INTERNATIONAL NOOJ CONFERENCE 2020 ZAGREB, CORATIA

Annotation of Cause-Result Questions in Standard Arabic Using Syntactic Grammars

University of Manouba, TUNISIA

Essia Bessaies
Slim mesfar
Henda ben ghzela

Outline

- Introduction
- Motivation & contribution
- Related Works
- Our approach
- Experimentation and results
- Conclusion and perspectives

Introduction

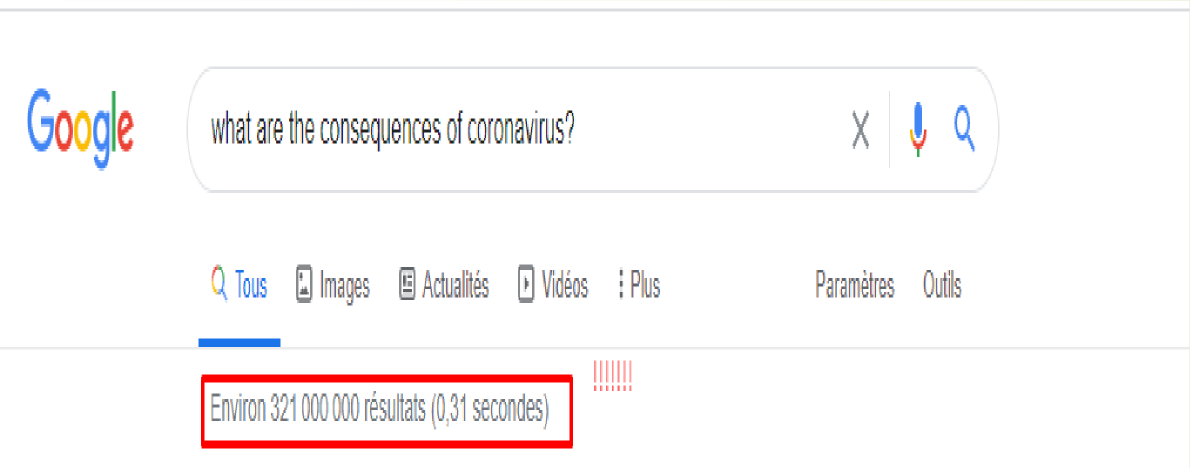
For information retrieval systems :

- Need to ask natural language questions?
- Need precise & domain related results ?

This task is especially complex when we need a short and precise answer.

Introduction

- ✓ **More than one million result by query**
- ✓ **Result documents are cross-domain**
- ✓ **1.2 trillion searches per year worldwide**
- ✓ **+ 20% of users are not satisfied with their initial keywords**



what are the consequences of coronavirus?

Environ 321 000 000 résultats (0,31 secondes)

Introduction

A Question answering system

Introduction

ARABIC NLP QA SYSTEM

Agglutinative morphology

syntactico-semantic ambiguity

Complex syntactic structures

Lack of available tools to assist researchers

Introduction

NL Questions are classified into different categories:

- **Factoid**
- **Binary questions**
- **Definition**
- **Cause/Result**
- **Procedural**
- **Opinion**
- **Evaluative or comparative**

Example :
What are the consquences of corona virus ?

Introduction

In our study domain (**medical domain**), this type of question is generally asked for information about medical treatments, and symptoms

In this domain, most of researches in QA systems have been developed for factual questions, In recent years, automatic extraction of semantic relations has become increasingly important for applications related to question answering

In particular, the cause-result relation is thought to play a very important part in human cognition due to its ability to influence decision making

This knowledge augmentation could be supremely valuable in many domains, especially for medical domain

Motivation & contribution

- Use the linguistic platform NooJ to build the required linguistic resources and rules.
- Present a method for analyzing medical cause- result questions.
- Analyze the asked question by means of a rule-based processing covering the morpho-syntactic level.
- Develop a question answering system in Arabic based on a linguistic approach.

Motivation & contribution

- Our contributions are to propose a Pattern recognizer model that employs a set of linguistic patterns identified based on a combination:
    - ☐ Keywords (Triggers)
    - ☐ Part-of-speech tags
    - ☐ Phrasal structures
- These combinations will be used to automatically identify sentences with the causal type.

• Related Works

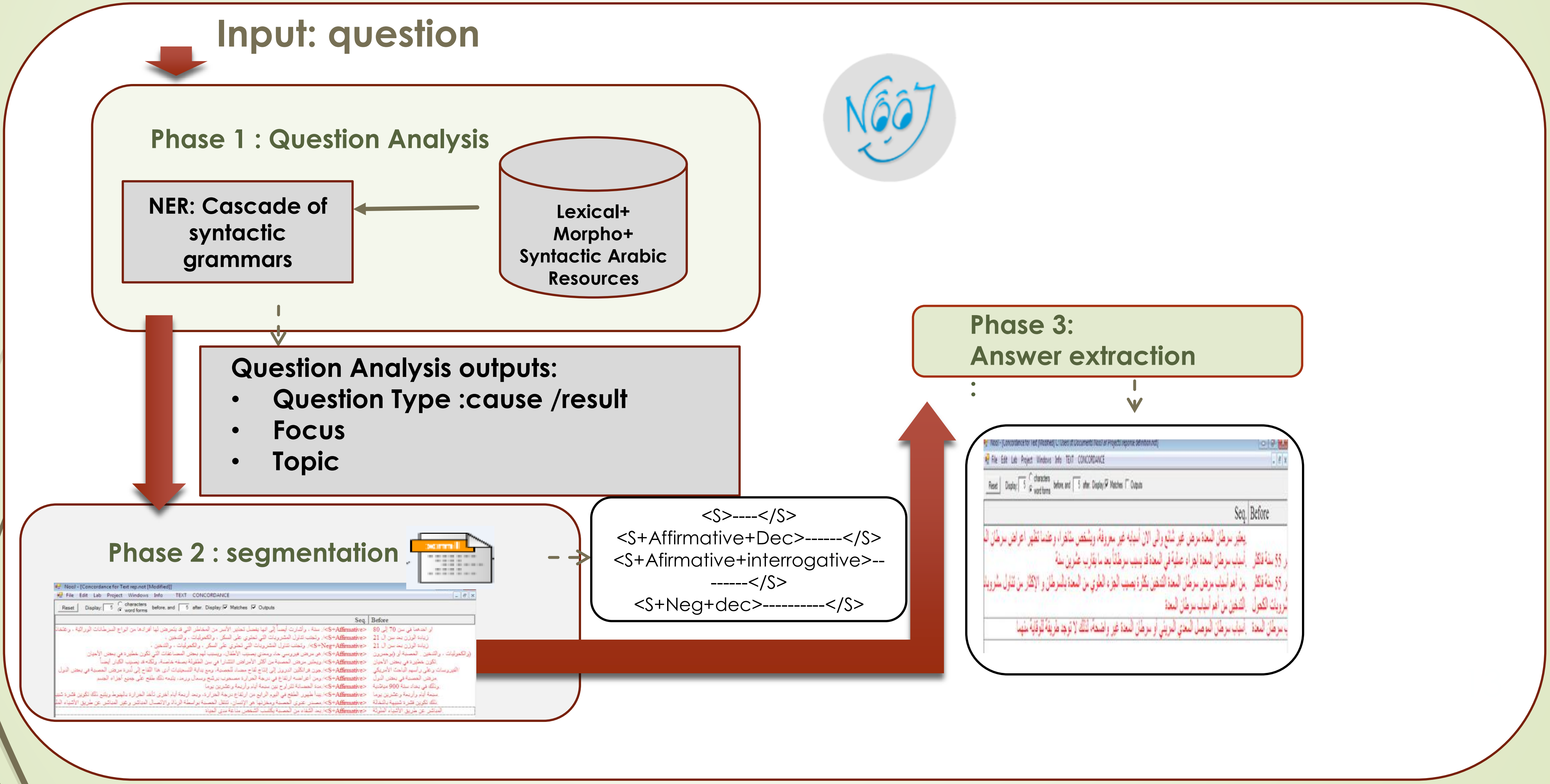
There are two main approaches for constructing patterns:

Rule based approaches	**Machine-learning approaches**
Khoo et al. 2000: A method for performing automatic extraction of cause-effect information from textual Medical documents.	Rink et al. 2010:This paper outlined a approach for discovering causal relations between events in text using graph patterns as features to a classifier. shows that this approach achieves better results when compared with a method that uses a flat representation on the same set of features.

• Related Works

Rule based approaches	**Machine-learning approaches**
Low et al. 2001: this paper study the application of a causation semantic template on the Hong Kong Stock market movement (Hang Seng Index) with English financial news from Reuters.the system shows that it can correctly analyzes single reason sentences with about 76% precision and 74% recall rates.	Do et al. 2011:develops a minimally supervised approach, based on focused distributional similarity methods and discourse connectives, for identifying of causality relations between events in context. We show that combining discourse relation predictions and distributional similarity methods.

Our proposed approach



Input: question

Phase 1 : Question Analysis

NER: Cascade of syntactic grammars

Lexical+ Morpho+ Syntactic Arabic Resources

Question Analysis outputs:

- **Question Type :cause /result**
- **Focus**
- **Topic**

Phase 2 : segmentation

<S>----</S>
<S+Affirmative+Dec>------</S>
<S+Afirmative+interrogative>--------</S>
<S+Neg+dec>----------</S>

Phase 3: Answer extraction

Our proposed approach

Phase 1 : Question Analysis

- Make a linguistic analysis of questions ➔ Add all annotations associated with all recognized forms (lexical , morphological, syntactic as well as distributional information) .

- Apply a NooJ syntactic grammar :
  - ✓ Extract the type of cause-result question
  - ✓ Identify and annotate topic of question.

Our proposed approach

Phase 1 : Question Analysis

- We have two types cause result question (N.ASHQAR,2016) :
  - Explicit

Example:

What are the effects of thyroid disease ?

ما هي تأثيرات مرض الغدة الدرقية ؟

  - Ambiguous

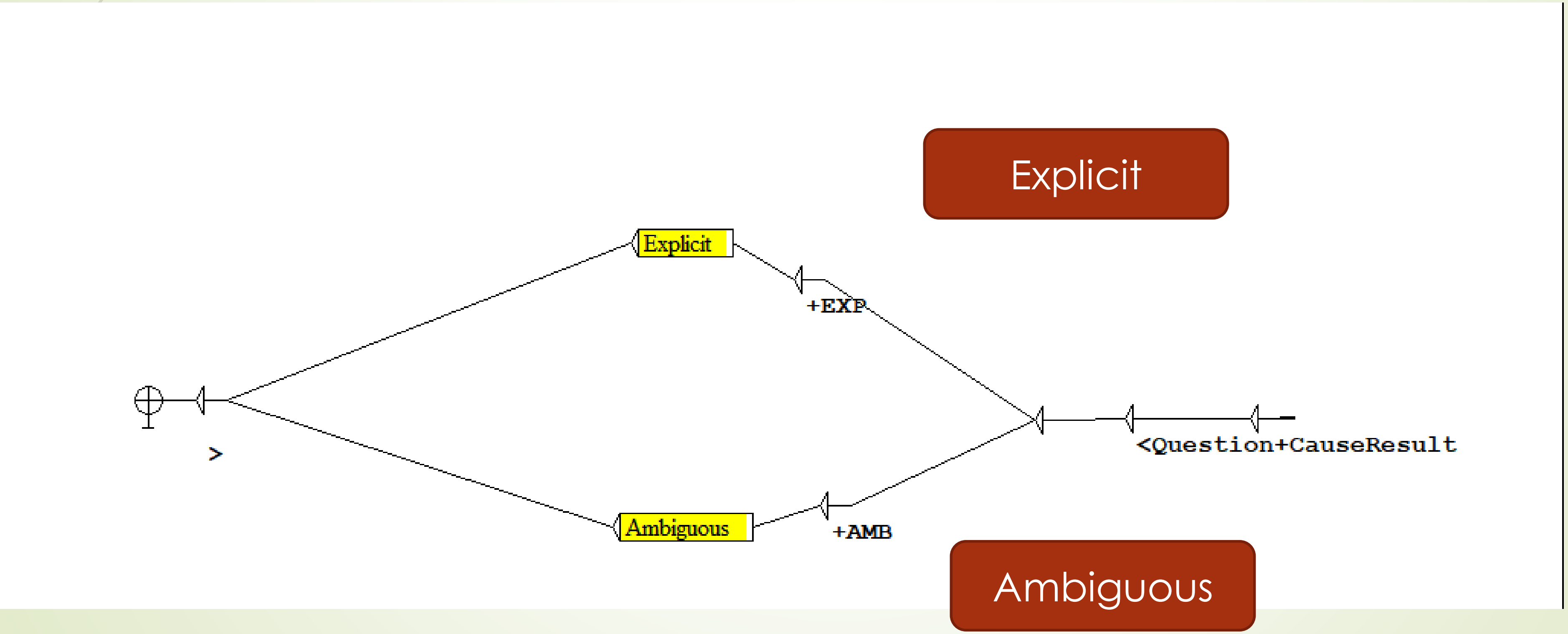
Example:

What could cause an allergic reaction ?

ما يمكن أن يسبب رد فعل تحسسي ؟

Our proposed approach

Phase 1 : Question Analysis



Explicit

Ambiguous

Our proposed approach

Phase 1 : Question Analysis

Example :

ما هي تأثيرات فيروس كورونا؟

↔

<ENAMEX+Medic>

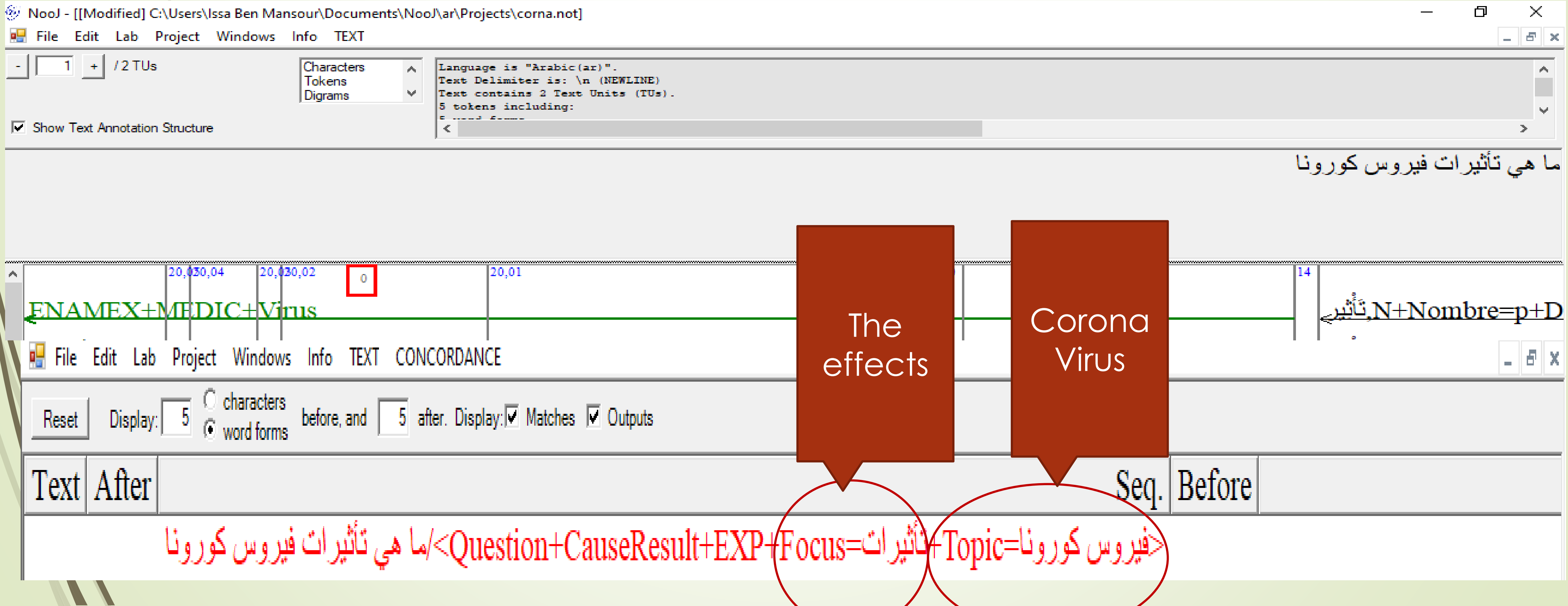
What are the effects of corona virus ?

ما هي : innterrogative mark
تأثيرات : Focus
فيروس كورونا: Topic ***<ENAMEX+Medic>***

Our proposed approach

Phase 1 : Question Analysis

Apply a syntactic grammar to identify and annotate the topic and focus of question.


NooJ - [[Modified] C:\Users\Issa Ben Mansour\Documents\NooJ\ar\Projects\corna.not]
File Edit Lab Project Windows Info TEXT
- 1 + / 2 TUs
Characters
Tokens
Digrams
☑ Show Text Annotation Structure
20,030,04
20,030,02
0
20,01
14
ENAMEX+MEDIC+Virus
تأثير,N+Nombre=p+D
File Edit Lab Project Windows Info TEXT CONCORDANCE
Reset
Display: 5
○ characters
● word forms
before, and 5 after. Display: ☑ Matches ☑ Outputs
Text
After
Seq.
Before
The effects
Corona Virus

```
Language is "Arabic(ar)".
Text Delimiter is: \n (NEWLINE)
Text contains 2 Text Units (TUs).
5 tokens including:
```

ما هي تأثيرات فيروس كورونا

ما هي تأثيرات فيروس كورونا/<Question+CauseResult+EXP+Focus=تأثيرات+Topic=فيروس كورونا>

Our proposed approach

Phase 2 : segmentation

- Integration of a segmentation grammar for Arabic texts ➔ an enhanced version of (S. Keskes & al., 2012)
- The segmentation tool will also identify the sentence style :
  - +Declarative,
  - +Imperative,
  - +Interrogative OR + Exclamative

➔ Generate XML <S> tags

Our proposed approach

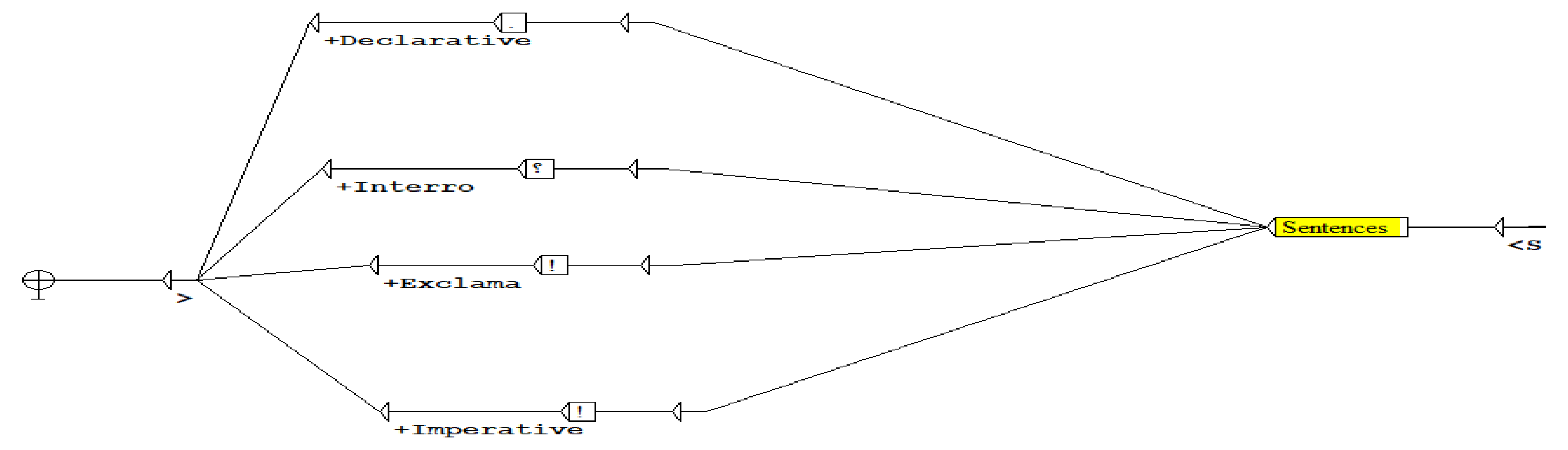
Phase 2 : segmentation

ما هي تأثيرات فيروس كورونا؟

يؤثر فيروس كورونا تقريبا على كل عضو في الجسم يصيب فيروس كورونا الرئتين. تظهر أعراض رئيسية هي الحمى والسعال الجاف المستمر. وتقول هيئة خدمة الصحة الوطنية البريطانية إن "السعال الجاف" يعني السعال القشري الذي لا يصاحبه أي بلغم (مخاط سميك). كما يؤدي ذلك إلى حدوث ضيق في التنفس، والذي يوصف غالبا بضيق شديد في الصدر، أو صعوبة في التنفس، أو شعور بالاختناق.

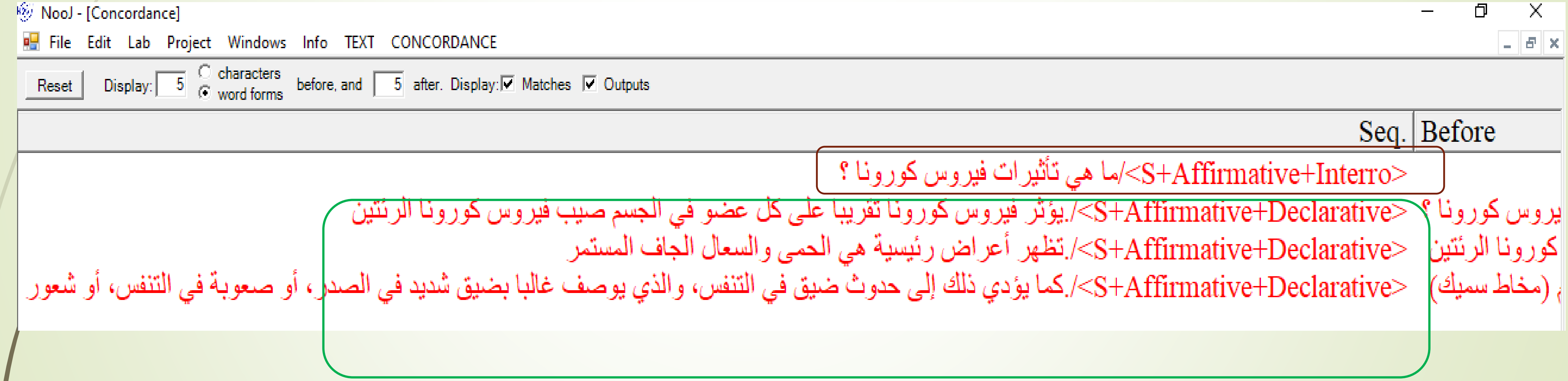
Our proposed approach

Phase 2 : segmentation



Our proposed approach

Phase 2 : segmentation



Display: 5 ○ characters ● word forms before, and 5 after. Display: ☑ Matches ☑ Outputs

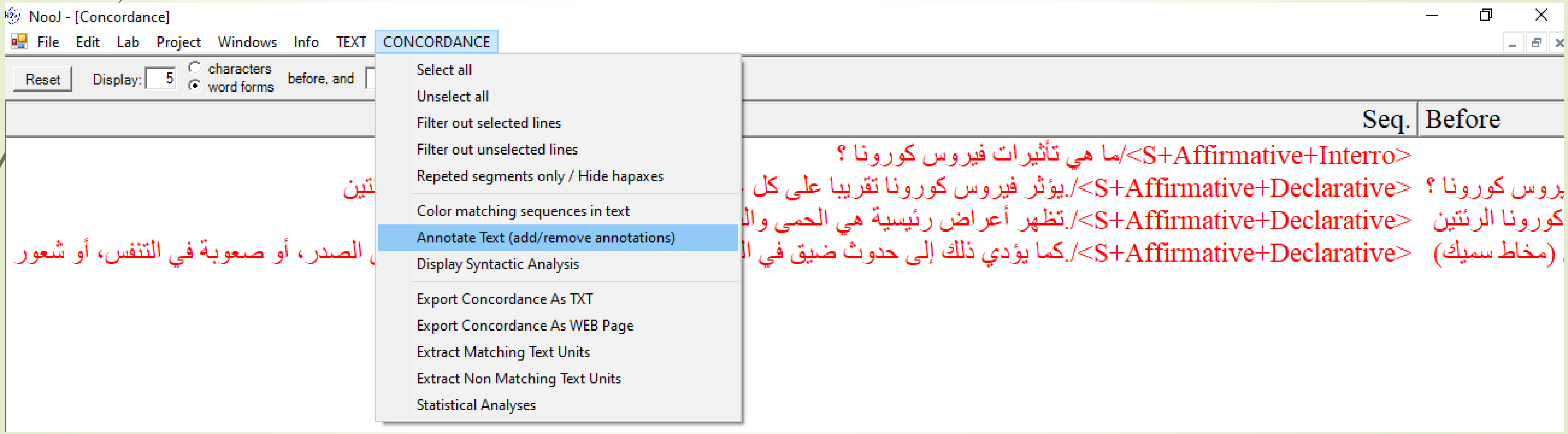
Seq.	Before
ما هي تأثيرات فيروس كورونا ؟/<S+Affirmative+Interro>	
يؤثر فيروس كورونا تقريبا على كل عضو في الجسم صيب فيروس كورونا الرئتين./<S+Affirmative+Declarative>	يروس كورونا ؟
تظهر أعراض رئيسية هي الحمى والسعال الجاف المستمر./<S+Affirmative+Declarative>	كورونا الرئتين
كما يؤدي ذلك إلى حدوث ضيق في التنفس، والذي يوصف غالبا بضيق شديد في الصدر، أو صعوبة في التنفس، أو شعور./<S+Affirmative+Declarative>	(مخاط سميك)

We Need only declarative sentences !!!

Our proposed approach

Phase 2 : segmentation

1. **NooJ>Concordance>Annotate Text (add/ remove annotation)**


NooJ - [Concordance]
File
Edit
Lab
Project
Windows
Info
TEXT
CONCORDANCE
Reset
Display: 5
characters
word forms
before, and
Select all
Unselect all
Filter out selected lines
Filter out unselected lines
Repeted segments only / Hide hapaxes
Color matching sequences in text
Annotate Text (add/remove annotations)
Display Syntactic Analysis
Export Concordance As TXT
Export Concordance As WEB Page
Extract Matching Text Units
Extract Non Matching Text Units
Statistical Analyses
Seq.
Before

<S+Affirmative+Interro>/ما هي تأثيرات فيروس كورونا ؟

يروس كورونا ؟ <S+Affirmative+Declarative>/.يؤثر فيروس كورونا تقريبا على كل … تين

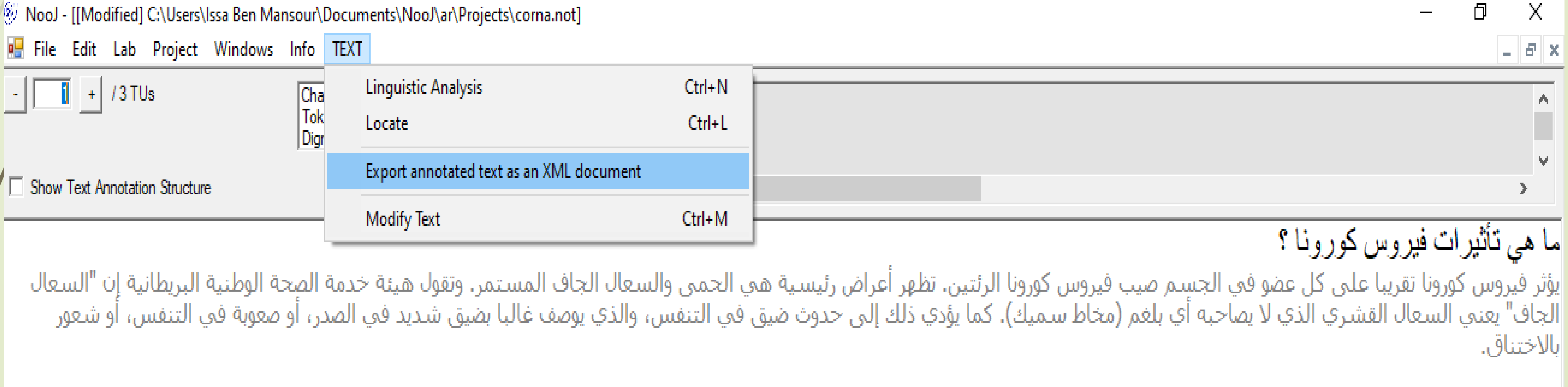
كورونا الرئتين <S+Affirmative+Declarative>/.تظهر أعراض رئيسية هي الحمى وال

(مخاط سميك) <S+Affirmative+Declarative>/.كما يؤدي ذلك إلى حدوث ضيق في ال … الصدر، أو صعوبة في التنفس، أو شعور

Our proposed approach

Phase 2 : segmentation

2. **Text> Exporate annotated text as an XML document**


NooJ - [[Modified] C:\Users\Issa Ben Mansour\Documents\NooJ\ar\Projects\corna.not]
File
Edit
Lab
Project
Windows
Info
TEXT
- 1 + / 3 TUs
Cha
Tok
Dig
☐ Show Text Annotation Structure
Linguistic Analysis Ctrl+N
Locate Ctrl+L
Export annotated text as an XML document
Modify Text Ctrl+M

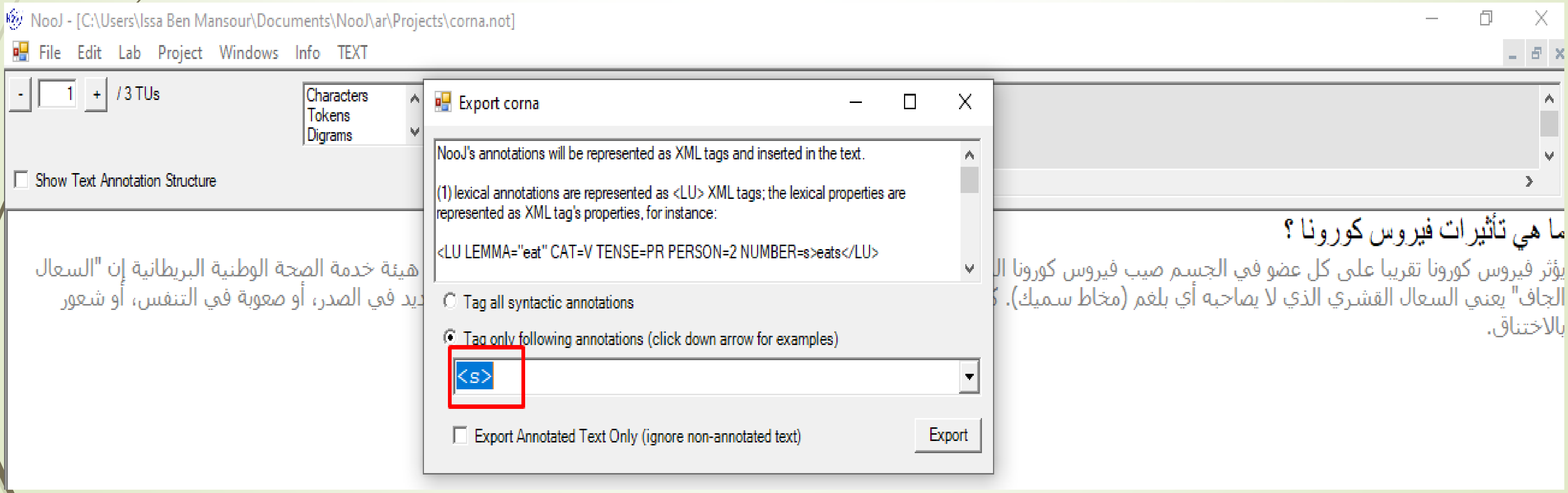
ما هي تأثيرات فيروس كورونا ؟

يؤثر فيروس كورونا تقريبا على كل عضو في الجسم صيب فيروس كورونا الرئتين. تظهر أعراض رئيسية هي الحمى والسعال الجاف المستمر. وتقول هيئة خدمة الصحة الوطنية البريطانية إن "السعال الجاف" يعني السعال القشري الذي لا يصاحبه أي بلغم (مخاط سميك). كما يؤدي ذلك إلى حدوث ضيق في التنفس، والذي يوصف غالبا بضيق شديد في الصدر، أو صعوبة في التنفس، أو شعور بالاختناق.

Our proposed approach

Phase 2 : segmentation

3.Added <S>


NooJ - [C:\Users\Issa Ben Mansour\Documents\NooJ\ar\Projects\corna.not]
File
Edit
Lab
Project
Windows
Info
TEXT
- 1 + / 3 TUs
Characters
Tokens
Digrams
☐ Show Text Annotation Structure
Export corna

NooJ's annotations will be represented as XML tags and inserted in the text.

(1) lexical annotations are represented as <LU> XML tags; the lexical properties are represented as XML tag's properties, for instance:

<LU LEMMA="eat" CAT=V TENSE=PR PERSON=2 NUMBER=s>eats</LU>

○ Tag all syntactic annotations

● Tag only following annotations (click down arrow for examples)

<s>

☐ Export Annotated Text Only (ignore non-annotated text)

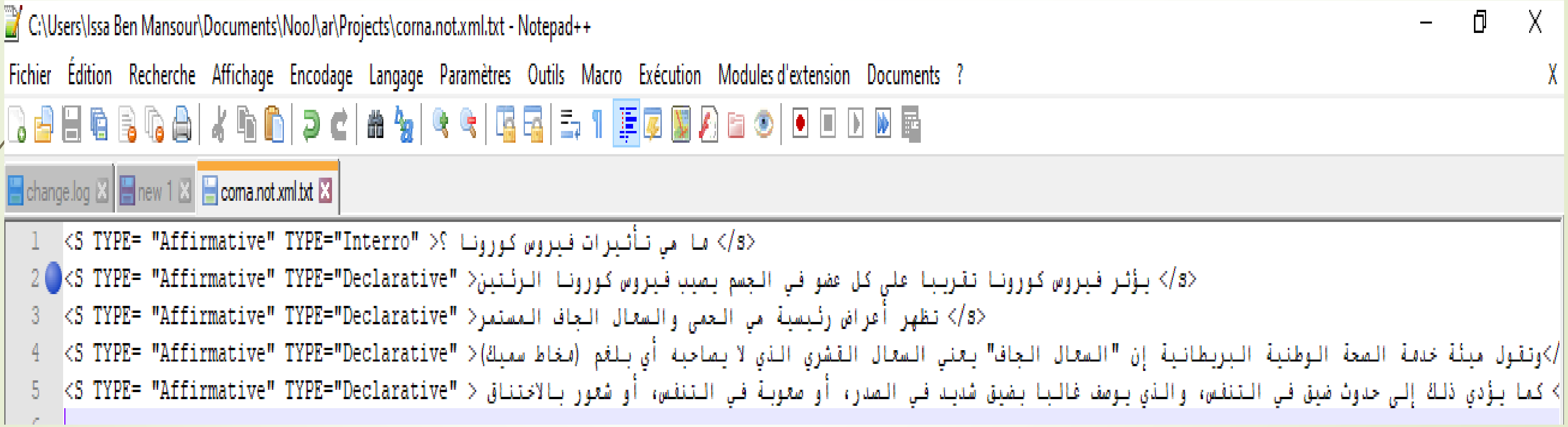
ا هي تأثيرات فيروس كورونا ؟

ؤثر فيروس كورونا تقريبا على كل عضو في الجسم صيب فيروس كورونا ا
الجاف" يعني السعال القشري الذي لا يصاحبه أي بلغم (مخاط سميك).
بالاختناق.

هيئة خدمة الصحة الوطنية البريطانية إن "السعال
يد في الصدر، أو صعوبة في التنفس، أو شعور

Our proposed approach

Phase 2 : segmentation



```
<S TYPE= "Affirmative" TYPE="Interro" >ما هي تأثيرات فيروس كورونا ؟ </s>
<S TYPE= "Affirmative" TYPE="Declarative" >يؤثر فيروس كورونا تقريبا علي كل عضو في الجسم يصيب فيروس كورونا الرئتين </s>
<S TYPE= "Affirmative" TYPE="Declarative" >تظهر أعراض رئيسية هي الحمى والسعال الجاف المستمر </s>
<S TYPE= "Affirmative" TYPE="Declarative" >وتقول هيئة خدمة الصحة الوطنية البريطانية إن "السعال الجاف" يعني السعال القشري الذي لا يصاحبه أي بلغم (مخاط سميك)</
<S TYPE= "Affirmative" TYPE="Declarative" > كما يؤدي ذلك إلى حدوث ضيق في التنفس، والذي يوصف غالبا بضيق شديد في الصدر، أو صعوبة في التنفس، أو شعور بالاختناق <
```

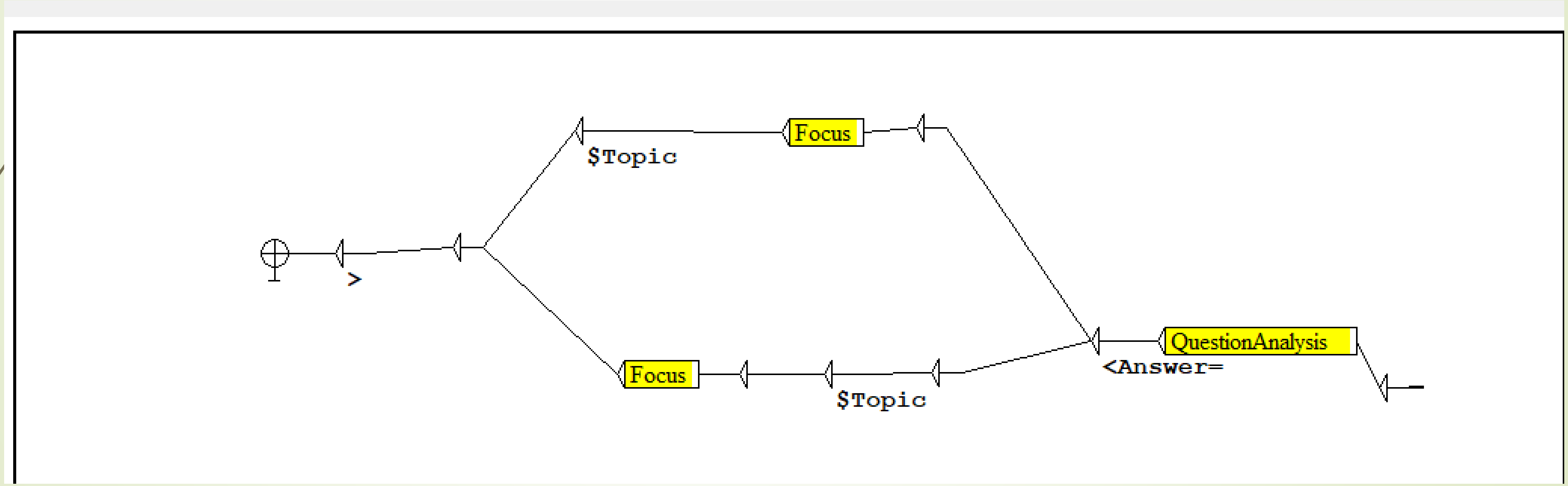
Our proposed approach

Phase 3 : Answer Extraction

- The third motivation behind the question analysis and segmentation task is to develop the linguistic patterns for the candidate passages.
- The passage retrieval is typically used as the first step in current question answering systems.
- After that we applying a grammar to extract the short and precise answer

Our proposed approach

Phase 3 : Answer Extraction:

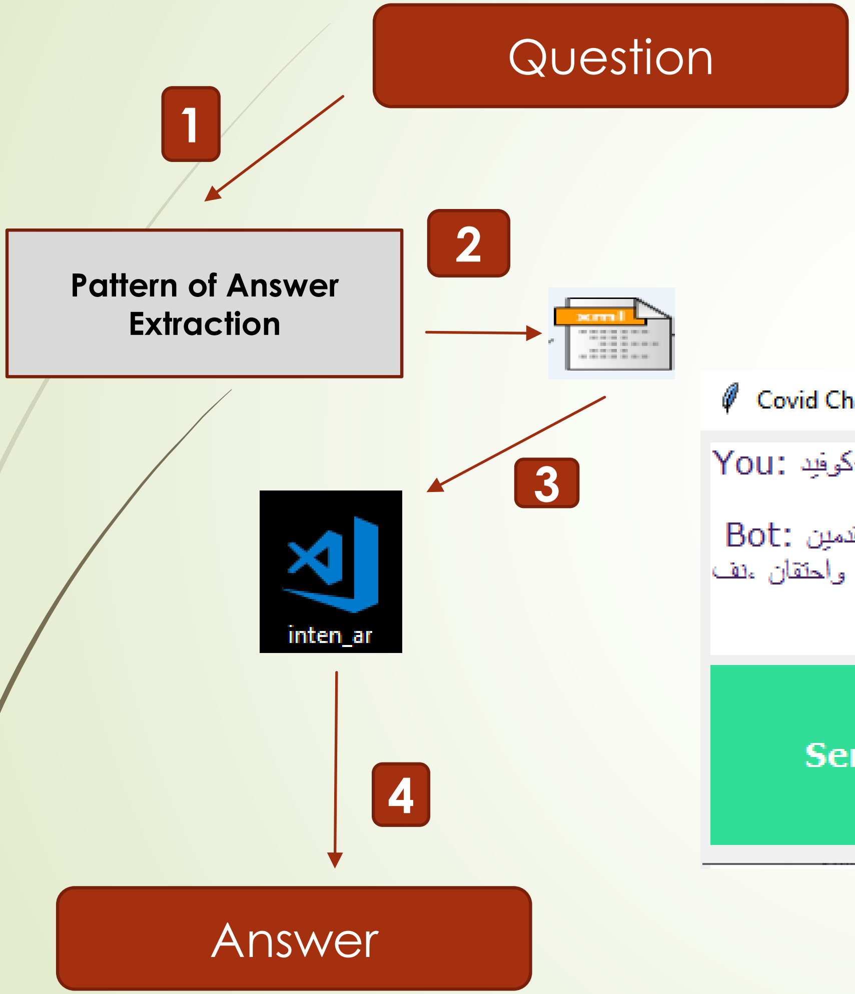


Experimentation and results

1. Apply a grammar to annotate Question and to extract short answer with NooJ
2. Export concordance as XML File
3. Parse XML File as Json python with xmltodict
4. We developed a chatbot with python

5. We used NooJ to enrich the chatbat training data in Json file

Experimentation and results



Question

1

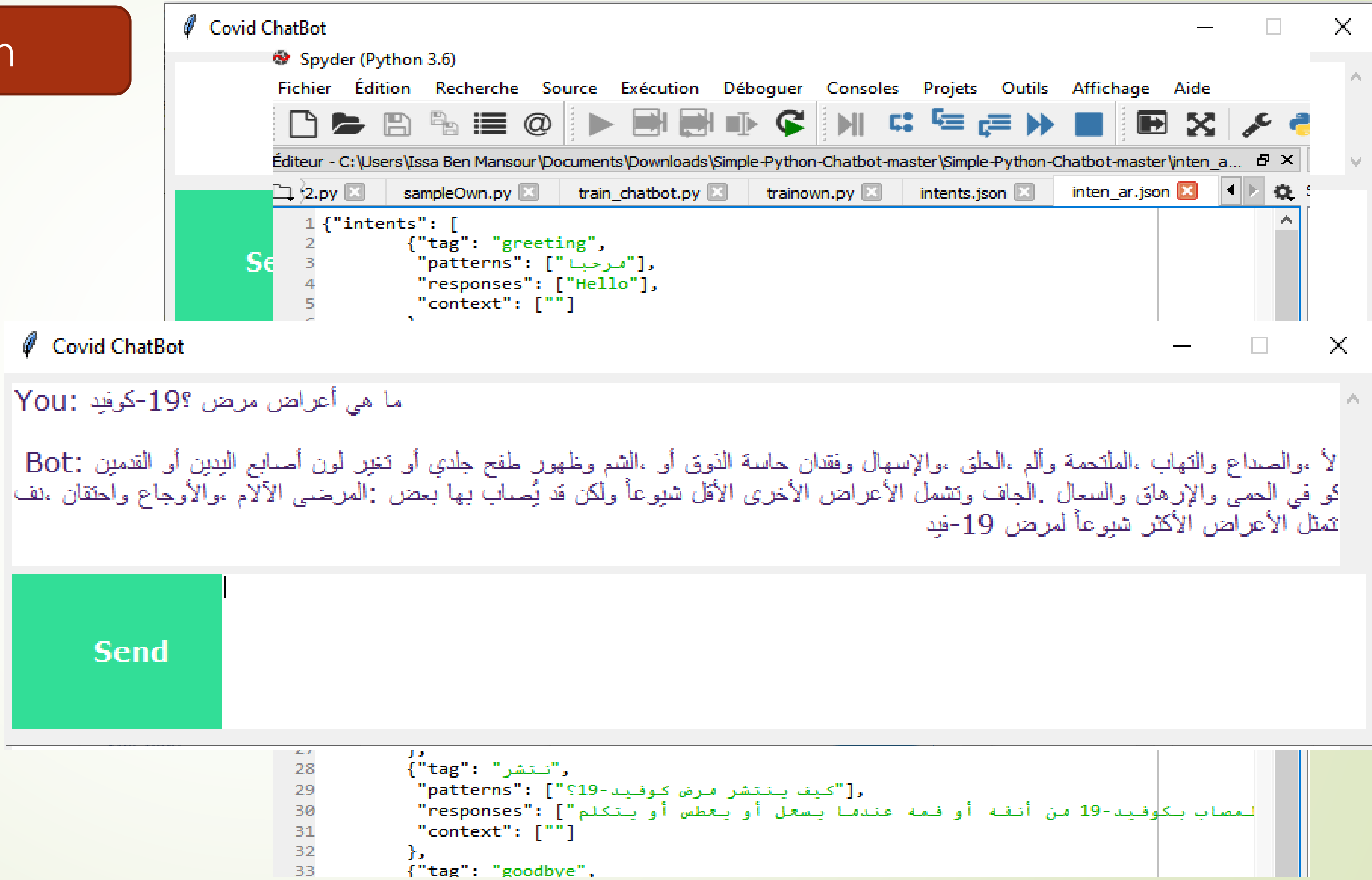
Pattern of Answer Extraction

2

3

4

Answer



```
{"intents": [
        {"tag": "greeting",
         "patterns": ["مرحبا"],
         "responses": ["Hello"],
         "context": [""]
```

You: ما هي أعراض مرض ؟19-كوفيد

Bot: لأ، والصداع والتهاب، الملتحمة وألم، الحلق، والإسهال وفقدان حاسة الذوق أو، الشم وظهور طفح جلدي أو تغير لون أصابع اليدين أو القدمين
كو في الحمى والإرهاق والسعال الجاف. وتشمل الأعراض الأخرى الأقل شيوعاً ولكن قد يُصاب بها بعض :المرضى الآلام، والأوجاع واحتقان، نف
تمثل الأعراض الأكثر شيوعاً لمرض 19-فيد

```
        {"tag": "نتشر",
         "patterns": ["كيف ينتشر مرض كوفيد-19؟"],
         "responses": ["لمصاب بكوفيد-19 من أنفه أو فمه عندما يسعل أو يعطس أو يتكلم
         "context": [""]
        },
        {"tag": "goodbye",
```

Conclusion and perspectives

- In this work, we developed a question answering which is based on a Linguistic approach.
- The use of the linguistic engine of Nooj in order to formalize the automatic recognition rules and then applying them to a dynamic corpus composed of Arabic medical journalistic articles of Coronavirus.
- Question analysis: apply a syntactic grammar to identify and annotate the topic and focus of question.
- After the phase of question analysis, we integrated the model with Nooj in our algorithm of chatbat.

References

- Christopher Khoo, Syin Chan, and Yun Niu. 2000. Extracting causal knowledge from a medical database using graphical patterns. In *Proceedings of 38th Annual Meeting of the ACL*. 336–343.
- Boon-Toh Low, Ki Chan, Lei-Lei Choi, Man-Yee Chin, and Sin-Ling Lay. 2001. Semantic expectation-based causation knowledge extraction: A study on Hong Kong stock movement analysis. In David Cheung, Graham J. Williams, and Qing Li (Eds.), *Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science:* Vol. 2035, Springer, Berlin,

References

- Quang Do, Yee Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of EMNLP'11*, 294–303.
- Bryan Rink, Cosmin Bejan, and Sanda Harabagiu. 2010. Learning textual graph patterns to detect causal event relations. In *Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference*. 265–270.