Automatic Generation of Questions in Romanian

Mihai Manolescu
mihai.manolescu@student.uni-tuebingen.de

Supervisor: Prof. Dr. Detmar Meurers

A thesis presented for the degree of
Bachelor of Arts

Seminar für Sprachwissenschaften
Eberhard-Karls Universität Tübingen
Germany

October 2019

Hiermit versichere ich, dass ich die vorgelegte Arbeit selbstständig und nur mit den angegebenen Quellen und Hilfsmitteln einschließlich des WWW und anderer elektronischer Quellen angefertigt habe. Alle Stellen der Arbeit, die ich in anderen Werken dem Wortlaut oder dem Sinne nach entnommen habe, sind kenntlich gemacht.

Mihai Manolescu

Acknowledgments

First of all, I like to thank my supervisor Prof. Dr. Detmar Meurers for showing me what this topic is about and helping me throughout the project. Furthermore he always tried to give me more ideas of what can be done. I would also like to thank Dr. Raluca Brăescu, lecturer of the department of linguistics at the University of Bucharest, for her insights on the Romanian language. And of course I need to thank my parents Radu and Anca, as well as Judith for their constant support over the whole period. Without you this would have not been possible to manage.

Abstract

This thesis shows a first approach towards an automatic question generation system for the Romanian language. The goal of the project is to be able to generate questions for the subject, object and root of a sentence automatically, based on the part-of-speech tags of the sentence and constructs the question following different generated templates, which are based on the syntax of a correct question in Romanian. The performance of the system is tested on four different Romanian texts with various language difficulty. In general, the system performed good for the fact that it is a first approach for this language.

Contents

1 Introduction

With the technological influence of the last decades, our world changed remarkably. The power of networking enabled more and more ideas to become reality and offer products and services globally. One challenge of linguistics is how to analyze a language better with the use of modern day technologies. Since the tools we use are getting better, it is possible for us to examine the characteristics of a language further and discover new features. Different studies in the field of computational linguistics show what fascinating capabilities there are to analyze different elements of a language and how these can help people in their daily life. One topic of great interest in Computational Linguistics is the automatic generation of questions based on texts in a language. Many different approaches to question generating systems have already been developed and have shown how efficient the different techniques used really are. Most of these systems were designed for the English language, which is good since English is the internationally most spoken language. But what about developing a system for a language spoken by a smaller group of people? The goal of my thesis is to show how a question generation system can be developed for Romanian, which is one of the five romance languages and for which, so far, no real approach has been done in this area. Therefore this thesis and project is the first and shows how such a system can be built for a more uncommon language. This thesis is split up in four different sections. It starts with background information about the Romanian language and question generation systems in general. The third section describes the resources and the approach to this topic and results are presented in the fourth section afterwards. Finally, the last segment discusses what needs to be done in this field to be able to improve the system.

2 Background

There is much information that can be found for the Romanian language and in order to understand how the system works, some language knowledge is introduced in this section as well as some basic information about question generation systems in general.

2.1 Language Background

Romanian is one of the five main Romance languages spoken today and belongs to the Balkan-Romance group. It is currently spoken by 23.4 million people (Ethnologue contributors, 2018). It's also the only Eastern-European Romance language left. Being a descendant of spoken Danubian Latin, the structure hasn't changed much from its origin. All non-Latin elements were transformed by the Romance pattern to correspond to the common form of Romance languages. As a literary lexis indicates, about 43% of Romanian words are borrowings from other Romance languages, mostly French. Another 20% are inherited Latin words, which is due to its descendance from Danubian Latin. Romanian has its own phonological features, that differ from other Romance languages (Pană Dindelegan, 2013).
The verb is one of the key features in the Romanian language. Romanian contains a high number of inflectional classes and sub-classes, that are characterised by suffixes and syncretisms and morphological alternations. There is, compared to other languages, a relatively small number of irregular, suppletive verbs. Another specific aspect is the fact that it has a complex of modal forms. It distinguishes between real and unreal, but also between assertive and injunctive with a mixed paradigm. Romanian shows a tendency towards a specialization and supplementary marking of epistemic modal values and evidentiality. The language has a class of verbs with two internal arguments, where it distinguishes a sub-class that is only taking a direct and indirect object, i.e. *'a întreba'* - to ask, and another sub-class that takes a direct object and a nominal phrase with an unmarked form, that is seen as an argument, called secondary object, i.e. *'aducere aminte'* where the verb is nominalized *'a aduce'* - *'aducere'*. As mentioned before, Romanian is based on Latin and therefore keeps specific components of the language. The classes of Latin verbs, which take two accusative objects, is preserved (one direct and one secondary object or one direct object and an objective predicative complement). The secondary object displays an unmarked case form (accusative - nominative).

The focus of my thesis is on generating questions for different aspects of a sentence and taking a closer look at subject and object in Romanian.
The subject in the Romanian language is very complex and difficult to define as one. It can be formed in many different ways which can also be misleading, even for native speakers.

(1) Îmi place muzica.
I like music
'I like music.'

(2) Nu-mi convine tratamentul
not=CL.DAT.1SG like treatment.DEF
'I don't like the treatment.'

There are many types of words that can be seen as a subject in a Romanian sentence, from a simple noun to a pronoun to even numerical values and even a whole syntactic group. The variability is very wide compared to other languages. There are a lot of problems, especially for taggers in Romanian, to find the correct subject in a sentence, since it can cause confusion and many ambiguities can be found. It is difficult to detect if a given word is the subject or an direct object at first. The sentence (1) is already a problem. Most of the people will instantly say that the pronoun *'Îmi'* is the subject, but this is wrong. Since the correct question to ask for a subject is 'Cine = ce?' (Who or what?), the correct subject in the sentence above is 'muzica'. An additional example of confusion is the sentence (2), which contains an intransitive verb 'convine' that can not have a direct object. Therefore, again like in the previous example, the correct subject is not the pronoun but rather the noun 'tratamentul'. The pronoun is a clitic form of the dative and therefore can not be the subject in a nominative case. Another aspect is the multiple subject in Romanian. It is commonly used and is difficult to be seen as the subject for computers, since it often only takes the first word of the multiple subject and tags it as the subject, as can be seen in (3). It happens fairly often that only the first word 'Copacii' is treated as the subject and the other parts are not processed in any way.

(3) Copacii, florile și animalele se
Trees.DEF, flowers.DEF and animals.DEF it=CL.REFL.ACC.3SG
bucură de vremea minunată de primăvară.
enjoys of weather wonderful of spring
'Trees, flowers and animals enjoy the wonderful spring weather.'

A further crucial point in subject recognition in Romanian is the unexpressed subject. The fact that the presence of the subject is not obligatory is characteristic of Romanian; its information can be contextually retrieved or impossible to (fully) recover. The feature known as 'null pronominal subject' / 'pro-drop language' is strictly related to the rich verbal inflection, which allows the verb to take over totally or partially, through agreement, the information encoded by the subject. Romanian displays the three characteristics which distinguish the pro-drop languages (subject non-realization, free subject inversion, extraction of the subject from the subordinate clause (Pană Dindelegan, 2013). Romanian differs from French and some northern Italian dialects, where the presence of pronominal clitics in subject position is usually obligatory, but resembles Portuguese, Spanish, varieties of central and southern Italian, in which the pronominal subject is not realized. This is a pretty common feature in the Romanian language, since many sentences are being shortened and therefore it becomes unclear where the subject really is. It needs to be deduced by neighboring sentences or from the context (Gramatica contributors, 2019).

(4) Să fim serioși!
Să.SUBJ be.SUBJ.1PL serious.M.Pl
'Let's be serious!'

(5) A intrat profesoara în clasă. Pare supărată.
Has entered teacher.DEF in class Seems upset
'The teacher entered the classroom. She seems upset.'

This form of subject is split into the included subject and the implied subject. In the included subject other parts of the sentence have the subject included in the word or as a context for the whole sentence. Sentence (4) has no direct subject given, but it implies that it mentions all of us and therefore the answer to the subject question would be 'noi' (we). For the implied subject on the other hand, the subject was already previously mentioned and is therefore unnecessary to be used again, as can be seen in (5). If only the second sentence would appear, there would not be any indication about the subject in this sentence. These are only a few cases of how a subject can be formed in the Romanian Language, but there are way more types of it. This thesis only covers a few to show examples for what has been programmed in the Question Generation system explained more detailed in the next section.

The direct object in Romanian can be found in a number of different settings. As it can be a part of the question generated by the system, it first needs to be discovered as a whole, which is not easy for parsers or taggers.

The direct object is usually selected by a transitive verb, which can have a finite or non-finite form (Pană Dindelegan, 2013). In Romanian this form of object can be constructed in a large number of ways. Only easier examples were used for this project in order to show what impact a direct object can have on a question generated by a system and since this is a first step into question generation for Romanian, it can be expanded in the future. Direct objects can be formed in different ways and can have a number of different constructions. Most common forms are prepositional constructions, which often describe the given circumstance clearer. The preposition 'pe' (on) is one of the most used in this case but needs to be distinguished with the direct object marker 'PE'. Romanian displays specific direct object marking – the PE prepositional marking, which has supplementary conditions: semantic (the feature [+specific]), lexical (the feature [+human]), pragmatic (the object's 'high' degree of prominence). Romanian is part of the group of languages characterised by Differential Object Marking. However these two forms share a number of features. Both select the accusative case and like other Romanian prepositions, it can block the definite article from occurring with unmodified nouns, but the noun phrase has a definite and specific reading. It also allows a relative clause to be its complement (Pană Dindelegan, 2013).

(6) Bogdan merge.
Bogdan walks
'Bogdan walks.'

(7) Bogdan bea apă.
Bogdan drink water
'Bogdan drinks water.'

All of the previous mentioned elements of the Romanian language play a role in how questions are formed in Romanian. The usual and most simple form of a question is ***Question word - root***. This form can be expanded with different kinds of information, like direct objects, adjuncts and many others. Romanian has different question words for specific cases. Questions follow specific rules in Romanian. First, all questions need to have one of the interrogative pronouns and adverbs, e.g. *Cine?, Ce? , Unde?, Când?*. Secondly the word order is fixed with the interrogative words being always at the beginning of the sentence. Furthermore in all total direct interrogative cases, the natural topic places the predicate enunciation first where in contrast in indirect cases the order of the constituents changes. This leads

to the subject, which is, if available, placed after the root element of the sentence. Another important thing worth mentioning is that the preposition, if there is one in the original sentence, is always placed before the interrogative word (Șerbanescu, 2002). Subject requires ***Cine*** (who) as a question word, with ***cine*** being the most used for questions like these. It asks directly for a person, a number or just a pronoun. Sentence (6) is a great example, where the surname *Bogdan* is the subject of this sentence and therefore the correct subject question would be ***Cine merge?*** (Who walks?). The second most common questions are asking for the object. This requires the interrogative word ***Ce*** (what). An example for the complement here can be seen in (7), which leads to the question '**Ce bea Bogdan?**' (What does Bogdan drink?).

This pattern of generating questions with a specific structure is not possible when asking for the root of a sentence. For this matter, there are two different possibilities how to ask for the verb in a sentence. The first one, which is applicable as a general question where the answer is either the whole sentence or just the core of the sentence - subject and root - is '**Ce se întâmplă?**' (What is happening?). The second option to ask for the root is to ask what the subject is doing. For the sentence (7), the question would be '**Ce face Bogdan?** (What does Bogdan do?). If the verb can stand by its own, then the answer to this question would be solely the root, if this is not possible then the subject needs to be included in the answer sentence. All of the mentioned features of the Romanian language will be analyzed in the following section.

2.2 Automatic Question Generation

Automatic question generation has been an interesting topic of Natural Language Processing for many years. The amount of research done in this field is very high, since it can be split up into multiple ways to generate questions automatically. As mentioned in *Automatic Question Generation: A Systematic Review* by (Soni et al., 2019), the paper by (Yao and Zhang, 2010) divides the subject into three different categories: Template-based QG, Syntax-based QG and Semantic-based QG.
Template-based systems purely work on the text and do not take syntactic or semantic information into consideration and construct and generate questions based on templates specified by the author. This approach can be used to create different simpler systems, which therefore generate less specialized questions.

Syntax-based approaches are the most common when designing a QG - System. The main idea here is to convert the declarative target sentence into an interrogative sentence by manipulation of a derived syntactic tree (Soni et al., 2019). One of the most important examples for this approach is Heilman's PhD thesis *Automatic Factual Question Generation from Text*. He focuses on generating factual WH-questions from an input text and generate questions to asses the reader's knowledge of the information in the text (Heilman, 2011). This is also one of the most important papers regarding QG, since it shows different methods to create such a system and also states the difficulties of them.
Semantic-based question generation systems are focused on semantic parse to generate questions from text. Therefore the questions are mainly factoid-based and systems can not generate questions of very good quality(Soni et al., 2019). Xuchen Yao and Gosse Bouma present a question generation system based on semantic rewriting. Their system works by understanding the meaning of the sentence and generating questions that are semantically purposeful (Yao et al., 2012).
To be able to generate deeper questions, these systems are merged together and especially the combination of syntax-based and template-based is very strong (Soni et al., 2019). These are only a few examples of how question generation systems are built, but there are of course other notable mentions, like (Piwek and Boyer, 2012) who also provide a good overview of the different methods for question generation. Question generation system are not only built for English, but also for other languages of the world and one of the most elaborate and complex ones was done by (Kolditz, 2015) for the German language. Although many languages tackled this topic, there is no real research done for Romanian, hence it is difficult to construct a system from scratch. The Romanian Automatic Question Generation System combines elements of a template-based and a syntax-based system to construct questions for the subject, the object and the verb from input text, which is shown in the next section.

3 Romanian Automatic Question Generation System

3.1 Resources

As basis of my project I am using the NLP-Cube (Boroș et al., 2018). NLP-Cube is a system created by Tiberiu Boroș and Ștefan Dumitrescu from the Research Institute for Artificial Intelligence 'Mihai Drăgănescu' which is part of the Romanian Academy located in Bucharest, Romania. Their project is an open-source Natural Language Processing Framework with support for various languages which are included in the Universal Dependencies[1] (UD) treebanks list. NLP-Cube provides a series of tools for language processing, which are necessary to be able to generate questions. Such tools are sentence segmentation, tokenization, part-of-speech tagging, lemmatization and dependency parsing. All these tools are used for my project and the generation of questions is done by using the part-of-speech tags, since there is a specific structure how a question needs to be formed in the Romanian language. To be able to parse and use text files, I am making use of the Natural Language Toolkit [2] (NLTK) library. NLTK is used to split sentences from a text file and pass them on to the NLP-Cube. Unfortunately there is no model for Romanian and therefore I am forced to use the English model for sentence splitting, which works very well.

3.2 Approach

The general idea of the project is to generate three different types of questions. The main questions in a language ask for the subject and the object of a sentence, since those are the most important parts. As mentioned in the background section, the system can also generate questions about the root of the sentence.
As can be seen in Figure 1, the idea is to extract the words and their given label into separate lists and generating a question by creating the questions firstly with the tags to create the correct form of a question in Romanian and then getting the matching word for each label from the list and constructing the question. Since the answer is also part of the list of words, it needs to be extracted separately for each type of question.

[1] https://universaldependencies.org/
[2] https://www.nltk.org/

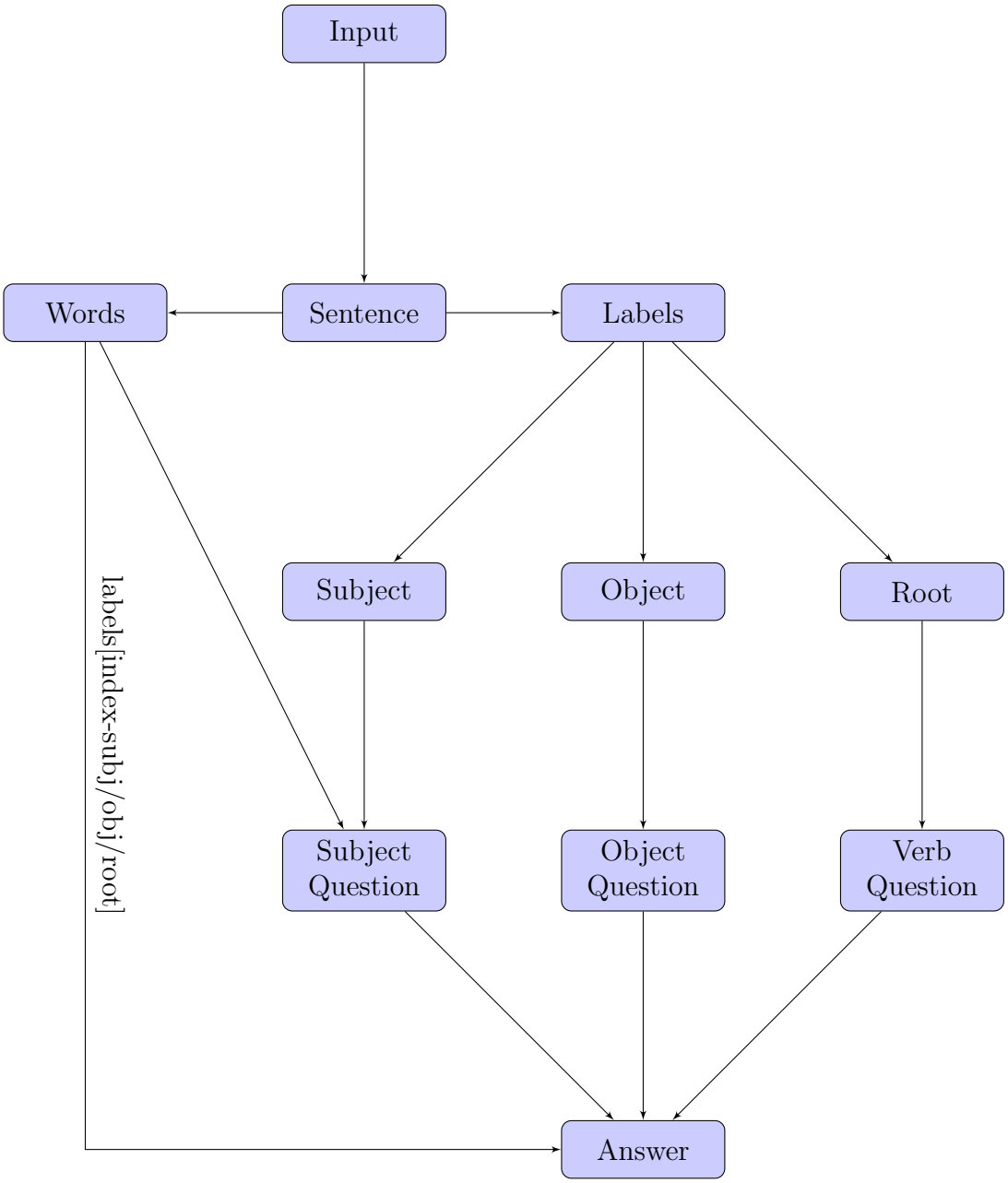

Figure 1: Construction of how the question generation works

Starting off with generating questions for the subject, the main form of the question in Romanian is defined as: '***Question word – verb – (object)?***'

To be able to always have access to the parts needed to generate such questions, the idea was to work with part-of-speech tags to be able to create the question for every possible sentence. NLP-Cube provides as output different information about the given word including its part-of-speech tag as well as the dependencies to the other words in the sentence. Since the subject is the important part here, the idea is to remove the subject from the sentence, store it separately and construct the question using the remaining elements of the sentence by positioning them in the correct way of how a question should look like in Romanian. There are, as mentioned before, many different cases of how a subject can be formed in a sentence. The idea is to take some of these cases and generate questions with the given resources and possibilities. Starting with simple cases, I looked at the easiest form of a sentence, which is: '***Subject + verb***'. Generating a question for this form proved to be rather easy, since there are only two elements and the question can be formed by putting the root of the sentence, the verb, after the correct question word for subject questions - *Cine*. This was the standard case of a correct question, from which I expanded to be able to generate questions for more specific and complex sentences. First an object was added to an easy sentence to see, how the system would tokenize it and to be able to create a slightly more complex question. The NLP-Cube identified the given argument as object or in some cases as an oblique nominal, which is normally used for a nominal functioning as a non-core (oblique) argument or adjunct (Universal Dependencies contributors, 2014-2017). When looking at generating subject questions with an object in the input sentence, the creation proves to be easier than expected. Using the same method as before, I start off by creating the base form of the question, but now adding the given object after the root. The same approach is used for questions including an oblique nominal. These usually also have a determiner associated with them, which also needs to be part of the final generated question. Therefore the idea here is to look at the labels in the list and generate the question following a specific pattern. First the question word is added followed by the root element. If there is a determiner associated with the object or the oblique nominal, it needs to added after the root element. Otherwise the object is being added after the root and thus completes a correct Romanian question.

One special feature that was already mentioned in the beginning of this paper is the multiple subject in Romanian. This special case is found pretty often in the normal day to day use, but parsers are unfortunately unable

to detect the multiple subject and tokenize it as one big subject. I tried to tackle this problem in different ways and were only successful in detecting a multiple subject containing two elements, since it's the only way to work with the output given by the NLP-Cube. The parser tokenizes such a feature in different ways. One of the options can be a collection of different coordination elements (UD: *cc*) elements with corresponding conjuncts (UD: *conj*).

(8) Bunica și mătușa...
Grandmother and aunt...
'Grandmother and aunt...'

(9) Șeful statului a semnat decretele.
Head state to sign decrees.
'The head of state signed the decrees.'

An example herefore is shown in (8). The parser identifies the subject, '*bunica*', correctly but does not add the following two words, '*și mătușa*' as well to the subject. Therefore I manually take the corresponding part of the multiple subject and store it with the subject from the parser as the correct answer to the subject question. The question here follows the same general pattern of questions as before, but now has the complete multiple subject as an answer to it, which without splitting all the elements of the multiple subject and saving them as one big subject would result in a messy question as well as in an incorrect answer.
Another specific case of the Romanian language are different nominal modifiers after the subject. As in the case of the multiple subject, the parser is not able to recognize it directly and again have to manually add it to the given subject. This is the case for different sentences, an example here would be (9). The nominal modifier *statului* is part of the subject and does not need to be put in the question for this sentence. Therefore nominal modifiers of the subject are removed from the list of words and saved with the subject of the sentence, as the answer to the question that is being formed like in previous cases.

After having successfully implemented the different cases of subject, the next point was to tackle the questions that need to be generated for the direct object. The approach is very similar to the one mentioned before. The first thing is to check if the sentence even has an object by looking for the part-of-speech tag in the list of tags generated for every sentence that is being parsed. If this is true, then the easiest question that can be generated that asks for the object is '**Ce - root - subj?**. Depending on the complexity of

the given sentence, this simple form of the question can be expanded without any problem. Since sentences can be pretty complex, the system first checks for the root and, if found in the sentence, an auxiliary of the verb. If the auxiliary is not placed before the root in the question for the object, then the question would not make any sense and also the answer would be messed up. The resulting form of the question is '**Ce - (*auxiliary*) - root - subj?**'. As well as before for the subject, determiners or nominal modifiers of the subject need to be included in the question. We check for these additions of the subject and if there are found in the sentence, they need to be added before the subject in the interrogative sentence. The form of the question can be expanded to '**Ce - (auxiliary) - root - (*determiner*) - subject - (*nominal modifier*)**'. The answer to these questions is stored before creating the question and separately in a variable. Like the subject, the object can also have determiners and other extensions. If a determiner is present, it is taken away from the sentence and stored with the object since without it, the question would not make sense. If none of the above can be found, then the object is stored as one simple element.

As beforehand mentioned, asking about the root of a sentence can be done in two different ways. The way this works is again similar to the subject and object questions, only that this time the questions are more manually created since the correct words to ask about the root of the sentence can not all be found in the sentence itself. The first question that is asked is what is happening in general. Here the correct answer is the complete sentence, which of course includes the root, which gives the important information back to the user. The question here created by the system is '**Ce se întâmplă?**'. The second question that can be generated with more impact from the source sentence is including the subject from the original sentence in the interrogative sentence. The system takes the verb a face (to do) and generates the question '**Ce face**' and adds the subject from the original sentence. The answer for this question can either be the root element of the sentence by its own or combined with the subject of the source sentence, to get the minimal information from the sentence.

The system works with text files that can be set by the user with content selected by him. The program reads the file in and separates the sentences. Then each sentence is being analised by the NLP-Cube. All of the questions are being generated if at least the subject is found in the source sentence. The program is built around it and whenever it does not find a subject in the sentence, which is not such an uncommon occurrence in the Romanian language, the system prints out a missing subject message and continues on

with the next sentence. The output for every sentence is constructed the same for any sentence. First of all, the source sentence is being printed to give an overview over the sentence with the question asking for the subject following it. After these, the two different questions asking about the root of the sentence are shown and last but not least the question for the direct object. If no direct object was found in the given sentence, then the last step will be skipped and the system continues with the analysis of the next sentence in the source file.

4 Results

To test the question generation system, different text files with several difficulties of the sentences were created. Since this is a pioneer project, the system can identify and create questions better for clearly structured sentences, but can also generate questions for more complex constructions as we will see in the following.

The first test file contains easy and clearly structured sentences to be able to not only test the different specific cases of either the subject or object, but to also see how good the system performs on simple sentences. The simple form of a sentence is **Subject - root**. With this in mind the first example is '*Ion merge.*' (John walks.), which gets correctly tokenized by the NLP-Cube and the questions are being generated fast and correct. The question for the subject in this case is '*Cine merge?*' (Who walks?) and for the root both options are available and valid with the first one being '*Ce se întâmplă?*' (What is happening?) with the answer being the whole sentence, since it is such a short sentence. The other question generated is '*Ce face Ion?*' (What is John doing?) and the corresponding answer in this case is just the root, since it is clear what the action is. With the knowledge that these simple sentences work fine, the basic sentence is extended by adding an object at the end of the sentence. The example sentence here is '*Ion mănâncă mămăligă.*' (John eats polenta.) and the system has again no problem generating the correct questions for the subject and the root. With the NLP-Cube correctly recognising '*mămăligă*' as an object, the QG system constructs the correct question for it - '*Ce mănâncă Ion?*'. To be able to test how good the system recognizes the correct object, we test a sentence that has an extended object and is not only composed out of a single word. The new given sentence is '*Maria are grijă de Ion.*' (Maria takes care of John.), in which the tokenizer recognizes the term *6*grijă as the object, although the rest of the sentence after this word is also part of the direct object and is being manually added

by the question generation system to the previously tokenized object and builds the correct answer to the generated question: '*Ce are Maria?*'. Another specific thing that is not easy for the parser is the multiple subject. We tried to see what the subject will be in a sentence with a clear case of multiple subject. The analysis sentence is '*Bunica și mătușa au venit din nou la mine.*' (Grandmother and aunt come to me again.). NLP-Cube tokenizes only Bunica as the subject, although the correct grammatical form would be 'Bunica și mătușa. The system corrects this and saves the second part of the subject to the subject recognized by the tokenizer and builds the correct subject question: '*Cine vine din nou la mine?*'. As for the last example, the idea is to add different elements like complements and determiners to the subject as well as to the object in a more complex sentence. We created the sentence '***Un șef de bancă scoate o hârtie de 50 de lei.***' (A bank manager pulls out a 50 lei bill.), where the different elements need to be edited by the question generation system to get the correct forms of the subject and the object, since both of them have determiners, nominal modifiers and other parts which are not directly recognized by the NLP-Cube. This shows that the NLP-Cube is capable to detect the elements, but due to the difficulty of the Romanian langauge it can not detect the complete subject and object. The correct subject question generated by the system after cleaning up is '*Cine scoate o hârtie de 50 de lei?*' with the whole correct answer '*Un șef de bancă*'. The correct object question therefore generated by the system is '*Ce scoate un șef de bancă?*' and the corresponding answer '*O hârtie de 50 de lei*'. These were only a few example sentences that were examined in order to understand, which specific cases needed to be improved and have an overview of how the system actually performs on these sentences. Overall about 90% of the questions were correct with only a few questions having small grammatical mistakes or false tokenization of parts of the sentence. Since these were specifically made to test the system, this high number of correct questions is no surprise at all, even if the Romanian language contains some very specific cases.

The second file tested consisted of a small text from one of the biggest Romanian newspapers *Romania Libera*[3]. It consisted of multiple sentences, which have a more complex structure than the first text file and consists of more challenging sentences, since it is normal spoken Romanian. The system begins to struggle with these complicated sentences, which also shows in how the questions are formed for the subject and the object. For five out of the seven sentences given in the input file the system generated the correct question

[3]https://www.romanialibera.ro/

for the subject. An example sentence from the input file is '*După ciocnirile de duminica trecută, poliția a decis să interzică manifestații, avertizând populația contra riscului de grave incidente.*' (After the clashes last Sunday, police decided to ban demonstrations, warning the population against the risk of serious incidents.). In comparison to the sentences mentioned in the first input file, there is a clear difference. Nevertheless the system tokenizes the subject correctly - *poliția* - and generates the correct question for it: '*Cine a decis să interzică manifestații avertizând populația contra riscului de grave incidente?*'
The two sentences for which no correct question was formed have the same pattern and also the problem why the correct question is not formed is the same. NLP-Cube recognises two forms of the subject, the normal nsubj and nsubj:pass. This is a subtype of the normal subject and can be found in some languages that have a grammaticalized passive transformation. This is not the common case and the question generation system works with the normal subject and therefore tokenizes another part of the sentence as the subject. In the sentence '*Pentru a trece peste această interdicție, apeluri au fost lansate pentru organizarea de adunări religioase, care nu necesită autorizații.*' (To overcome this ban, appeals have been launched for the organization of religious assemblies, which do not require authorization.) the subject recognized by NLP-Cube is the word ***care*** and the subtype of the subject is ***apeluri***. This occurence is pretty uncommon and is not being treated further by the system.
Because of the misunderstanding with the two different forms of the subjects, the object questions for the same sentences are also generated wrong, but for the first example the system generated the correct question. The questions regarding the root element of the sentence work very good, but it also follows a very simple pattern. Since the patterns are clearly defined by the system, such specific are not taken into consideration yet and therefore the questions are poorly generated. All in all the question generation system performed good for the difficulty of the sentences and their complexity.

The third test file contained sentences from a story for children. The story is called 'Ursul păcălit de vulpe' and was written by Ion Creangă, who is a famous Romanian writer and one of the main figures of the Romanian literature. The reason behind choosing a story for children is that it should contain sentences that are written in a way for children to understand the context and learn grammatical features at the same time. There were fourteen sentences chosen from the story and the question generation system performed again well. For nine out of the fourteen sentences correct subject questions were formed. What was new for this test file is that it contained sentences

with no subject at all. This is not such an uncommon thing in the Romanian language, since the subject can sometimes be found in one or two sentences before the current one and therefore it does not need to be mentioned again in the sentence that is currently being parsed and tokenized. Only eight sentences also contained an object and the system was able to generate four correct questions. A good example from this input file is: '*Atunci ea rădică puțin capul și, uitându-se la vale, în lungul drumului, zărește venind un car tras de boi.*' (Then she raised her head slightly and, looking down at the valley, along the road, saw an ox cart coming.)
The subject, which in this case is the pronoun ***ea***, is being correctly recognized by the tagger as well as the object ***capul***. The resulting generated questions by the system are '*Cine rădică puțin capul și uitându-se la vale în lungul drumului zărește venind un car tras de boi?*' for the subject and '*Ce rădică ea?*' respectively for the object. A point worth mentioning is the root in this case. The correct verb form should be '*ridică*, but the parser understands this older form of the verb and is able to correctly tag it as the root. Using older types of verbs is a common feature found in older literature in the Romanian language and in some cases is not very easy to understand. Testing the system with sentences from a children story proved to be more difficult than expected, since there are some specific cases that can not be tested very well yet. To be able to correctly generate questions for sentences without a subject, the context of the analyzed sentence needs to be taken into consideration, so that the system can try to generate the correct forms of the question.

For the final test file, the chosen text is written by Ion Luca Caragiale, who is like a Ion Creangă one of the most important people of the Romanian literature. We extracted nine long sentence from his work *'Mamă...'* [4] to test the question generation system on a more advanced text and collect useful information about the performance. The system generated correct subject questions for four out of the nine sentences in the input file. This is not a surprise, since these sentences are more advanced when it comes to their complexity and use of language. The interesting part here is nevertheless the tagging of the word ***să*** in the sentence '*Dar să știi ca la-ntoarcere trebuie să găsesc și copil in casă; de unde nu, pâine și sare cu mine nu mai mănânci!*' as the subject of this sentence. This is like in the previous test file a case of the subject being told one or two sentences before the given sentence and the word ***să*** is here used to replace the already beforehand given information. This is rather common as in Romanian, people tend to shorten up their sen-

[4]https://ro.wikisource.org/wiki/Mamă...

tences when the other person knows about the context already. Questions about the object were correctly generated for the two sentences out of the five that had an object. As in the examples from Creanga, there again are sentences without a subject at all and therefore the generation of questions is not possible. When looking at the questions for the verb, both methods work, as they also have for all the other test files, very good. Since the length of the sentence is significantly longer than in the other files, it is no surprise that the system does not do as good of a job as it has for the other test files. Nevertheless the system performs relatively good when it comes to such complex sentences, when it is made of clear parts and the sentence is clearly structured.

All in all the question generation system performs well on a number of different sentences with diverse structure, complexity and language used (both older and newer versions of words). It takes care of some of the specific aspects of the Romanian language very well and can identify and modify the answers from the input sentence, so that the answer is structured correctly for most of the sentences. Creating such a system for a complex language like Romanian is not easy and can be improved with a number of different features, which will be discussed further in the next section.

5 Discussion

This system is one of the first of its kind for the Romanian language. Since it is a basic project, there are many things that can be done to make the system better and more efficient. It can be used as a pilot project and on its basis, an improved version can be created with more complexity and structure. First of all, this system works for subject, object and root questions with a small selection of patterns for these forms. Seeing that Romanian has a high number, as mentioned in the background section, of how for example a subject can be formed in a sentence, there are more patterns that can be implemented and these that are already implemented can be improved to be able to detect almost every form in every sentence, no matter how difficult the sentence is. Romanian is not an easy language with easy to understand forms and therefore matching every possibility of the forms will always be a problem, when trying to perform Natural Language Processing methods on Romanian sentences. There are of course other parts of the sentence for which questions can be generated. Aspects of time, mode or place are other structures that can be interesting to create questions for, especially if such a system is used in an educational environment, where the students are just starting to learn the language. Beside the usual structural forms of a sentence, it could be useful to generate general questions about a text in general, to see if the student understands the message or the core of the input. Elements of all the mentioned types of question generation systems can be used to also create questions based on semantics and syntax to have more context.

Another point worth mentioning is that there is a high number of question generation systems for English and other more common languages. Unfortunately there is a lack of essential tools for more uncommon languages. A quick search shows the lack of support for this language. Common tools like the Stanford CoreNLP [5] or any NLTK tools do not provide support for Romanian. This system uses the the English sentence tokenizer from the NLTK package, which fortunately works relatively good also for Romanian, but is not completely reliable, if the system is more complex and tries to understand full multi-page input texts. Mainly speaking, this brings more difficulties in finding libraries or resources to create a question generation system from the core. To be able to have the same quality of systems like for English, these tools need to provide support for uncommon languages, so that everyone can have the possibility to create such system for their language, if there is interest.

[5] https://stanfordnlp.github.io/CoreNLP/

I believe this specific field of question generation for Romanian is very broad and can further developed. This project can be therefore used as a base for a more complex system in the future, with more complexity and more knowledge which is needless to say an interesting field in Romanian computational linguistics.

Bibliography

Tiberiu Boroș, Ștefan Daniel Dumitrescu, and Ruxandra Burtică. NLP-cube: End-to-end raw text processing with neural networks. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 171–179, Brussels, Belgium, October 2018. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/K18-2017`.

Andra Șerbanescu. *Întrebarea - Teorie și practică*. Editura POLIROM, Iași, România, 2002. ISBN 973-683-958-3.

Ethnologue contributors. Romanian, 2018. URL `https://www.ethnologue.com/language/ron`. [Online; last accessed 16-September-2019].

Gramatica contributors. Subiectul neexprimat, 2019. URL `https://gramaticalimbiiromane.ro/sintaxa/sintaxa-propozitiei/subiectul/subiectul-neexprimat/`. [Online; last accessed 16-September-2019].

Michael Heilman. Automatic factual question generation from text. *Language Technologies Institute School of Computer Science Carnegie Mellon University*, 195, 2011.

Tobias Kolditz. Generating questions for german text. master thesis in computational linguistics. *Master thesis in computational linguistics*, 2015.

Gabriela Pană Dindelegan. *The Grammar of Romanian*. Oxford University Press, 2013. ISBN 978–0–19–964492–6.

Paul Piwek and Kristy Elizabeth Boyer. Varieties of question generation: introduction to this special issue. *Dialogue & Discourse*, 3(2):1–9, 2012.

Sonam Soni, Praveen Kumar, and Amal Saha. Automatic question generation: A systematic review. *Available at SSRN 3403926*, 2019.

Universal Dependencies contributors. Universal dependencies - obl, 2014-2017. URL `https://universaldependencies.org/u/dep/obl.html`. [Online, last accessed 10-September-2019].

Xuchen Yao and Yi Zhang. Question generation with minimal recursion semantics. In *Proceedings of QG2010: The Third Workshop on Question Generation*, pages 68–75. Citeseer, 2010.

Xuchen Yao, Gosse Bouma, and Yi Zhang. Semantics-based question generation and implementation. *Dialogue & Discourse*, 3(2):11–42, 2012.