

International NooJ Conference 2017
Morocco, 18 – 20 May 2017

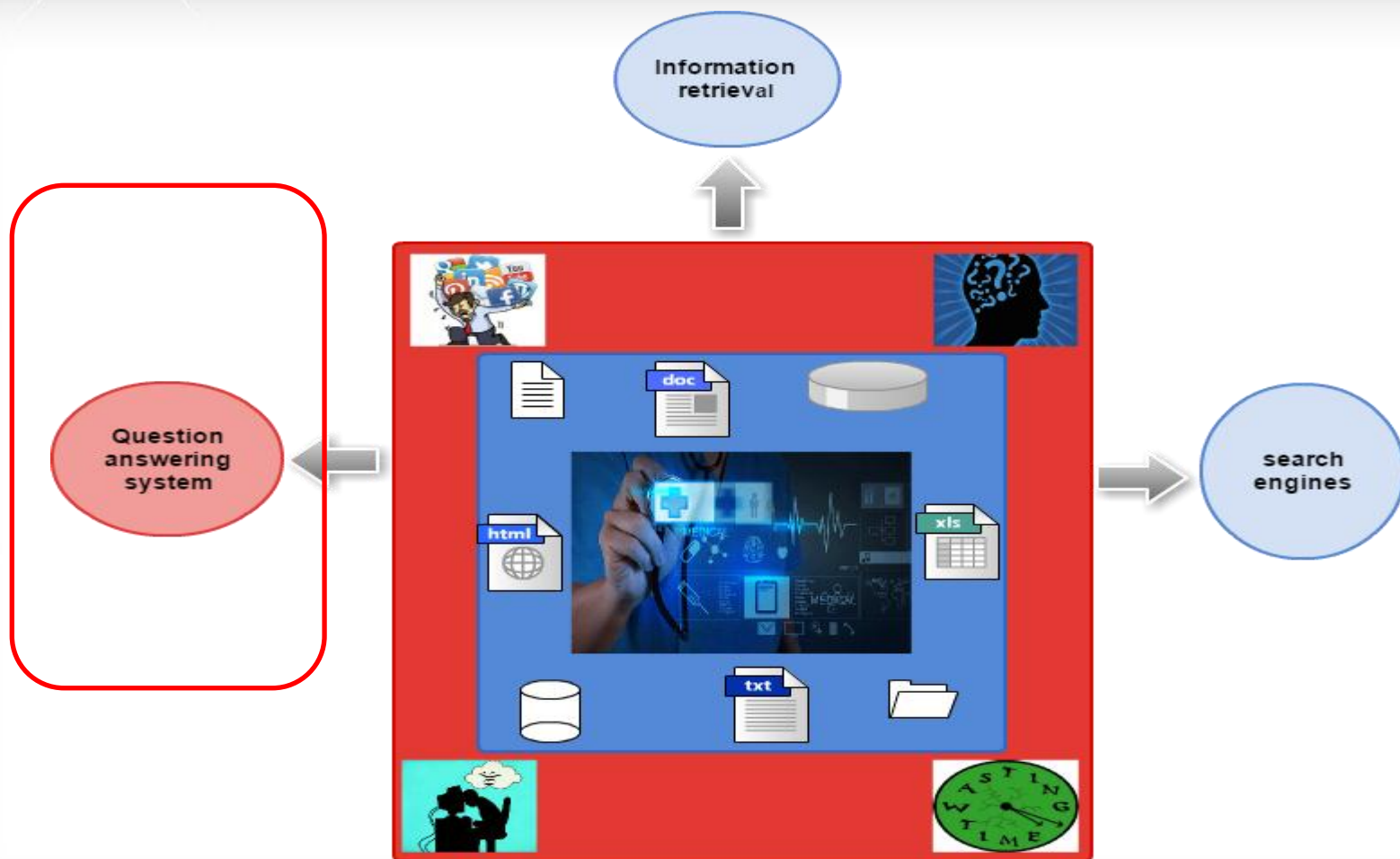
**RECOGNITION AND ANALYSIS OF
BINARY QUESTIONS FOR STANDARD
ARABIC**

Essia Bessaies, Slim Mesfar, Henda Ben Ghezala
University of Manouba, TUNISIA





Introduction





Introduction

- There has been a lot of research in the field of English +- & some European language Question Answering Systems.
- However, Arabic Question Answering Systems could not match the pace due to some inherent difficulties with the language itself as well as due to lack of tools available to assist the researchers.



Introduction

For this purpose, the developed question answering system is based on a linguistic approach.



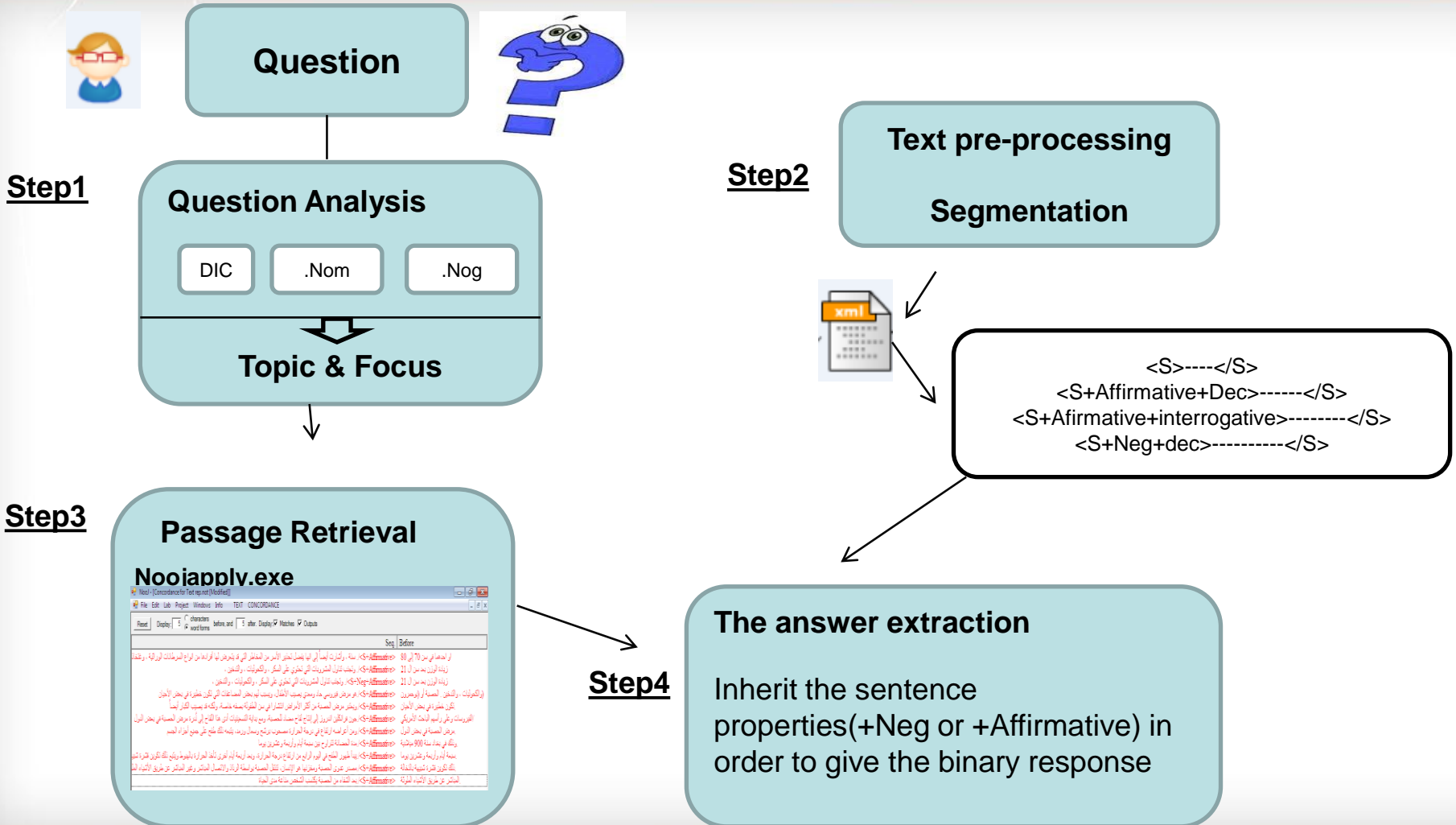
Formalize the automatic recognition rules →
apply them to a dynamic corpus composed of
medical journalistic articles.



Related Works

System	Objective	Domain	Dataset	Results	Shortcomings
AQuASys (S. BEKHTI and M. AL-HARBI,2013)	answer unformatted factbased questions written in an Arabic natural language.	Close domain	ANERcorp: 150,000 tagged tokens) as well as few gazetteers (ANERgazet) available online	recall rate of 97.5% and 66.25% as a precision rate	The system focused only on factoid questions
DefArabicQA (O. Trigui,2010)	provide short answers for Arabic natural language questions	Close domain	collection of Arabic text documents	not presented	Does not include the other types of question (How and Why)
Yes/No Arabic Question Answering System (H. Kurdi,et al,2014)	design a formal model for a semantic based yes/no Arabic question answering system based on paragraph retrieval	Open domain	20 documents which used to test the system and a collection of 100 different yes/no question	The results of using documents technique:85% when 20 documents are used. The result of using paragraphs technique: 88% when 20 documents are used	The system focused only on yes/no questions. and the corpus size is small (20 documents)
JAWEB (W. N. Bdour and N. K. Gharaibeh , 2008)	provide short answers for Arabic natural language questions	Close domain	an extended version of the Arabic corpus	The system provided 15- 20% higher recall	The system focused only on factoid questions

Our approach





Our approach

Step 1 :Question Analysis

- Make a linguistic analysis of questions → Add annotations associated with all recognized forms (lexical , morpho-logical as well as syntactic information)
- Apply a syntactic grammar to identify and annotate the **topic** and **focus** of question.



Our approach

Step 1 : Question Analysis

Example:

Is measles a contagious disease?

هل يعتبر مرض الحصبة معدياً؟

<ADJ> <ENAMEX+MEDIC> <ADV+Interro>

هل, Is = Binary ininterrogative mark

معدياً, Contagious = Focus

مرض الحصبة, Measles disease = Topic



<مدي=Focus1+مرض الحصبة=Topic1+Binary+Question>/هل يعتبر مرض الحصبة معدي



Our approach

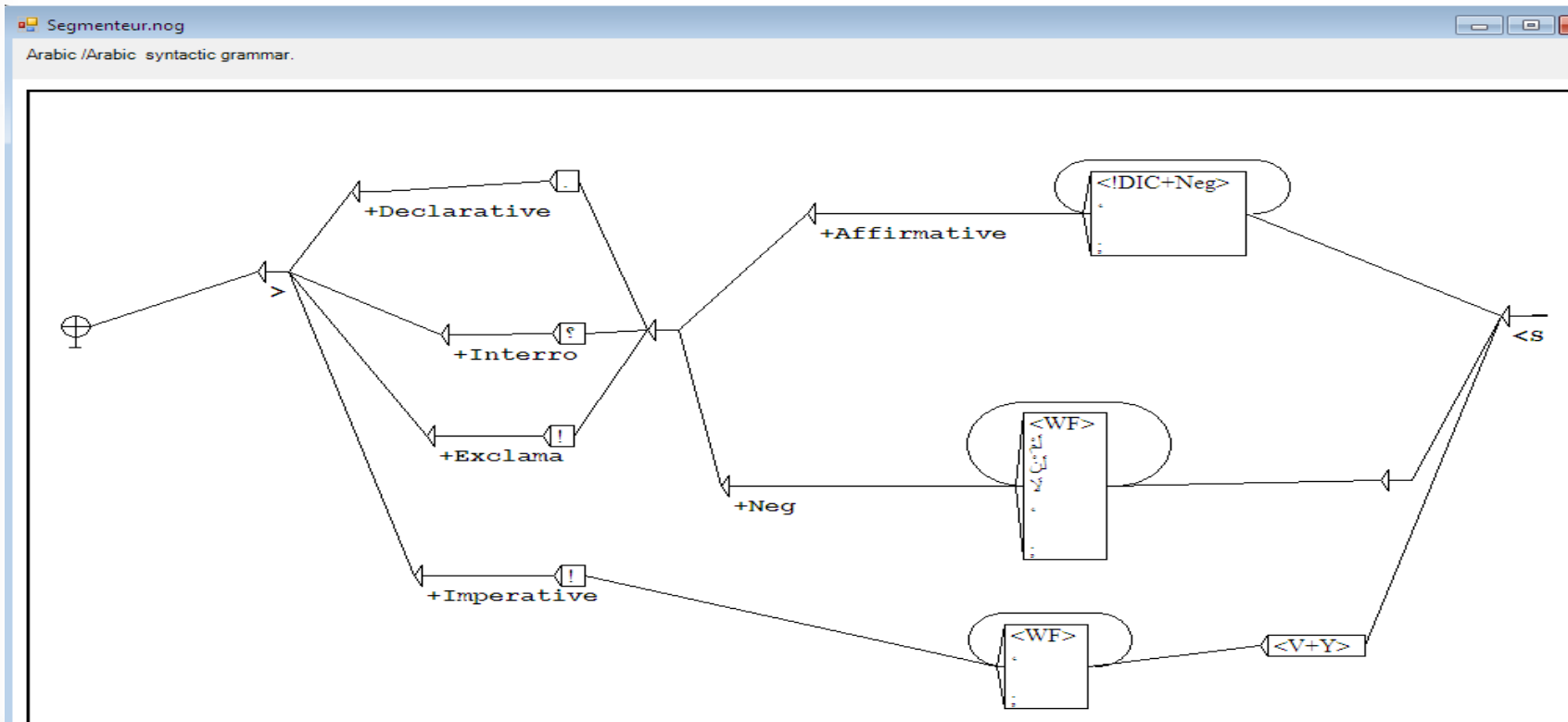
Step 2 : segmentation tool for Arabic Texts

- Integration of a segmentation tool for Arabic texts → an enhanced version of (Nadia Ghezaiel and Kais Haddar. 2016)
- The segmentation tool will also identify :
 - The negation status of the sentence : +Negative OR +Affirmative
 - The sentence style : +Declarative, +Imperative, +Interrogative OR + Exclamative



Our approach

Step 2 : Segmentation tool for Arabic texts





Our approach

Step 2 : Segmentation tool for Arabic texts

Example:

هل يعتبر مرض الحصبة معدي؟ الحصبة هو مرض فيروسي حاد ومعدي يصيب الأطفال، ويسبب لهم بعض المضاعفات التي تكون خطيرة في بعض الأحيان. ويعتبر مرض الحصبة من أكثر الأمراض انتشارا في سن الطفولة بصفه خاصة، ولكنه قد يصيب الكبار أيضاً.



Our approach

Step 2 : Segmentation tool for Arabic texts

Example:

Nool - [Concordance for Text C:\Users\st\Documents\Nool\ar\Projects\example.not]

File Edit Lab Project Windows Info TEXT CONCORDANCE

Reset Display: 5 characters before, and 5 after. Display: Matches Outputs

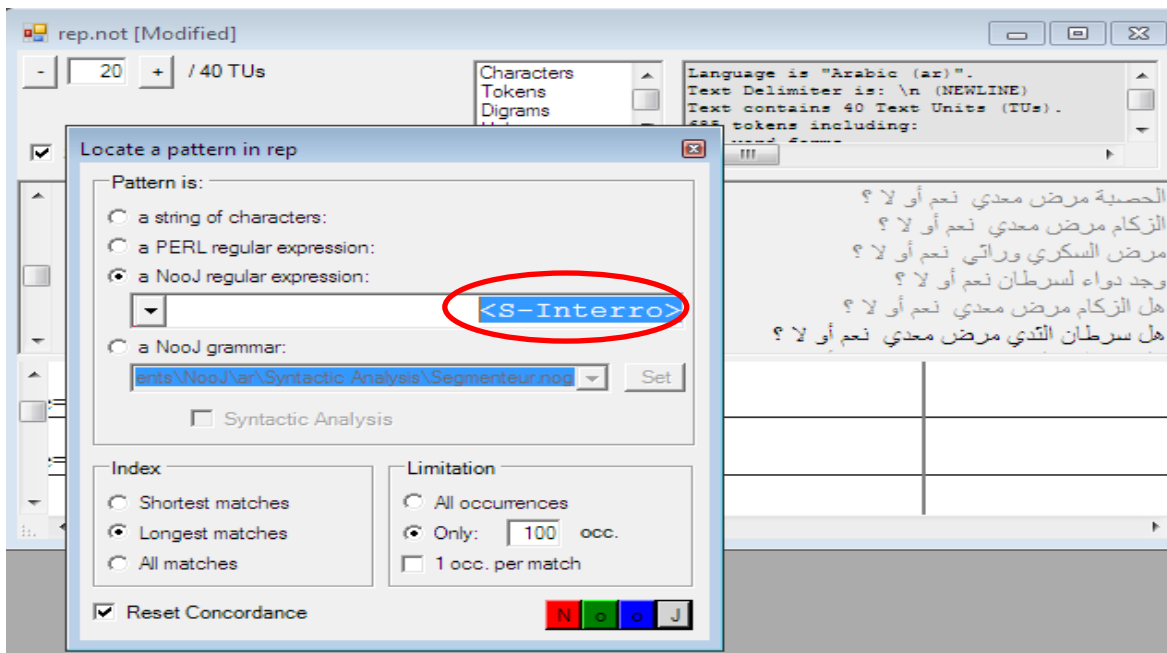
word forms

After	Seq.	Before
الحصية هو	<S+Affirmative+Interro>	هل يعتبر مرض الحصية معدي؟
ويعتبر مره	<S+Affirmative+Declarative>	الحصية هو مرض فيروسي حاد ومعدي يصيب الأطفال، ويسبب لهم بعض المضاعفات التي تكون خطيرة في بعض الأحيان
ن خطيرة في بعض الأحيان	<S+Affirmative+Declarative>	ويعتبر مرض الحصية من أكثر الأمراض انتشارا في سن الطفولة بصفه خاصه، ولكنه قد يصيب الكبار أيضا



Our approach

Step 2 : Segmentation tool for Arabic texts



- Apply a NooJ regular expression :
<S-Interro>



Our approach

Step 2 : Segmentation tool for Arabic Texts

NooJ - [Concordance for Text [Modified] C:\Users\st\Documents\NooJ\ar\Projects\example.not]

File Edit Lab Project Windows Info TEXT CONCORDANCE

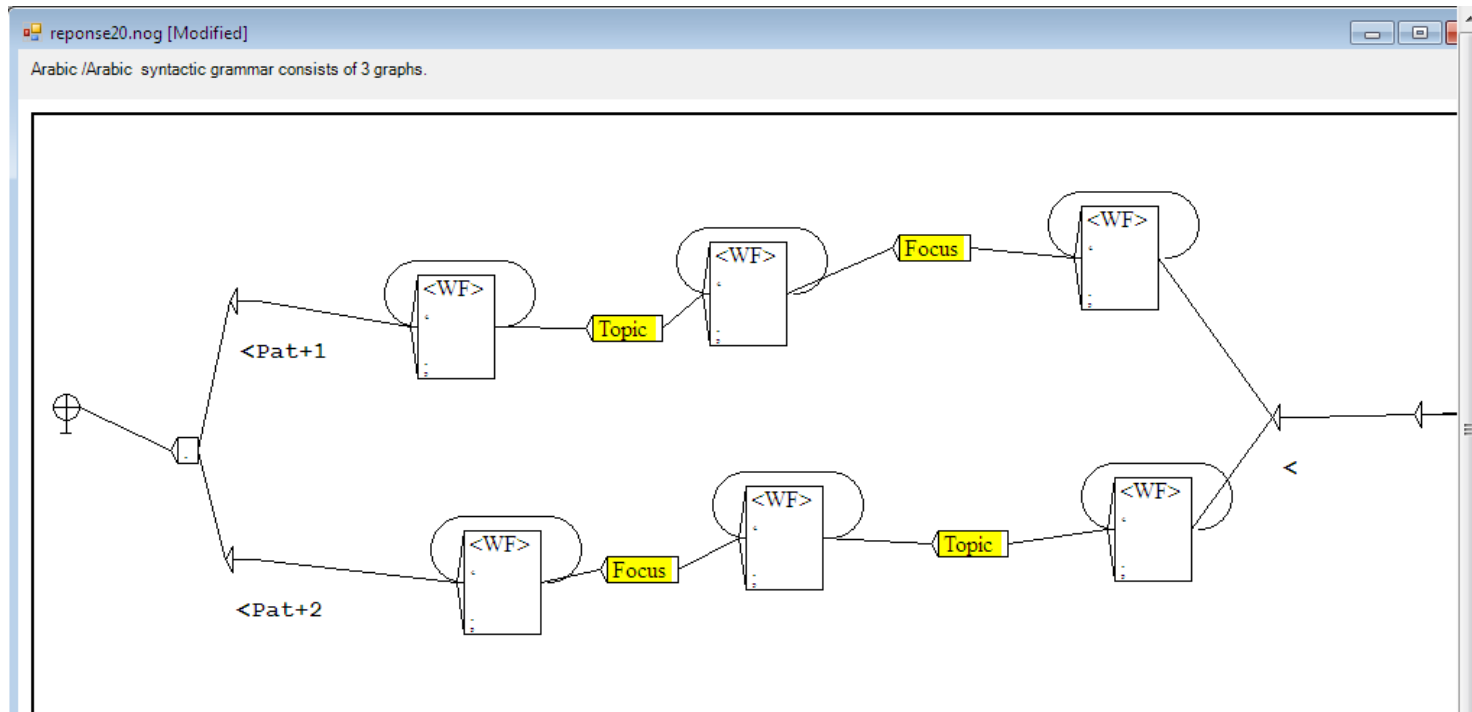
Reset Display: 5 characters before, and 5 after. Display: Matches Outputs word forms

Text	After	Seq.
<i>ويعتبر مرض الحصبة من أكثر</i>		
<i>الحصبة هو مرض فيروسي حاد ومعدّي يصيب الأطفال، ويسبب لهم بعض المضاعفات التي تكون خطيرة في بعض الأحيان</i>		
<i><S+Affirmative+Declarative></i>		
<i>ويعتبر مرض الحصبة من أكثر الأمراض انتشاراً في سن الطفولة بصفه خاصة، ولكنه قد يصيب الكبار أيضاً</i>		
<i><S+Neg+Declarative></i>		



Our approach

Step3 : Passage Retrieval





Our approach

Step 3 : Passage Retrieval

NooJ - [Concordance for Text [Modified] C:\Users\st\Documents\NooJ\ar\Projects\example.not]

File Edit Lab Project Windows Info TEXT CONCORDANCE

Reset Display: 5 characters before, and 5 after. Display: Matches Outputs
 word forms

Text	After	Seq.	Before
ويحتبر مرض الحصبة من أكثر يحتبر مرض الحصبة معدي. يحتبر يحتبر المرض معدي	الحصبة هو مرض فيروسي حاد ومعدي يصيب الأطفال، ويسبب لهم بعض المضاعفات التي تكون خطيرة في بعض الأحيان ويحتبر مرض الحصبة من أكثر الأمراض انتشارا في سن الطفولة بصفه خاصة، ولكنه قد يصيب الكبار أيضاً		هل يحتبر مرض الحصبة معدي؟ تكون خطيرة في بعض الأحيان ولكنه قد يصيب الكبار أيضاً



Our approach

Step 4 : The answer extraction

- In this 4th step, we inherit the negation status of the sentence (+Neg OR + Affirmative) given by the segmentation tool
- As a preliminary result, we suppose that the given status represents correctly the binary answer of our question.
- Then, we are working on the integration of similarity scores in order to better rank the retrieved passages.



Conclusion and perspectives

- We are developing a question answering system 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
- Question analysis: apply a syntactic grammar to identify and annotate the topic and focus of question.



Conclusion and perspectives

- Segmentation tool: will also identify the negation status of the sentence and The style of the sentence
- we are working on the integration of similarity scores in order to better rank the retrieved passages.
- Finally, as a long term ambition, we intend to consider studying the processing of the “why” and “how” question types.



Thank you!

Any question ?



References

- S. BEKHTI and M. AL-HARBI, “Aquasys: A question-answering system for arabic,” in WSEAS International Conference. Proceedings. Recent Advances in Computer Engineering Series, no. 12. WSEAS, 2013.
- H. Kurdi, S. Alkhaider, and N. Alfaifi, “Development and evaluation of a web based question answering system for arabic language,” Computer Science & Information Technology (CS & IT), vol. 4, no. 2, pp. 187– 202, 2014.



References

- O. Trigui, H. Belguith, and P. Rosso, “Defarabicqa: Arabic definition question answering system,” in Workshop on Language Resources and Human Language Technologies for Semitic Languages, 7th LREC, Valletta, Malta, 2010, pp. 40–45.
- W. Brini, M. Ellouze, S. Mesfar, and L. H. Belguith, “An arabic question-answering system for factoid questions,” in Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009. International Conference on. IEEE, 2009, pp. 1–7.