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
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
Introduction

A Question answering system
Introduction

- Agglutinative morphology
- Lack of available tools to assist researchers
- ARABIC NLP QA SYSTEM
- Syntactico-semantic ambiguity
- Complex syntactic structures
Introduction

NL Questions are classified into different categories:

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

Example:
What are the consequences of coronavirus?
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:

<table>
<thead>
<tr>
<th>Rule based approaches</th>
<th>Machine-learning approaches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Khoo et al. 2000: A method for performing automatic extraction of cause-effect information from textual Medical documents.</td>
<td>Rink et al. 2010: This paper outlined an 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.</td>
</tr>
</tbody>
</table>
### Related Works

<table>
<thead>
<tr>
<th>Rule based approaches</th>
<th>Machine-learning approaches</th>
</tr>
</thead>
<tbody>
<tr>
<td>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 analyze single reason sentences with about 76% precision and 74% recall rates.</td>
<td>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.</td>
</tr>
</tbody>
</table>
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

Phase 3: Answer extraction

$<$S$>$ ---- $<$S$>$
$<$S+Affirmative+Dec$>$ ---- $<$S$>$
$<$S+Affirmative+interrogative$>$ ---- $<$S$>$
$<$S+Neg+dec$>$ ---- $<$S$>$
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؟

What are the effects of corona virus?

ما هي: interrogative 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.
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 تقريباً 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 يصاحب أي بلغم (مخاط سميك). كما يؤدي ذلك إلى حدوث ضيق في التنفس، والذي يوصف غالباً بضيق شديد في الصدر، أو صعوبة في التنفس، أو شعور بالاختناق.

"السعال الجاف" يعني السعال القشري الذي لا يصاحب أي بلغم (مخاط سميك). كما يؤدي ذلك إلى حدوث ضيق في التنفس، والذي يوصف غالباً بضيق شديد في الصدر، أو صعوبة في التنفس، أو شعور بالاختناق.
Our proposed approach

Phase 2: segmentation
Our proposed approach

Phase 2: segmentation

We Need only declarative sentences !!!
Our proposed approach

Phase 2: segmentation

1. NooJ>Concordance>Annotate Text (add/remove annotation)
Our proposed approach

Phase 2: segmentation

2. Text> Exporate annotated text as an XML document
Our proposed approach

Phase 2: segmentation

3. Added <S>
Our proposed approach

Phase 2: segmentation
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 chatbot training data in Json file
Experimentation and results

1. Question

2. Pattern of Answer Extraction

3. Covid ChatBot

4. Answer
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


- 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