases	CP=unmv[!>VP|$−− /,/|<ADV] & [!<,(C $.. S)]
The subject of subordinate clauses that act as a complementizer phrase (e.g., **O que Pedro disse que está demitida?* <What did Peter day that is fired?> from *Pedro disse que **a Maria** está demitida* <Peter said that **Mary** is fired>.).	CP<,C<(S<(NP=unmv!$,,VP) |<(VP <NP=unmv))
The clauses which are descendants of verb phrases and are offset by commas, also signaled by adjuncts, which supplement the main body of the sentence.	VP<<(S=unmv $,, /,/)
Phrases under subordinate conjuctions which generally include the question phrase words.	CONJP<CONJ <<NP|VP|ADVP|PP=unmv
Prepositional phrases appearing within prepositional phrases (e.g., *O que Joana beijou Pedro no parque das?* <What did Joana kiss Peter in the park of?>) from *Joana beijou Pedro no parque das **flores** <Joana kissed Peter in the park of **flowers**>*.	PP <<PP=unmv
Phrases under conjuctions (e.g., *Quem Manuel encontrou e Maria?* <Who did Manuel find and Mary?> from the sentence *Manuel encontrou **Pedro** e Maria* <Manuel met **Peter** and Mary>).	(/.*/<,CONJ|<,(CONJ'<CONJ))$−−(NP|PP|VP=unmv) |<–(NP|PP|VP=unmv)
Clauses that serve as predicate of the main clause with a copula verb.	S<(VP<+(VP)(V<é|era) <<(VP|AP|PP|ADVP=unmv))
Reflexive clitics appearing under noun phrases after the verb (e.g., *Quem ele ofereceu um livro?* <Who did he offer a book?> from the sentence *Ele ofereceu-**me** um livro.* <He offered **me** a book>).	VP|V'<+(VP|V')(V$ (NP=unmv<CL))
Propagating rules: rules that ensure that nodes below (child of) unmovable nodes can not be selected as valid answer phrases and also ensure that redundant nodes are not considered as answer phrases.	1. NP|PP|AP|ADVP|CP<<NP|AP|ADVP|VP|N'|CP =unmv 2. @UNMOVABLE<<NP|ADJP|VP|ADVP|PP|AP| CONJP|CP=unmv

noun phrases, prepositional phrases and subordinating clauses[*2] to be candidates to "answer phrases". Accordingly, the "question phrases" supported are mainly from interrogative pronouns, respectively *quem, (o) que, qual(is), quanto(a)(s), quando* and *Onde* (meaning: who, what, how much, when and where).

Different types of questions can be generated from a single sentence, whether simple or complex depending on the answer we expect. Questions from sentences that demand explanation or reasoning, like *porquê* <*why*>, require more than syntactic transformations thus invoking semantic learning of the sentence. Though the system could be expanded to accomodate other types of questions, in this work we only generate mentioned questions and leave related challenges to future works.

The output of a question generator for a single sentence may lead to a set of possible questions (that may potentially be correct) due to the fact that a sentence may contain multiple answer phrases, for example, from the sentence *A pequenina exigiu que o chapéu ficasse sobre a banca* <The little girl demanded that the hat be placed on the table>, the question generator would produce the following questions:

- *Quem exigiu que o chapéu ficasse sobre a banca?* <Who demanded that the hat be placed on the table?>
- *Exigiu a pequenina que o chapéu ficasse sobre a banca?* <Did the little girl demand that the hat be place on the table?>
- *(O) que exigiu a pequenina?* (What did the little girl demand?)
  or *Qual exigiu a pequenina?* <What did the little girl demand?>
- *Sobre o que a pequenina exigiu que o chapéu ficasse?* <What did the little girl demand that the hat be placed on?>
  or *Sobre qual a pequenina exigiu que o chapéu ficasse?* <What did the little girl demand that the hat be placed on?>

The reason there are alternatives ('or' conjunctions) for some questions is because in Portuguese the meaning of *what* is ambiguous, i.e. *o que* and *qual*, and furthermore this may lead to the generation of questions with wrong question phrases. We approach this ambiguity problem, in future works, by introducing the ranking stage.

In short, the question generator stage takes in a declarative sentence and:

(1) Marks unmovable phrases (i.e, phrases that cannot be used as answer phrases).
(2) Replaces identified answer phrases with corresponding question phrases.
(3) Performs subject-verb inversion.
(4) Performs Question phrase insertion.

This sequence of operations is fundamental to generating questions, we will notice that not necessarily every transformation is applied in the generation process. In **Fig. 3**, we show in detail the operations performed (by this stage) in order to generate questions.

3.1.1 Marking Unmovable Phrases

The noun phrases, prepositional phrases, and subordinating phrases node heads that can not be considered as valid answer phrases are marked as unmovable by the *Tregex* expressions presented in Table 1. The *Tregex* expressions match unmovable nodes and *Tsurgeon* makes sure that it relabels those nodes by adding a label ("UNMOVABLE-") in the beginning of each node.

In the case of the rules defined in Table 1, any change to the attributes of a particular node in a monotonous manner, namely from *movable* to *UNMOVABLE*. The compatibility of the order of application follows from the connectivity of the set operator "∪" (union). The rules defined in Table 1, under *propagating rules* ensures that the matches made by *Tregex* can be manipulated by *Tsurgeon* to reflect unmovability constraints. These rules mark

[*1] In this work, embbeded sentences are not extracted, thus indistinguishable.
[*2] Noun phrase tag is NP, prepositional phrase tag is PP and subordinating clause tag is CP.

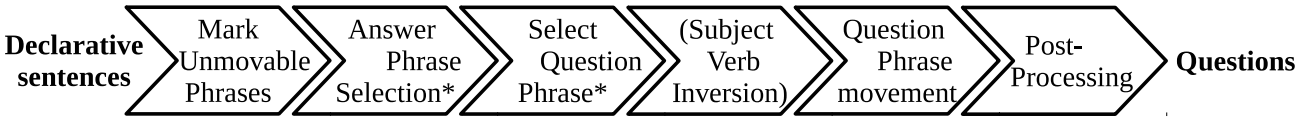

Fig. 3 Describes the operations performed on a declarative sentence until it outputs questions, in question generator stage (1). Some operations may be skipped depending on the location of the *answer phrase*. This diagram is an adaptation of M. Heilman and N. A. Smith (2009)[20].

as unmovable all nodes under an unmovable node and also make sure that phrases having descendants similar to head nodes are marked as unmovable (avoids redundancy).

In the following sentence, *"A capital de Moçambique é Maputo"* <The capital of Mozambique is Maputo>, we noticed that it would be incorrect to generate the question **O que é a capital de Maputo?* <What is the capital of Maputo> because *"A capital de Moçambique"* <The capital of Mozambique> itself is a unique answer phrase in the subject, for that reason "a prepositional phrase appearing in the subject can not be considered as answer phrase" constraint is employed in order to mark it as unmovable.

3.1.2 Answer Phrases Generation

In short, a declarative sentence is converted into the Treebank (hierarchical tree) format and then a set of rules are given to the Tregex tool for pattern matching of the nodes that cannot be considered as valid answers and after matching is done, the Tsurgeon tool manipulates the tree by marking matched nodes with the *UNMOVABLE-* prefix and recurrently the nodes below (child of) the *UNMOVABLE-* prefixed node are also marked as unmovable. In the end, the nodes not marked as unmovable are suffixed sequentially with numbers, i.e. the noun phrases (NP), prepositional phrases (PP) or subordinating phrases (CP). **Figure 4** shows the process decribed in 3.1.1 and 3.1.2 that is performed by our system.

3.1.3 Question Phrase Selection

For each generated answer phrase will be converted into an equivalent question phrase, which will be the head of the interrogative sentence. This conversion does not apply to yes-no questions.

The process of selecting a question phrase is backed up by some constraints that allow coverage on most of the question types. In our system, a declarative sentence is given to the FREELING (L. Padró and E. Stanilovsky, 2012)[23], about 90% of accuracy, for annotating with entity labels. The FREELING tool uses its named entity (NE) module[*3] to recognize and classify entity words (such as PERSON, ORGANIZATION, LOCATION), it also detects time and monetary units.

Our system generates a single (the most likely) question phrase for a given answer phrase, and accordingly the questions are generated. In **Table 2**, we jotted down constraints that may lead to the selection of a specific type of question.

3.1.4 Subject-Verb Inversion

Subject-verb inversion is only performed when the answer phrase is not in the subject and for yes-no questions. In European Portuguese (D.I.Chutumiá, 2013)[24], the interrogatives-*Q* , in general, require a movement of Verb (V) and the alteration of the order of words, except in cases in which we have *D-linked* (Pesetsky, 1987) [25] (Q + N) constituents or interrogatives Q with *é que*.

In Portuguese the presence of clitics (e.g: *me, te, se, etc...*) in verbs is common, in such cases the clitic will always precede its verb after subject-verb inversion has been performed. Any adverb appearing before the verb will also precede the (clitic, if it occurs, and) verb after subject-verb inversion. Our system relies on the power of regular expressions to identify answer phrases that do not appear in the subject, and built-in NLP transformation tools for insertion, deletion and modification of tagged declarative sentences (we described main operations in **Table 3**).

Our system, does not distinguish auxiliar verbs from main verbs due to limitation in Portuguese language tools used by the QG system, performs inversion and provides results in the following order:

- *[<Adverb>][<Clitic>]<Verb><Subject>*.[*4]

For example, from operations performed in this section, the system outputs a question like *"Onde se localiza Moçambique?"* (Where is Mozambique located?), where *'se'* corresponds to the *clitic*, *'localiza'* corresponds to the *verb*, and *'Moçambique'* to the *subject*.

3.1.5 Question Phrase Movement

The selected question phrase is moved to the beginning of the question after subject-verb inversion has been performed, excluding cases in which yes-no questions are being generated. This phase also makes use of regular expressions and built-in NLP functions to achieve the desired and expected result.

3.1.6 Post-processing

In this phase, the system cleans the generated questions in order to ensure proper formatting and turns the questions into human readable form. This also includes correcting contractions and prepositional combinations as well as processing questions ending with adverbs.

For example, the system could generate a question such as *A onde foi o Pedro?* <where did Peter go to?>, where *A* + *onde* = *Aonde* <To where>.

4. Experiments

In **Fig. 5**, provides an illustrative example of the system's question generator process. In this paper, we set our experiments to analyze the output of the Question Generator (1), which corresponds to the initial stage of the overall system described in Fig. 1. In these experiments, we will supply to the QG system a set of sentences and in return we will analyse the generated questions. We are interested in the distribution of the interrogatives-Q and length features of the generated questions.

[*3] NER means named entity recognition.
[*4] '[ ]' means may or may not occur.

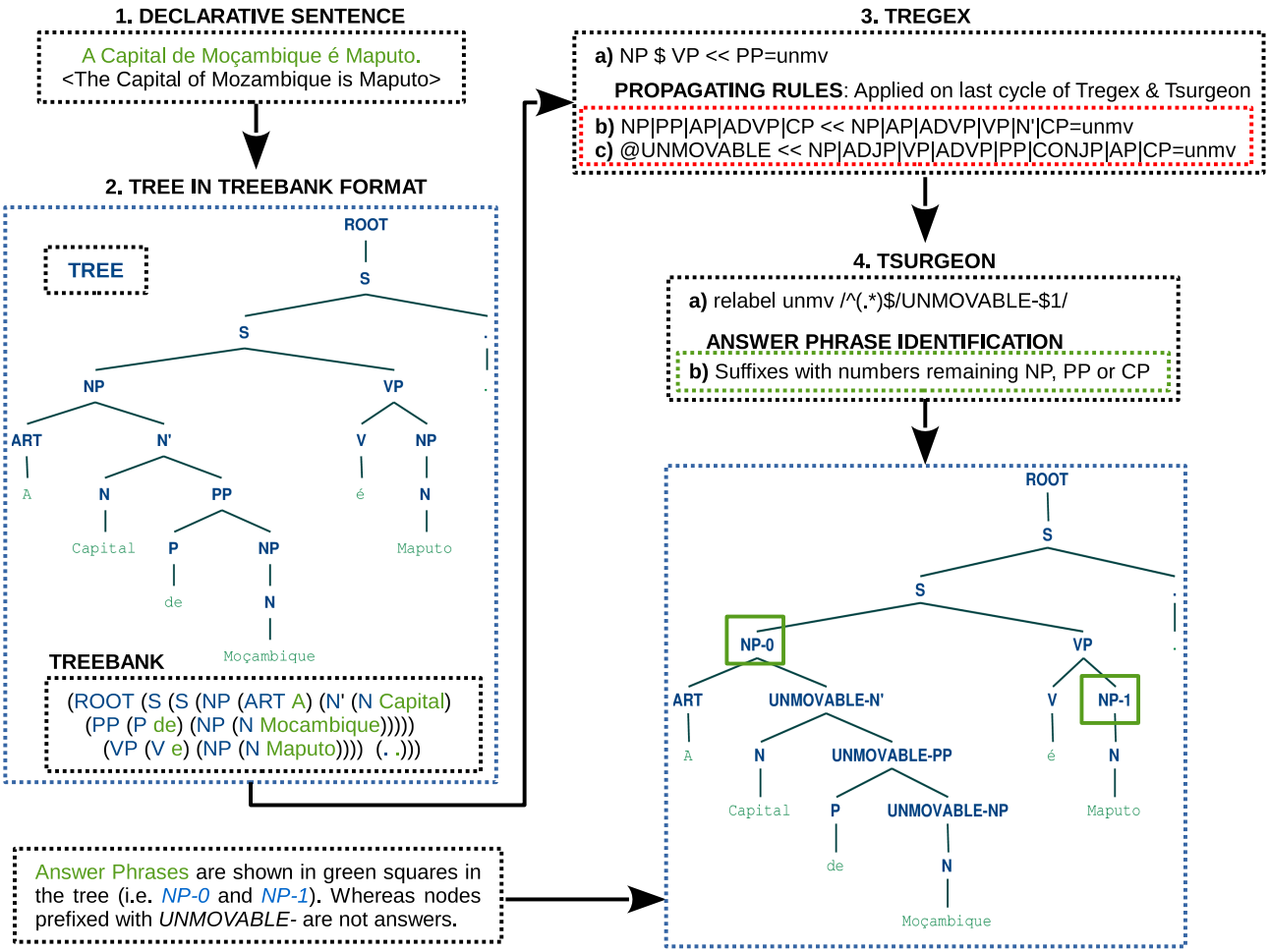

Fig. 4 Illustrates how potential answer phrases are generated.

Table 2 Conditions to generate questions with certain question phrases (* means its absence is irrelevant).

Question Phrase	Possible constraints on answer phrases
Quem <Who>	Usage: used always to ask about persons. Tag(s): noun phrase NER: PERSON *Preposition(s): com, a, de, para
(O) Que <What>	Usage: used to talk about things, undefined fenomenas and seldomly on persons. Tag(s): noun phrase, preposition phrase, subordinating phrase *NER: LOCATION, ORGANIZATION
Qual(is) <What>	Usage: used to talk about things and occasionally to talk about persons; it points out a quality and implies a choice. Tag(s): noun phrase *NER: LOCATION, ORGANIZATION
Quanto(a)(s) <How much>	Usage: it refers to things or persons, whether for countable or uncountable names. Tag(s): noun phrase (NP-$d <(CARD $ /.*N.*/))
Quando <When>	Usage: when refering to time. Tag(s): noun phrase, prepositional phrase Preposition(s): até, desde, para
Onde <Where>	Usage: it refers to places. Tag(s): noun phrase, preposition phrase NER: LOCATION Preposition(s): em, a, para, de.

4.1 Corpora

The dataset consisted of a total of 1323 sentences from news and novels articles, respectively. The sentences were collected from CINTIL-TreeBank (Branco et al., 2011[26]) corpus which contained four corpora.

The first sub-corpus, also labeled as “sentences for regression testing”, contained 266 sentences. This sub-corpus was developed from various sources.

The second sub-corpus was extracted from the CINTIL-International Corpus of Portuguese and contained 444 sentences. The content of this sub-corpus is a combination of sentences from the news and novels domain.

The third sub-corpus contained 43 sentences and was generated from Penn TreeBank (translation), and originated from the news domain.

The fourth sub-corpus was extracted from CETEMPúblico and

Table 3 Describes operations and possible transformations.

Operation	Definition
InSubject	Usage: Determines whether the answer phrase occurs in subject or not. Regular Expression(s): "(*ROOT* (*S* (*NP*–" or "(*ROOT* (*S* (*S* (*NP*–" Return Type: Boolean
Insertion, deletion and modification	Usage: Responsible for transforming the declarative sentence into a question. Algorithm(s): Uses *Quicksort algorithm* to sort the verbal phrase tags. Built-in tools: Uses the power of python lists to perform insertion, deletion and modification and rearrange the sentence in the desired order.

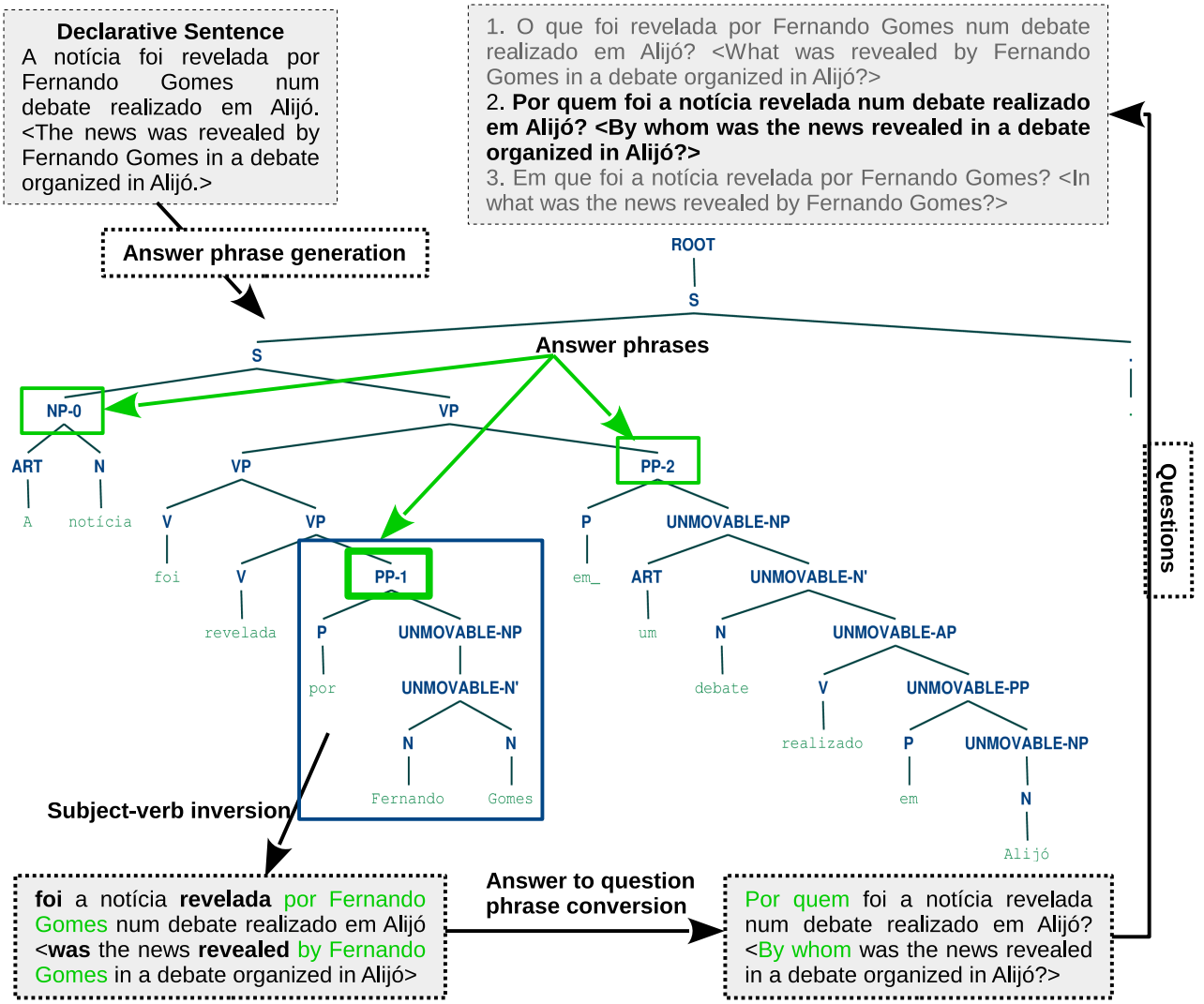

Fig. 5 An illustration of the question generator. The tags *NP-0, PP-1, PP-2* represent the answer phrases, whereas nodes starting with *UNMOVABLE-* tag represent phrases that the QG system did not consider as answer phrases.

Table 4 Corpora used in our experiments.

Sub-corpus	Sentences	Tokens
Sentences for regression testing	266	1415
CINTIL-International Corpus of Portuguese	444	3147
Penn TreeBank (translation)	43	393
CETEMPúblico	570	4707
Total	1323	9662

contained 570 sentences, and compiled from news domain.

The details of the CINTIL-TreeBank corpora are shown in **Table 4**.

4.2 Results of QG

To evaluate results of the QG system, we calculated the lengthy features and analyzed the distribution of the interrogatives-Q. The number of generated questions was 2347, which is 1.77 times the number of the input sentences given to the system.

In **Table 5**, we noticed that the majority of the generated questions are from interrogative word *Que* (What), the reason is that this type of question relies less on *named entities (NE)* and can easily replace a different type of question.

In **Table 6** and **Fig. 6**, we calculated the average parameters for the questions and answer phrases and plotted their respective distribution.

Table 5 The notations (# - Q means number of Questions, # - NE means *number of references to the named entinty*, and # - P. in QP means *presence of preposition in the question phrase*).

Q-Type	# - Q	# - NE	# - P. in QP
Quem	481	481	70
Que	1696	589	347
Qual	29	8	0
Quanto(a)(s)	32	0	0
Quando	15	15	0
Onde	94	94	10
Total	2347	1187	427

Table 6 The average number of tokens in the answer phrase (AP) and in questions (Q) was 1.77 and 6.77, respectively.

Features	Average
AP tokens	1.77
Q - tokens	6.27

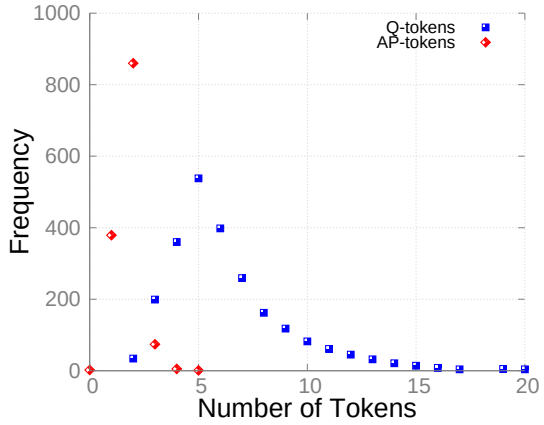

Fig. 6 In the figure, the frequency of questions (Q) with a specific number of tokens and answer phrases (AP) is shown against the average number of questions. We could refer to this as the number of tokens per question and the number of tokens per answer phrase, respectively.

5. Future Works

This work approached the question generation problem with the intent to construct a corpus of portuguese questions generated from declarative sentences. The corpus is characterized by the presence of interrogative-Q in different types of questions.

This paper enabled us to use its results to perform further research in QG systems and develop a question ranking framework that will rank the questions according to their grammaticality or question acceptability. In short, the corpus constructed here will be given raters for manual annotations of question score according to question acceptability and along with features extracted from stage 1 of the QG system, we will develop a model and train, in stage 2, our corpus. That will allow our system to automatically rank questions.

Further research in the QG stage could do better at extending the coverage and accuracy of our *rule-set* to allow complex and different types of questions.

6. Conclusion

We started out by introducing our QG system for Portuguese Language, with the intent to develop an automatic question generation system, then we went on to the implementation, and in the experiments we noticed that our QG system generated questions with a 1.77 ratio to the original corpora.

We also noticed that for interrogative questions of type *Que* <what>, the system generated an overall of 72.3% questions.

The results of our QG system prove that our rule-based approach provides coverage to the types of questions that we defined.

Acknowledgments We thank anonymous readers and the NLP group of Nagaoka University of Technology, specially Professor Kazuhide Yamamoto, for the valuable comments. This work was supported by Nagaoka University of Technology.

References

[1] D. Jurafsky and James H. Martin. 2009. *Speech and Language Processing.* Second edition.

[2] J. Brown, G. Frishkoff, and M. Eskenazi. 2005. *Automatic question generation for vocabulary assessment.* In Proc. of *HLT/EMNLP*.

[3] M. A. Walker, O. Rambow, and M. Rogati. 2001. *A trainable sentence planner.* In Proc. of *NAACL*.

[4] I. Langkilde and Kevin Knight. 1998. *Generation that exploits corpus-based statistical knowledge.* In Proc. of *ACL*.

[5] Wikipedia, a enciclopédia livre. 2016. *Filosofia.* "https://pt.wikipedia.org/wiki/Filosofia"

[6] J. Weizenbaum. 1966. *ELIZA - A Computer Program for the Study of Natural Language Communication between Man and Machine* Communications of the Association for Computing Machinery 9, Massachusells Institute of Technology, Cambridge, Mass.

[7] Echihabi, A. and D. Marcu. 2003. *A noisy-channel approach to question answering.* In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, pages 1623, Sapporo.

[8] Hickl, A., J. Lehmann, D. Moldovan, and S. Harabagiu. 2005. *Experiments with interactive question-answering.* In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL05), pages 205214, Ann Arbor, MI.

[9] Wyse, B., and Piwek, P. 2009. *Generating Questions from OpenLearn study units.* In Proceedings of the 2nd Workshop on Question Generation In Craig, S.D. & Dicheva, S. (Eds.) (2009) AIED 2009: 14th International Conference on Artificial Intelligence in Education: Workshops Proceedings.

[10] Mostow, J., and Chen, W. 2009. *Generating Instruction Automatically for the Reading Strategy of Self-Questioning.* In Proceeding of the 2009 conference on Artificial Intelligence in Education, 465472. Amsterdam, The Netherlands: IOS Press.

[11] Chen, W.; Aist, G.; and Mostow, J. 2009. *Generating Questions Automatically from Informational Text.* In Proceedings of the 2nd Workshop on Question Generation In Craig, S.D. & Dicheva, S. (Eds.) (2009) AIED 2009: 14th International Conference on Artificial Intelligence in Education: Workshops Proceedings.

[12] Schwartz, L.; Aikawa, T.; and Pahud, M. 2004. *Dynamic Language Learning Tools.* In Proceedings of the 2004 InSTIL/ICALL Symposium.

[13] Sag, I. A., and Flickinger, D. 2008. *Generating Questions with Deep Reversible Grammars.* In Proceedings of the First Workshop on the Question Generation Shared Task and Evaluation Challenge.

[14] Yao, X. 2010. *Question Generation with Minimal Recursion Semantics.* Masters thesis, Saarland University & University of Groningen.

[15] D. Diéguez, R. Rodrigues and P. Gomes, 2011. *Using CBR for Portuguese Question Generation.* ISBN: 978-989-95618-4-7. CISUC, University of Coimbra, Portugal, Higher School of Computer Engineering, University of Vigo, Spain.

[16] Santorini, B. 1990. *Part-of-speech tagging guidelines for the Penn Treebank Project.* Technical report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.

[17] D. Chen and C. D. Manning. 2014. *A Fast and Accurate Dependency Parser using Neural Networks.* In Proc. of *EMNLP*.

[18] J. Silva, A. Branco, S. Castro and R. Reis. 2010. *Out-of-the-Box Robust Parsing of Portuguese.* In Proc. of *9th ICCPP*.

[19] R. Levy and G. Andrew. 2006. *Tregex and Tsurgeon: tools for querying and manipulating tree data structures.* In Proc. of *5th ICLRE*.

[20] M. Heilman and N. A. Smith. 2009. *Question Generation via Overgenerating Transformations and Ranking.* Language Technologies Institute, Carnegie Mellon University Technical Report CMU-LTI-09-013.

[21] D. M. Gates. 2008. *Generating Look-Back Strategy Questions from Expository Texts.* Language Technologies Institute, Carnegie Mellon University.

[22] Wikipedia, 2016. *Portuguese Language.*
[23] L. Padró and E. Stanilovsky. 2012. *FreeLing 3.0: Towards Wider Multilinguality.* In Proc. of the Language Resources and Evaluation Conference (LREC 2012) ELRA. Istanbul, Turkey. May, 2012.
[24] D. I. Chutumiá. 2013. *As Interrogativas-Q do Português de Moçambique: Contribuição para uma análise comparativa com o Português Europeu e o Português Brasileiro.* Faculdade de Letras, Universidade do Porto.
[25] Pesetsky, David. 1987. *Wh-in-situ: Movement and Unselective Binding.* In: Reuland & ter Meulen (eds.), *The Representation of (In)definiteness*, Cambridge, Mass.: CUT, 98-129.
[26] Branco et al. 2011. *CINTIL-TreeBank.* Departamento de Informtica NLX - Grupo de Fala e Linguagem Natural, Faculdade de Ciłncias da Universidade de Lisboa, Edifcio C6, Campo Grande 1749-016 Lisboa.