Automatic Ontology-Based Document Annotation for Arabic Information Retrieval

Ashraf I. Kaloub
Alaqsa-Community College
Khan Younis, Gaza Strip

Rebhi S. Baraka
Faculty of Information Technology
Islamic University of Gaza

1

The 3rd Palestinian Symposium on Computational Linguistics
and Arabic Content (iArabic'2014)
April 12, 2014

OUTLINE

- **Introduction**

- **Methodology and model**

- **System Realization, Experimental Results and Evaluation**

- **Conclusion and Future Work**

INTRODUCTION

- The need for semantically enriched Information Retrieval (IR) and searching are among the most important issues of the semantic web.

- Semantic IR try to overcome the limitations of the traditional IR model which suffers from misunderstanding the query and its context and on the keyword which cannot represent the semantic information of resources therefore obtaining a lower recall and precision.

INTRODUCTION

o Using ontology in the field of IR improves the retrieval accuracy and reduces irrelevant results.

o An Ontology is a formal explicit description of concepts in a domain of discourse classes (concepts). Properties of each concept to describe various features and attributes of the concept (slots), and restrictions on slots (facets) ontology together with a set of individual instances of classes constitutes a knowledge base.

4

INTRODUCTION

- We develop an **automatic ontology-based document annotation and retrieval model** for Arabic documents.

- The model will be used to improve the accuracy of Arabic retrieved documents depending on Arabic Ontology Domain "فقه الصلاة" (Prayer jurisprudence).

- All Documents in this domain are written in Arabic language and stored in a corpus.

METHODOLOGY AND MODEL

To build the model various steps performed:

1. Preparing the corpus

2. Building Arabic Ontology Domain "فقه الصلاة" (Prayer jurisprudence)

3. Documents annotation

4. Processing annotated documents

5. Indexing and searching

6

Methodology and model

Preparing the corpus

- The corpus is a collection of documents in the domain "فقه" الصلاة (Prayer Jurisprudence).

- We collect these documents from IslamWeb website related to Fatwa questions in the field of Islamic issues.

- Collected documents are converted to xml type when we load it into Gate in order to facilities the processing of documents annotation and retrieval.
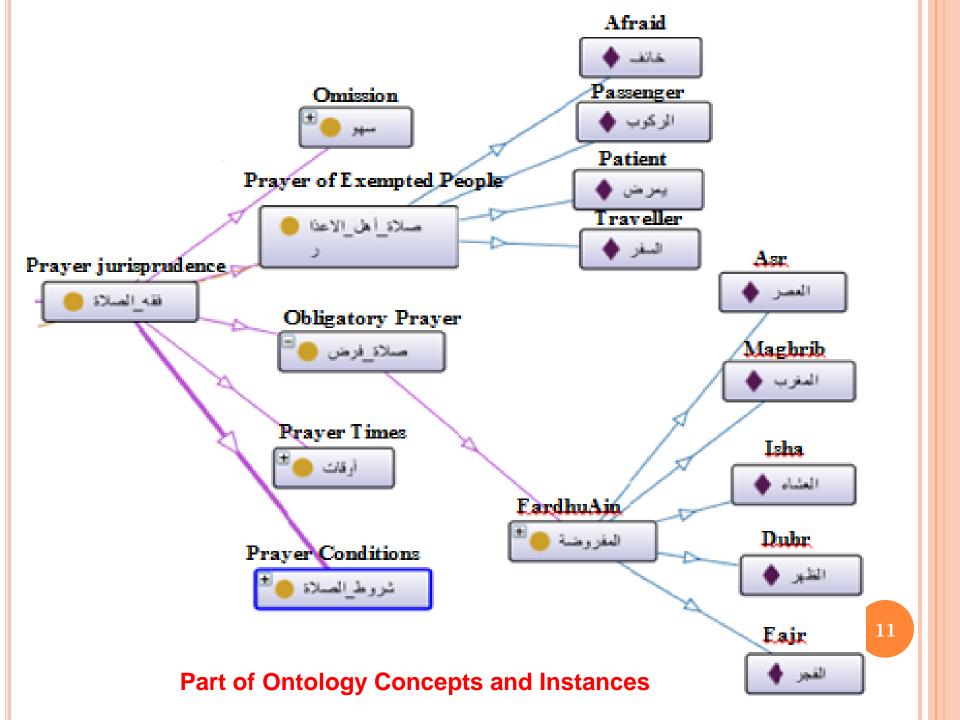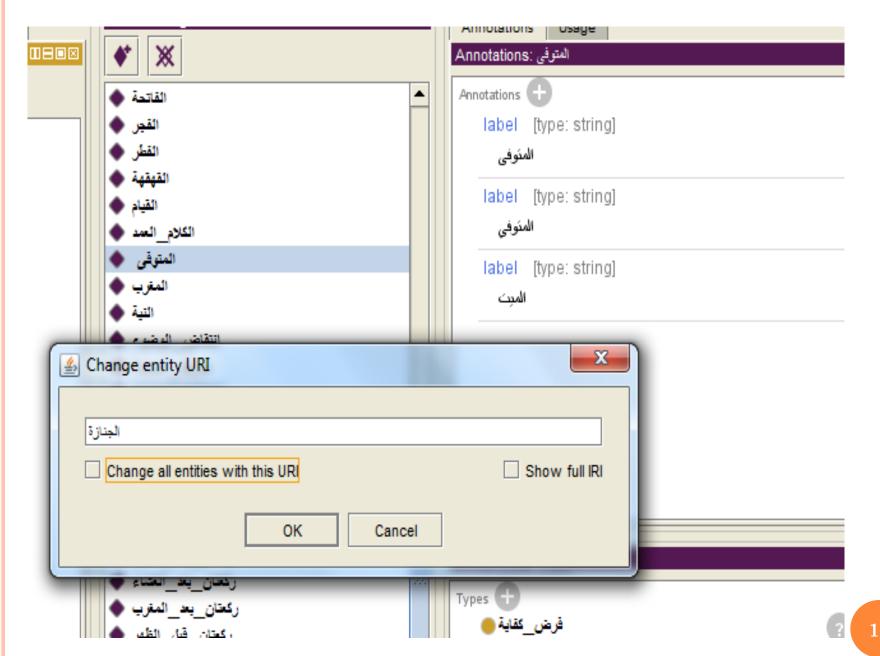
METHODOLOGY AND MODEL

Building Arabic Ontology Domain "فقه الصلاة" (Prayer jurisprudence). The development of ontology consists of the following stages:

- Define concepts, i.e., classes based on studying and analyzing the domain.

- Define instances, i.e., real elements in our domain.

- Define relations among classes as a requirement to come up with the ontology.

- Enrich ontology with Synonyms and Stemming words.
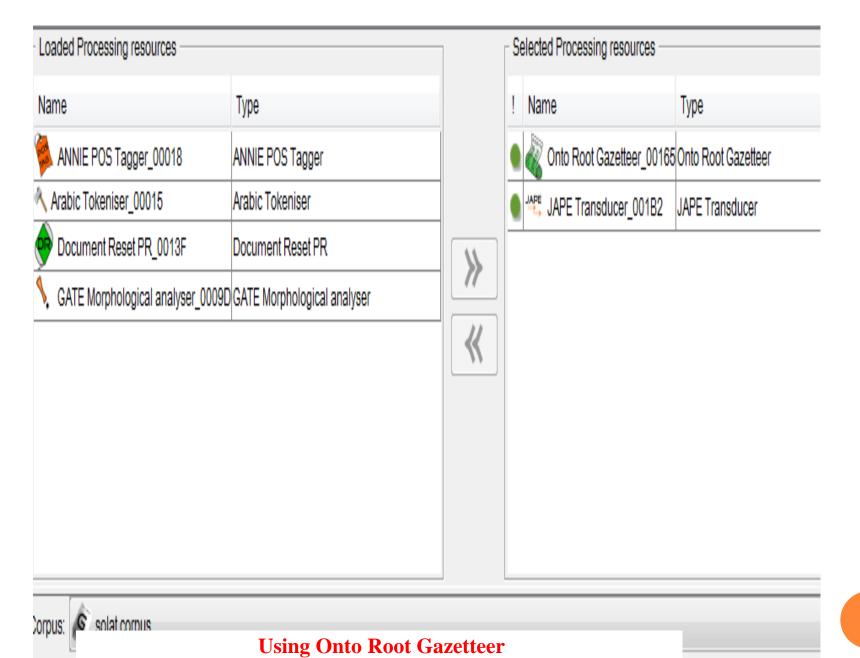
- Ontology Evaluation

8

No.	Classes /Arabic	Classes /English	Description
1	وقت الصلاة	Prayer Time	The time of FardhuAin prayer
2	الآذان	Aladan	Aladan is the call to prayer itself, and the person who calls it is called the muadhan.
3	السهو	Omission	Forget one of the prayer steps
4	سهو زيادة	Increase Omission	Either increase in acts or statements when the person does the prayer
5	سهو شك	Omission Doubt	Doubt between the two things, whichever is signed throughout the prayer
6	سهو نقصان	Decrease Omission	Either increase in acts or statements when the person does the prayer
7	شروط الصلاة	Prayer Conditions	Matters that are not part of the prayer, but must be satisfied before starting the prayer
8	شروط صحة	Validity Conditions	Conditions of prayer being valid refer to that on which the validity of prayer depends, such that if one of these conditions is broken, then prayer is not valid as a result.
9	شروط وجوب	Obligation Conditions	Conditions of prayer must be available in the person who want to pray to be his prayer right.
10	صلاة التطوع	Voluntary Prayer	It is the optional prayer can do beside the obligatory prayer
11	رواتب سنن	AlRoateb Sunan	Beyond the five daily required prayers, Muslims often engage in optional prayers before or after the regular prayers (FardhAin). These are known as "AlRoateb Sunan".

Ontology Classes

No.	Classes /Arabic	Classes /English	Description
12	رواتب بعدية	Post-Roateb	It is done after the FardhuAin prayer
13	رواتب قبلية	Pre-Roateb	It is done before the FardhuAin prayer
14	صلاة العيد	Eid Prayer	Eid prayer is performed on the morning of <u>Eid ul-Fitr</u> and <u>Eid ul-Adha</u>.
15	صلاة أهل الأعذار	Prayer of Exempted People	Persons who have a problem which can't do the prayer in suitable way.
16	صلاة فرض	Obligatory Prayer	The prayer must done by every person
17	فرض عين	FardhuAin	It is the main five prayers that done by person who want to pray.
18	فرض كفاية	FardhuKifayah	Prayer that carried out by one fall for others
19	مكونات الصلاة	Prayer Components	The main components for prayer and must be found in it include (Staff, Disliked, things which invalidate and Musthbat).
20	أركان	Staff	It is one of the important components of prayer related with the practical side.
21	مكروهات	Disliked	Things that are unlike in prayer
22	مبطلات	Things which Invalidate	Things make prayer wrong
23	مستحبات	Musthbat	Things that are preferred in the prayer

Afraid خائف

Passenger المركوب

Patient يمرض

Traveller السفر

Omission سهو

Prayer of Exempted People صلاة_أهل_الأعذار

Prayer jurisprudence فقه_الصلاة

Obligatory Prayer صلاة_فرض

Prayer Times أوقات

Asr العصر

Maghrib المغرب

Isha العشاء

FardhuAin المفروضة

Duhr الظهر

Prayer Conditions شروط_الصلاة

Fajr الفجر

Part of Ontology Concepts and Instances

11

Synonyms Words for Instance "الجنازة"(Funeral)

Loaded Processing resources

Name	Type
ANNIE POS Tagger_00018	ANNIE POS Tagger
Arabic Tokeniser_00015	Arabic Tokeniser
Document Reset PR_0013F	Document Reset PR
GATE Morphological analyser_0009D	GATE Morphological analyser

Selected Processing resources

!	Name	Type
●	Onto Root Gazetteer_00165	Onto Root Gazetteer
●	JAPE Transducer_001B2	JAPE Transducer

Corpus: solat corpus

No processing resource selected...

Using Onto Root Gazetteer

The Annotation Process Result

THE MODEL STRUCTURE

User Interface Part

Document annotation and Retrieval Part

Synonyms and Stemming for ontology elements

List of Documents

Input Query

Ontology

Corpus

Information Retrieval Process

Results

Annotator
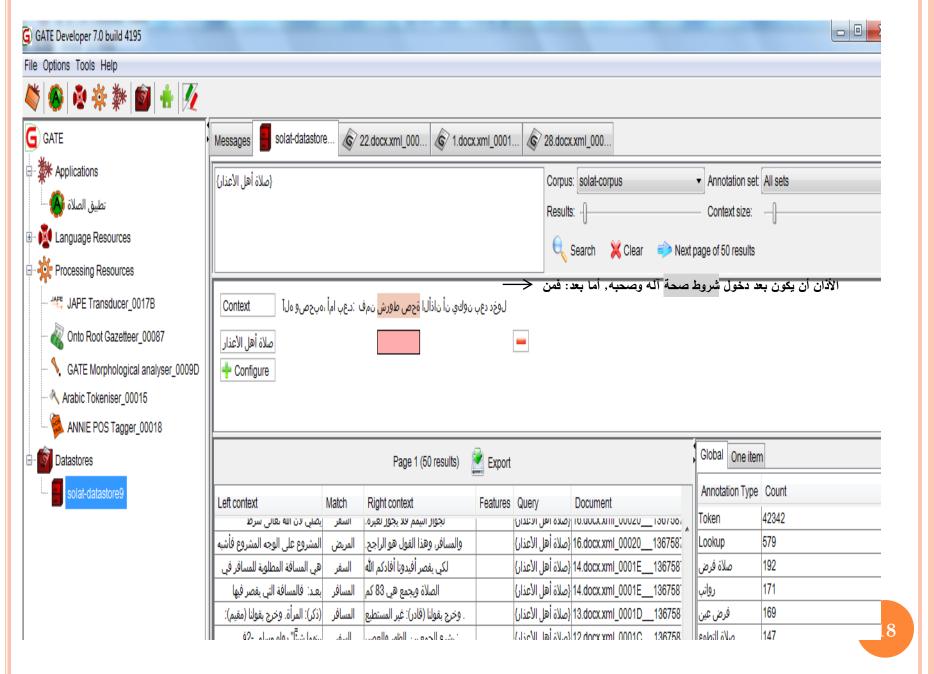
Apply Jape rules

List of Annotated Indexed Documents

15

Tools and Programs

o For indexing and keyword searching we use Lucene Datastore search engine.

o Protégé for ontology building.

o Gate as environment to execute all our work.

16

SYSTEM REALIZATION, EXPERIMENTAL RESULTS AND EVALUATION

System Interface

- **Applications:** in this part we execute our application which we name it "تطبيق الصلاة"(Prayer Application), by adding the plugins and Jape rules in its pipeline.
- **Language Resources (LRs):** represent entities such as lexicons, corpora or ontologies.
- **Processing Resources (PRs):** represent entities that are primarily algorithmic such as parsers.
- **Data stores**: specialized folder on a hard drive used to store the annotated corpus and improve processing times for large collections of documents.
- **Text area:** view the document before and after the annotation.
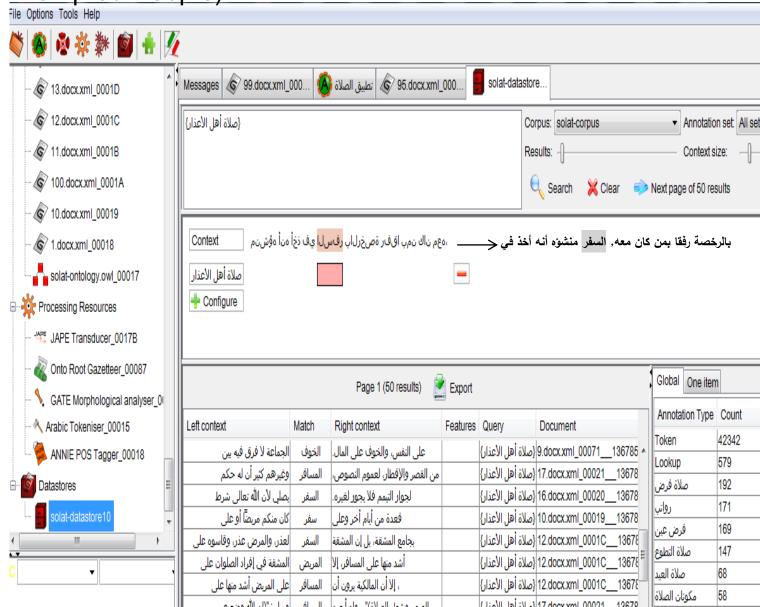
System Gate Interface

Experiments

- We performed a series of experiments to demonstrate the ability of our system to retrieve the related documents.

- All our experiments depend on the annotation types (ontology classes) that created from the processing of annotated documents using Jape rules.

- We give some examples to demonstrate and test the prototype and search using the annotation types that come up with the process of documents annotation.
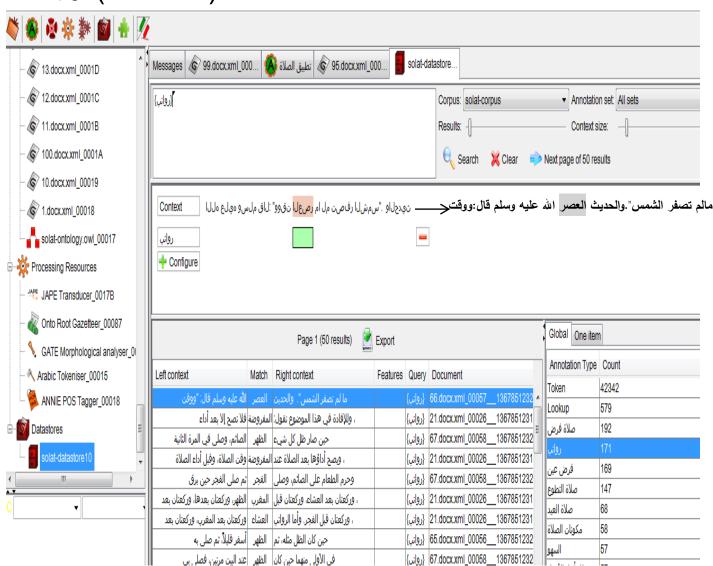
19

EXPERIMENTS

- The first three examples showing the results of a search using three annotation types "صلاة أهل الأعذار"(Prayer of Exempted People), "رواتب"(Roateb) and "الآذان"(Aladan).

- The last example for using the word "الآذان"(Aladan) as keyword (traditional way) in the search.

20

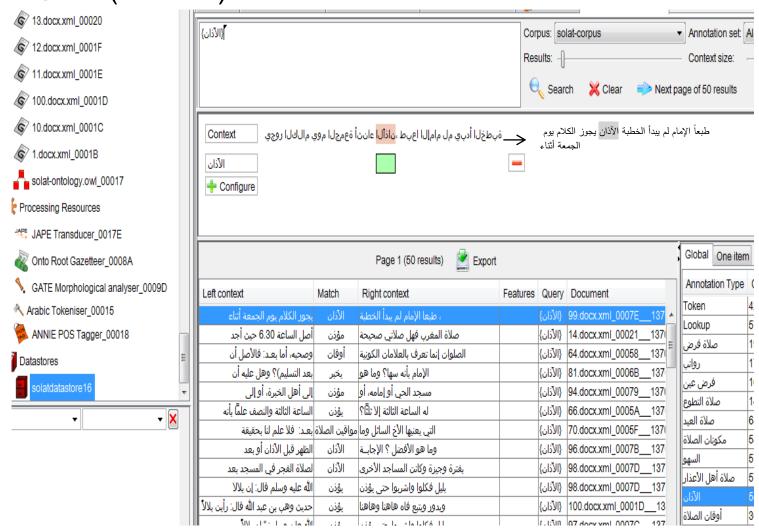**Example 1**. Searching using annotation type "صلاة أهل الأعذار"(Prayer of Exempted People).
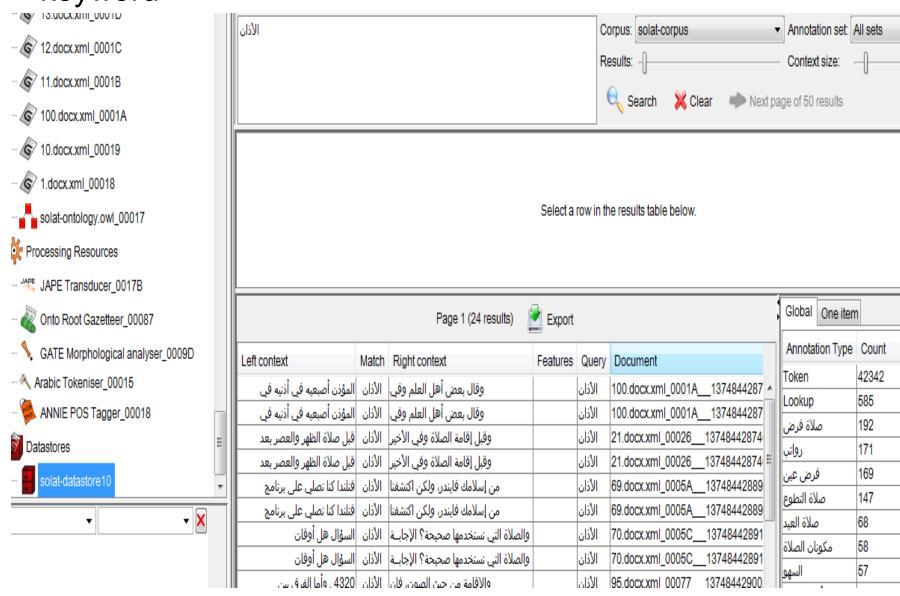
Example 2. Searching using annotation type "رواتب"(Roateb).

Example 3. Searching using annotation type "الآذان"(Aladan).

Example 4. Searching using the word "الآذان"(Aladan) as keyword

SYSTEM EVALUATION

• System evaluation depends on finding all related documents to the ontology components. We use 100 documents in our related Arabic Ontology Domain "فقه الصلاة" (Prayer jurisprudence) then we used the Gate tool to automatically annotate these documents, based on the Onto Root Gazetteer annotator.

• We depend on two important measures which are commonly used to evaluate such a system: precision and recall.
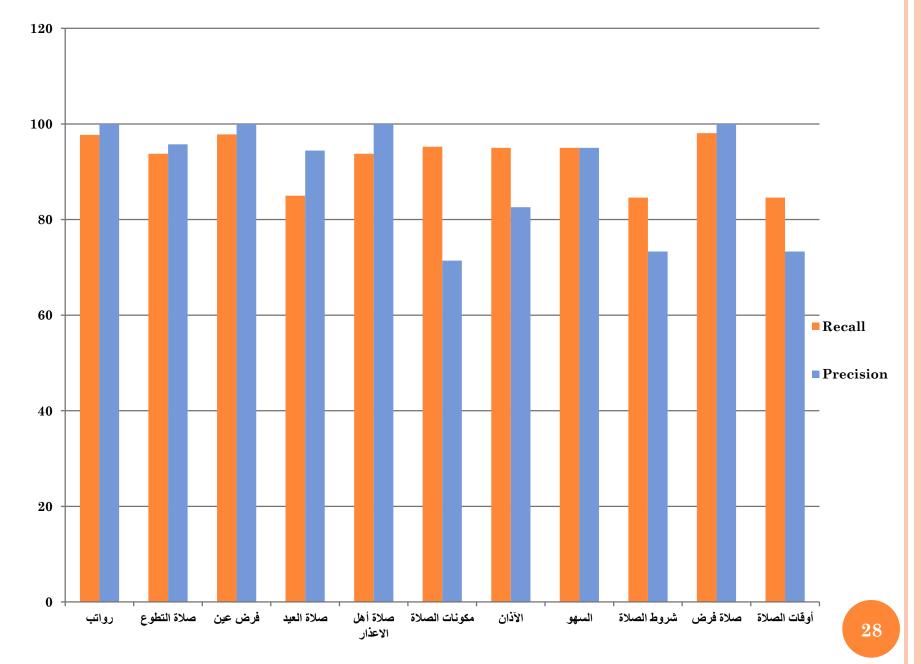
SYSTEM EVALUATION

Recall: is defined as the number of relevant documents retrieved by a search divided by the total number of existing relevant documents (which should have been retrieved .

$$\text{Recall} = \frac{\text{Number of retrived relevant docuemnts}}{\text{Number of relevant docuemnts}}$$

Precision: is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search.

$$\text{Precison} = \frac{\text{Number of retrievd relevant documents}}{\text{Number of retreivd docuemnts}}$$

26

Precision and Recall Results

Annotation Types	Recall	Precision
رواتب	97.72	100
صلاة التطوع	93.75	95.74
فرض عين	97.82	100
صلاة العيد	85	94.44
صلاة أهل الاعذار	93.75	100
مكونات الصلاة	95.23	71.42
الآذان	95	82.60
السهو	95	95
شروط الصلاة	84.61	73.33
صلاة فرض	98.07	100
أوقات الصلاة	84.61	73.33

Recall and Precision for Every Annotation Type

CONCLUSION AND FUTURE WORK

Our contribution in this work includes the following:

- Building and evaluating a domain specific ontology " فقه" (الصلاة)(Prayer jurisprudence)

- Building automatic ontology-based document annotation for Arabic information retrieval model used in the process of documents annotation and retrieval.

- A model that covers an important issue in the field of " فقه" (الصلاة)(Prayer jurisprudence) for users who are interested in the part of Islamic issues.

- Adaptation of GATE to work with Arabic documents specially when we use lucene Datastore search engine.

CONCLUSION AND FUTURE WORK

This work can be improved in multiple directions:

• Extending the ontology by adding the other parts that have relation with the domain "فقه الصلاة" to include other issues related with Islam.

• Increasing corpus to retrieve more documents in the domain and obtain more accurate results.

• Extending system model to be online to help retrieve more and new documents in the field we work in it. This requests building in independent system out of the Gate environment.

THANK YOU

31