Passage retrieval and intellectual property in legal texts

Paolo Rosso
Joint work with Santiago Correa, Davide Buscaldi,
Natural Language Engineering Lab.
Dept. SIC, Universidad Politécnica de Valencia, Spain

In collaboration with Alfonso Rios
Maat Knowledge, Spain

FLACOS, Toledo, 24-25 September 2009

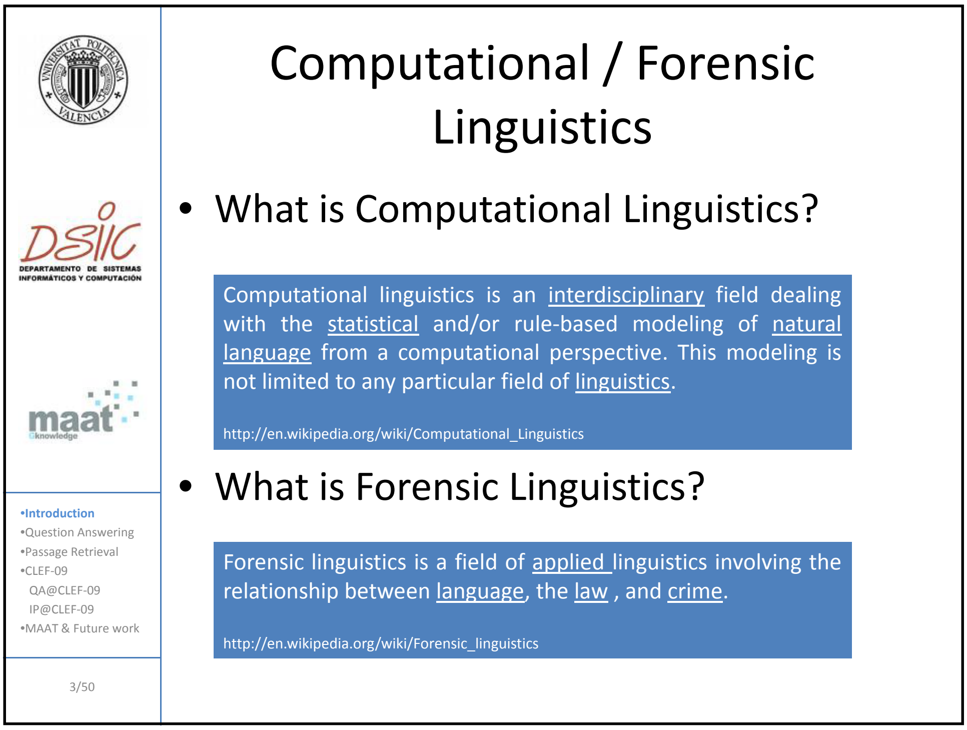
Overview

•Introduction
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

Introduction:

- Computational linguistics and forensic linguistics
- Tracks on legal texts
- Tracks on intellectual property (patents)
- Tracks on plagiarism detection (plagiarism of ideas)

- Question Answering
- Passage Retrieval
- CLEF-09
- QA@CLEF-09
- IP@CLEF-09
- Passage Retrieval at MAAT and Future work



Computational / Forensic Linguistics

- What is Computational Linguistics?

Computational linguistics is an interdisciplinary field dealing with the statistical and/or rule-based modeling of natural language from a computational perspective. This modeling is not limited to any particular field of linguistics.

http://en.wikipedia.org/wiki/Computational_Linguistics

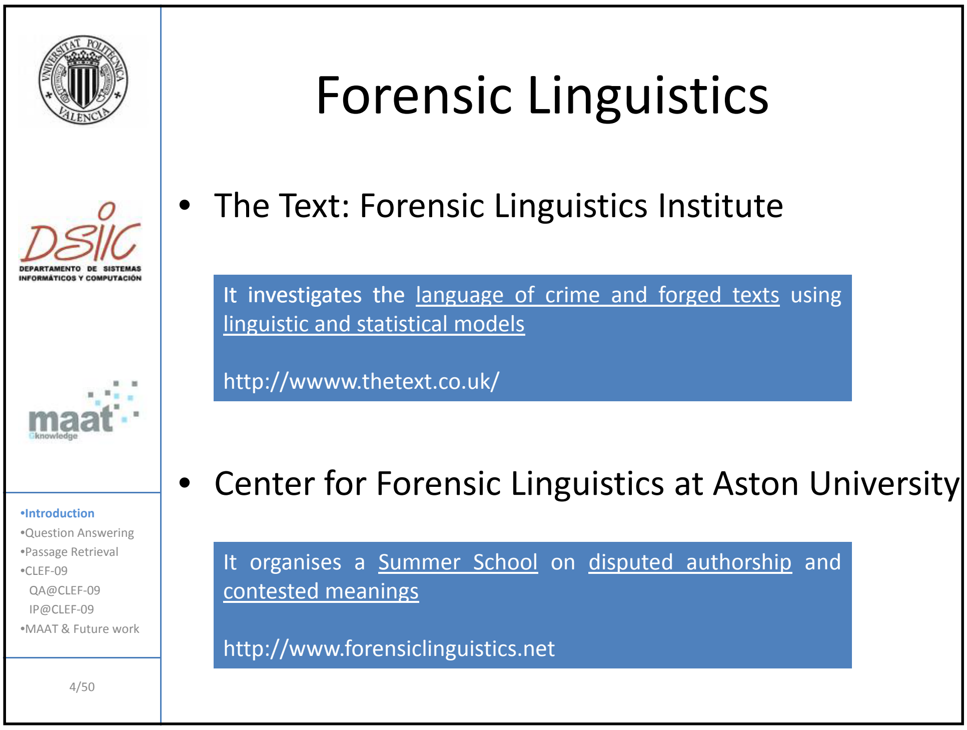
- What is Forensic Linguistics?

Forensic linguistics is a field of applied linguistics involving the relationship between language, the law , and crime.

http://en.wikipedia.org/wiki/Forensic_linguistics

•**Introduction**
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

3/50



Forensic Linguistics

- The Text: Forensic Linguistics Institute

It investigates the language of crime and forged texts using linguistic and statistical models

http://wwww.thetext.co.uk/

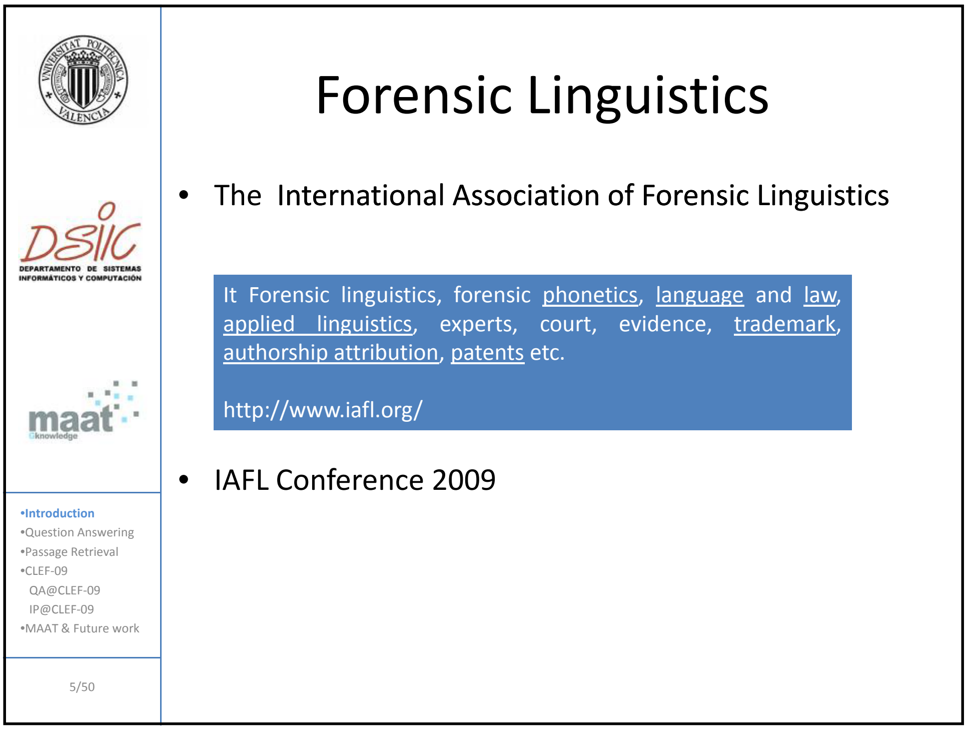
- Center for Forensic Linguistics at Aston University

It organises a Summer School on disputed authorship and contested meanings

http://www.forensiclinguistics.net

•**Introduction**
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

4/50



Forensic Linguistics

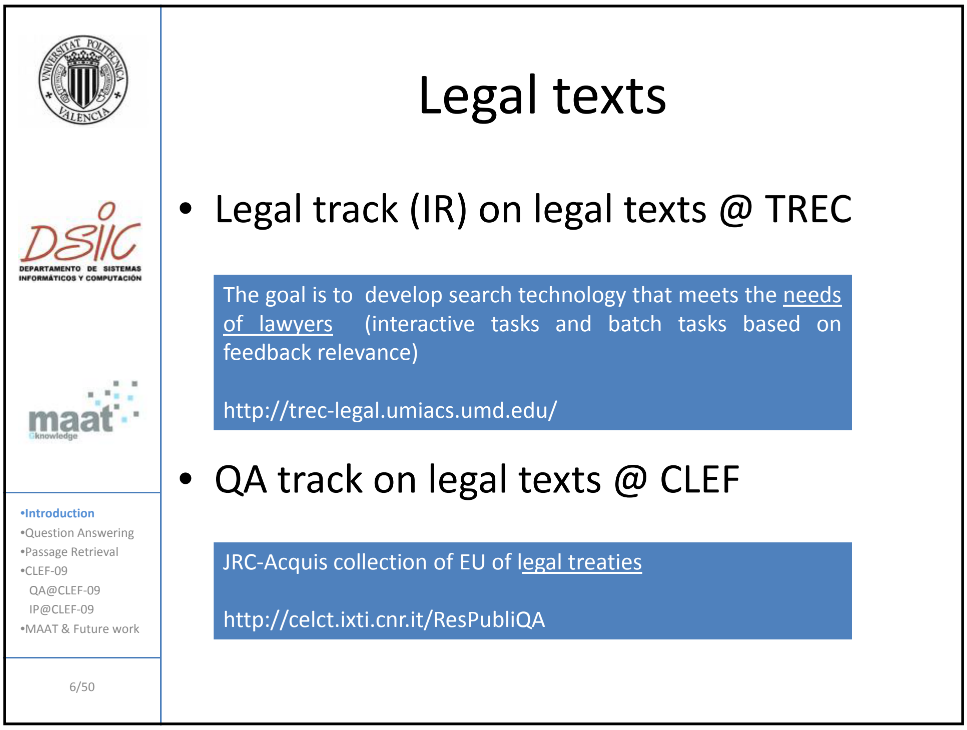
- The International Association of Forensic Linguistics

> It Forensic linguistics, forensic phonetics, language and law, applied linguistics, experts, court, evidence, trademark, authorship attribution, patents etc.
>
> http://www.iafl.org/

- IAFL Conference 2009

•Introduction
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

5/50



Legal texts

- Legal track (IR) on legal texts @ TREC

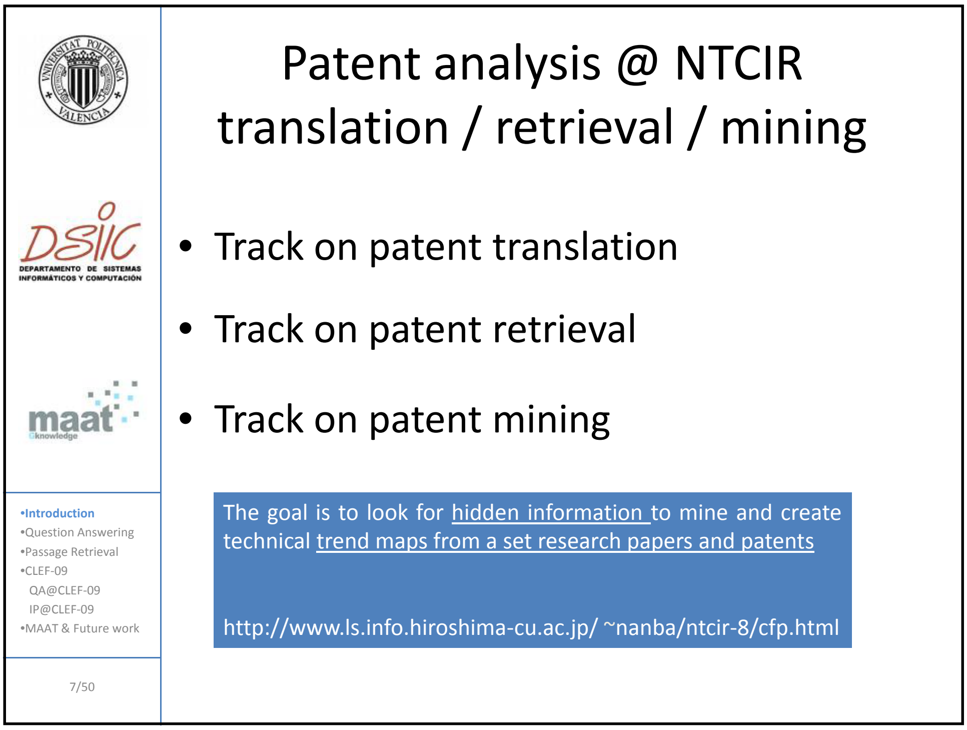
> The goal is to develop search technology that meets the needs of lawyers (interactive tasks and batch tasks based on feedback relevance)
>
> http://trec-legal.umiacs.umd.edu/

- QA track on legal texts @ CLEF

> JRC-Acquis collection of EU of legal treaties
>
> http://celct.ixti.cnr.it/ResPubliQA

•Introduction
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

6/50



Patent analysis @ NTCIR translation / retrieval / mining

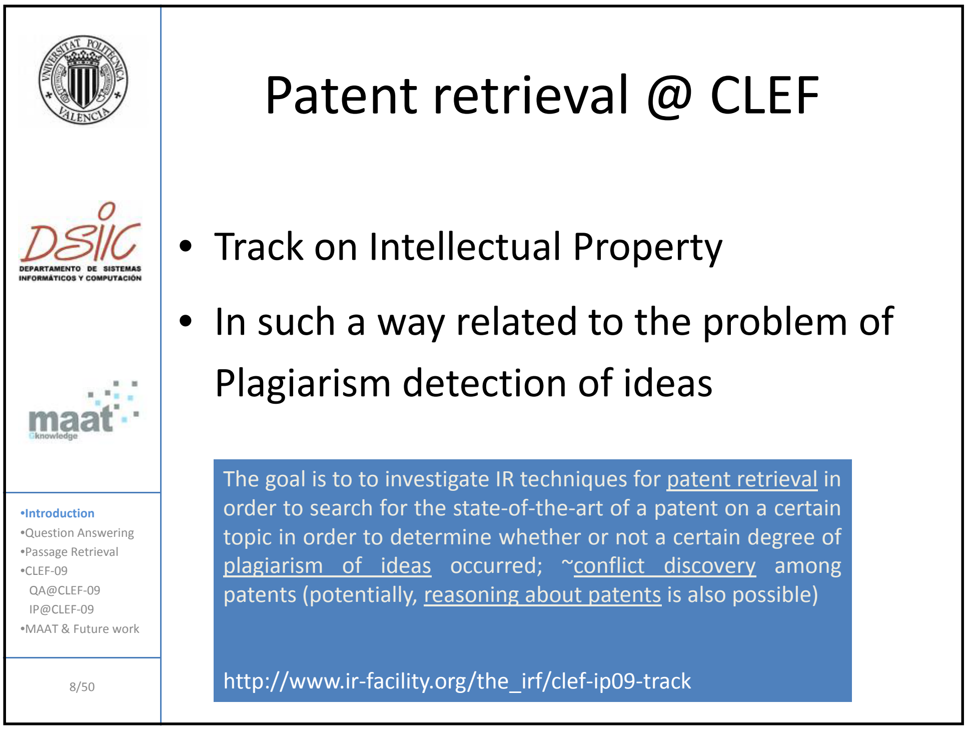
- Track on patent translation
- Track on patent retrieval
- Track on patent mining

The goal is to look for hidden information to mine and create technical trend maps from a set research papers and patents

http://www.ls.info.hiroshima-cu.ac.jp/ ~nanba/ntcir-8/cfp.html

•Introduction
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

7/50



Patent retrieval @ CLEF

- Track on Intellectual Property
- In such a way related to the problem of Plagiarism detection of ideas

The goal is to to investigate IR techniques for patent retrieval in order to search for the state-of-the-art of a patent on a certain topic in order to determine whether or not a certain degree of plagiarism of ideas occurred; ~conflict discovery among patents (potentially, reasoning about patents is also possible)

http://www.ir-facility.org/the_irf/clef-ip09-track

•Introduction
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

8/50

Intellectual Property (Patents)

- Information Retrieval Facility (IRF)

It is an open IR science institution: http://wwww.ir-facility.org

Mission:

-to bridge the gap between IR research and Industry and bring the latest IR technologies to the community of patent professionals

- to maintain a facility that enables large scale IR and in-depth patent processing

- to provide access to a large and high-quality corpus of patent data

-to support communication between the IP and IR communities

- to support and coordinate academic projects

Intellectual Property (Patents)

- Information Retrieval Facility (IRF)

Targets:

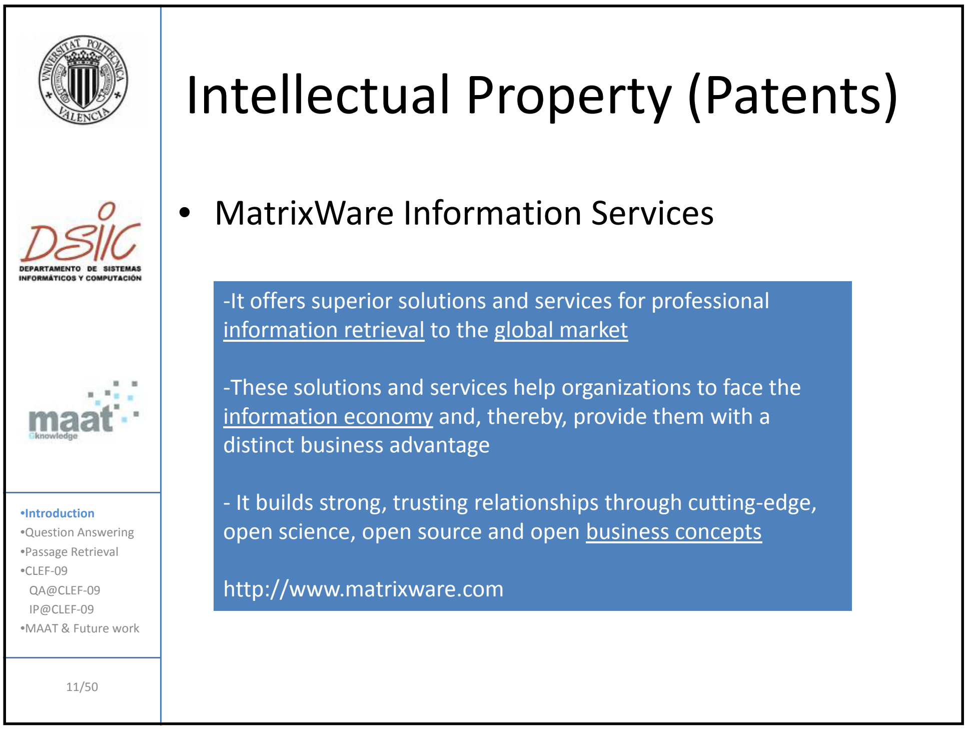
-Industrial information professionals

-Researchers in information retrieval

- Patent authorities and governmental institutions

Conference (next in Vienna, 2010):

http://www.ir-facility.org/irf-conference



Intellectual Property (Patents)

- MatrixWare Information Services

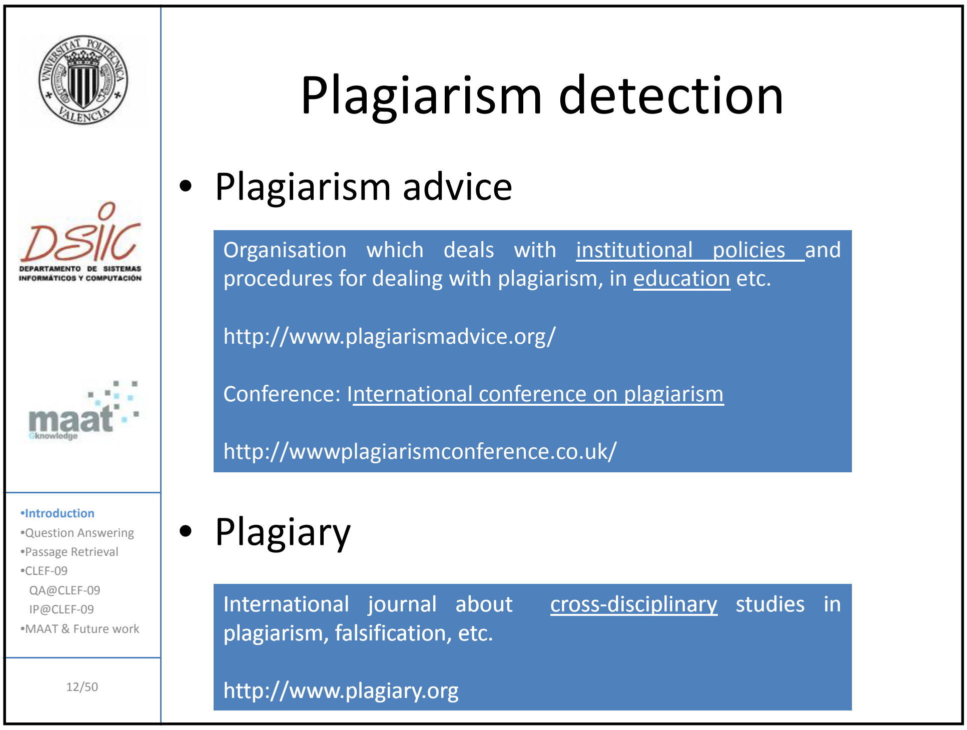
-It offers superior solutions and services for professional information retrieval to the global market

-These solutions and services help organizations to face the information economy and, thereby, provide them with a distinct business advantage

- It builds strong, trusting relationships through cutting-edge, open science, open source and open business concepts

http://www.matrixware.com

•Introduction
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work



Plagiarism detection

- Plagiarism advice

Organisation which deals with institutional policies and procedures for dealing with plagiarism, in education etc.

http://www.plagiarismadvice.org/

Conference: International conference on plagiarism

http://wwwplagiarismconference.co.uk/

- Plagiary

International journal about cross-disciplinary studies in plagiarism, falsification, etc.

http://www.plagiary.org

•Introduction
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

Plagiarism detection

- PAN workshop

Uncovering plagiarism, authorship and social software misuse (e.g. vandalism on Wikipedia)

3rd edition: http://www.webis.de/pan-09

- 1st Competition on plagiarism detection

Participants: 13 teams from (result ranking order): Germany, Czech Republic, Italy, Ukraine, Austria, Nigeria and South Korea, Greece, Israel, Brazil, Canada, Spain, UK, USA

Sponsor: Yahoo!Research

Aims: corpus creation, comparison of methods, introduction of (standard) measures

http://www.webis.de/pan-09/competition.php

•**Introduction**
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

13/50

Question Answering

- What is Question Answering?

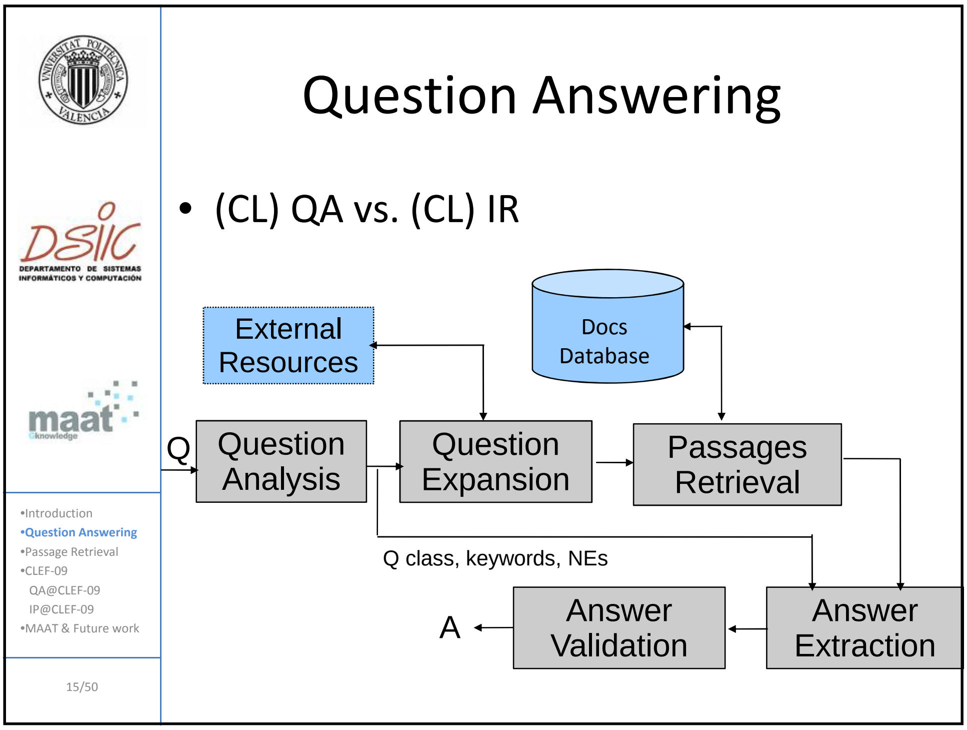
Question Answering can be viewed as a particular form of Information Retrieval (IR), in which the amount of information to return is the minimum required to satisfy the user needs expressed by a specific question.

In information retrieval, question answering (QA) is the task of automatically answering a question posed in natural language. To find the answer to a question, a QA computer program may use either a pre-structured database or a collection of natural language documents (a text corpus such as the World Wide Web or some local collection).

http://en.wikipedia.org/wiki/Question_answering

•Introduction
•**Question Answering**
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

14/50

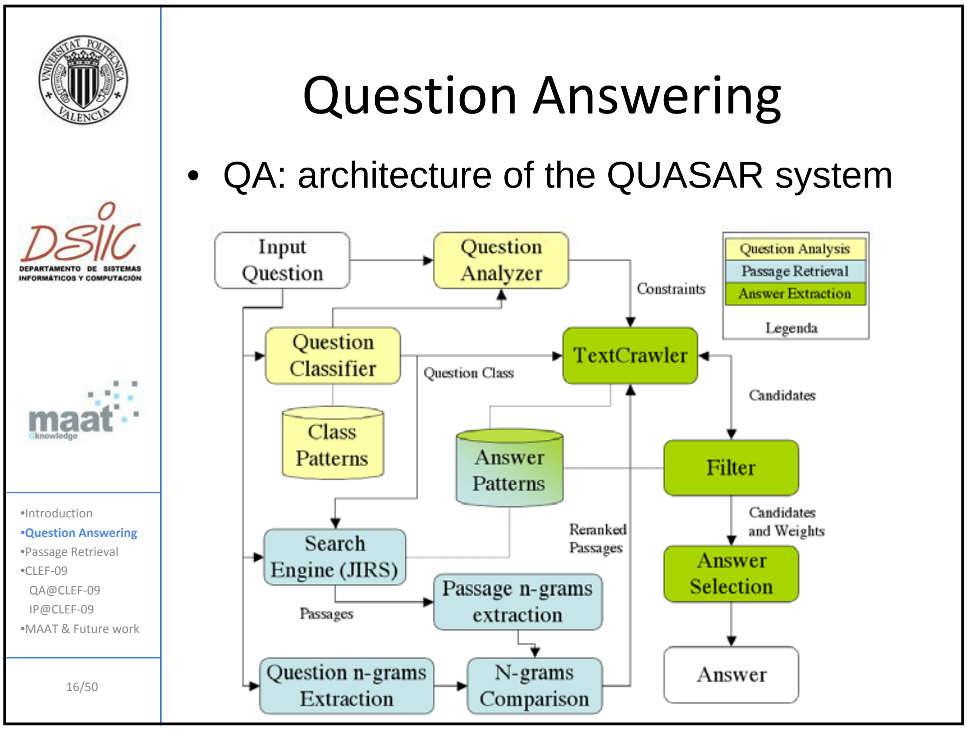


Question Answering

- (CL) QA vs. (CL) IR

•Introduction
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

15/50

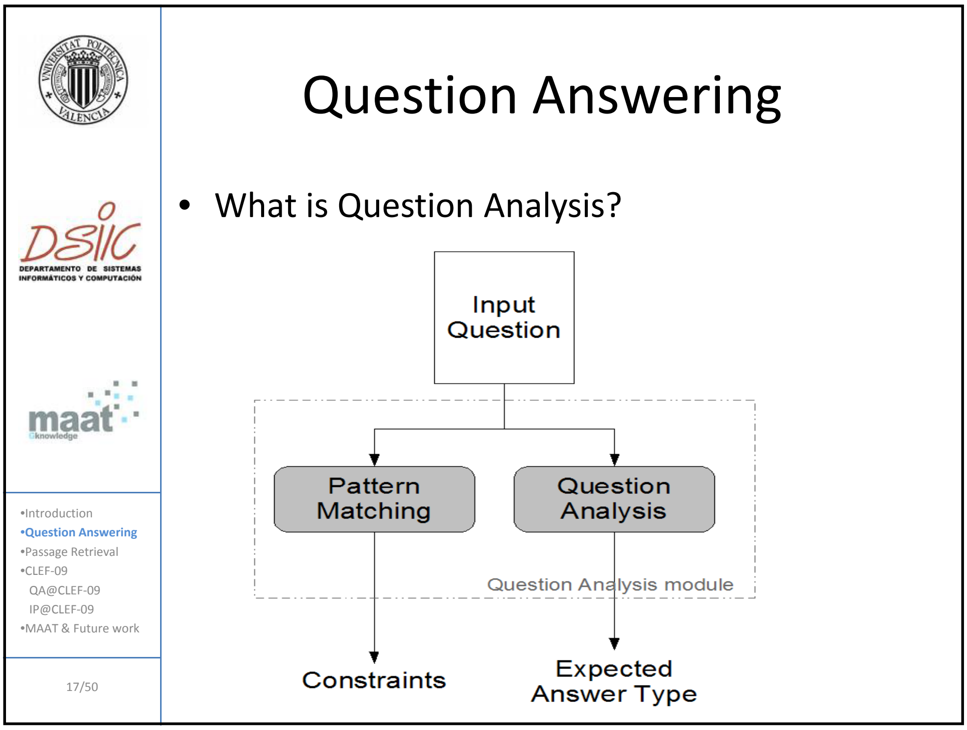


Question Answering

- QA: architecture of the QUASAR system

•Introduction
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

16/50



DSIC DEPARTAMENTO DE SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

maat Gknowledge

•Introduction
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

17/50

Question Answering

- What is Question Analysis?

Input Question

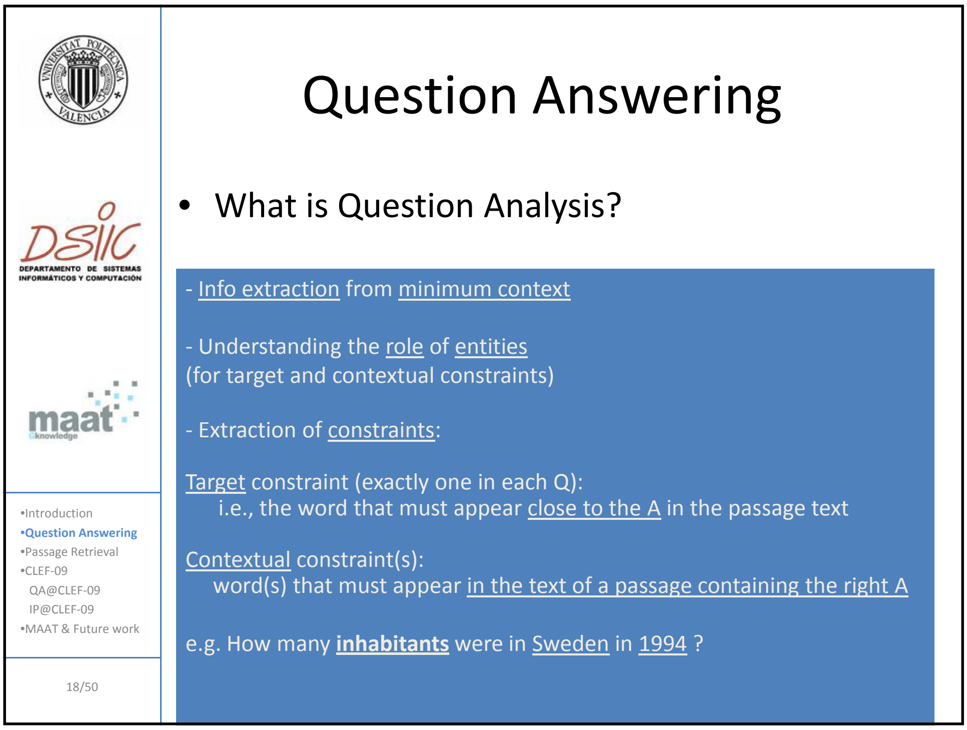
Pattern Matching

Question Analysis

Question Analysis module

Constraints

Expected Answer Type



DSIC DEPARTAMENTO DE SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

maat Gknowledge

•Introduction
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

18/50

Question Answering

- What is Question Analysis?

- Info extraction from minimum context

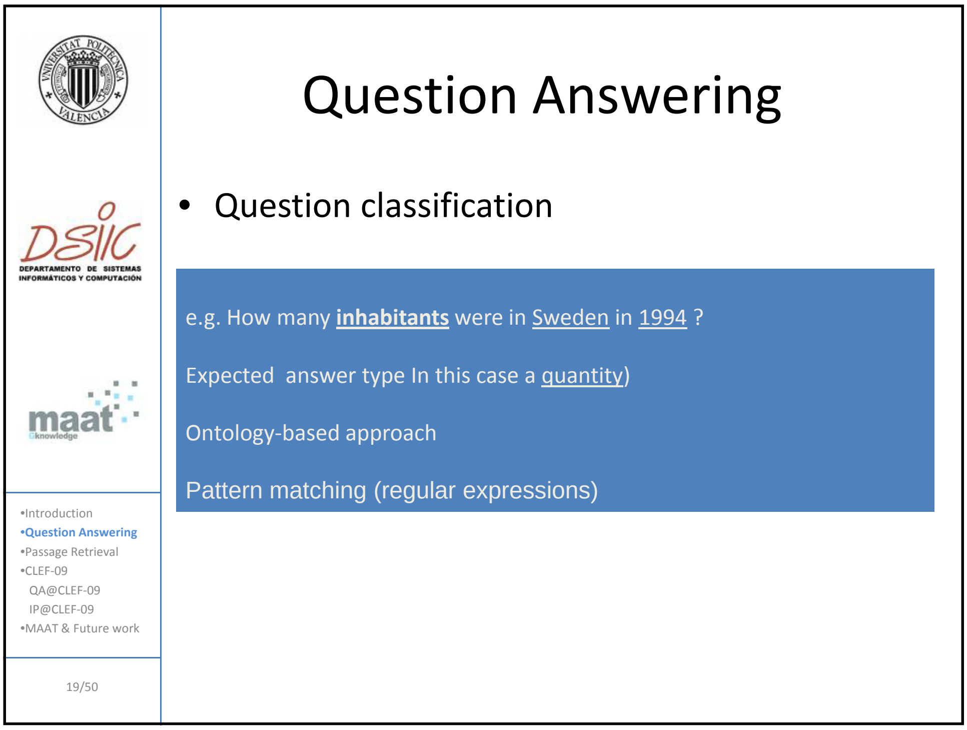
- Understanding the role of entities
(for target and contextual constraints)

- Extraction of constraints:

Target constraint (exactly one in each Q):
i.e., the word that must appear close to the A in the passage text

Contextual constraint(s):
word(s) that must appear in the text of a passage containing the right A

e.g. How many **inhabitants** were in Sweden in 1994 ?



Question Answering

- Question classification

e.g. How many **inhabitants** were in Sweden in 1994 ?

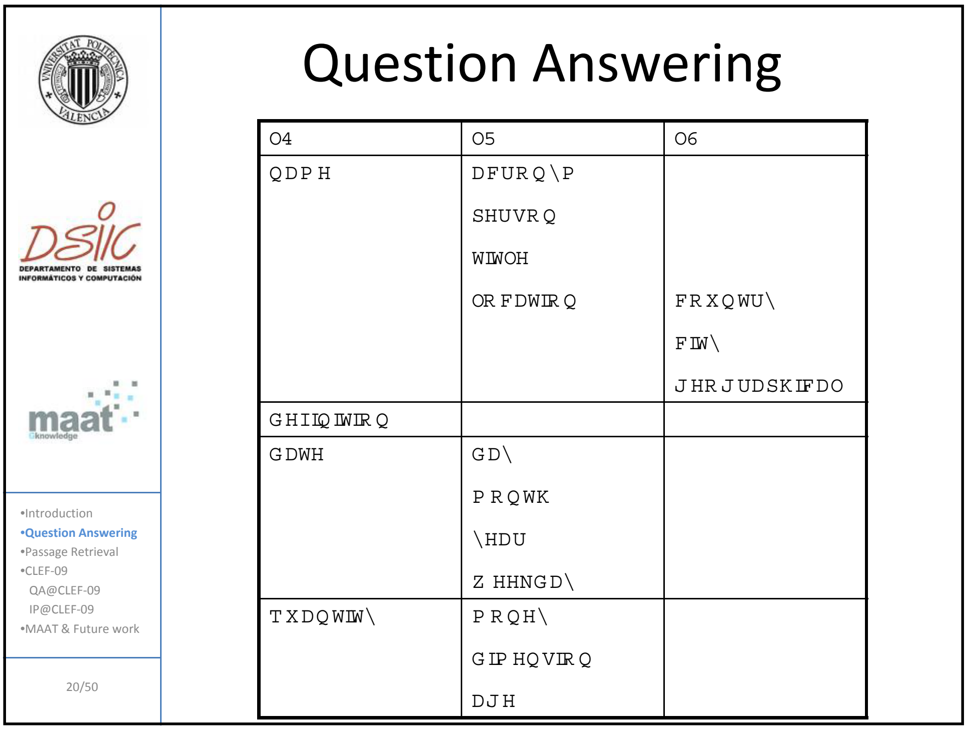
Expected answer type In this case a quantity)

Ontology-based approach

Pattern matching (regular expressions)

•Introduction
•**Question Answering**
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

19/50

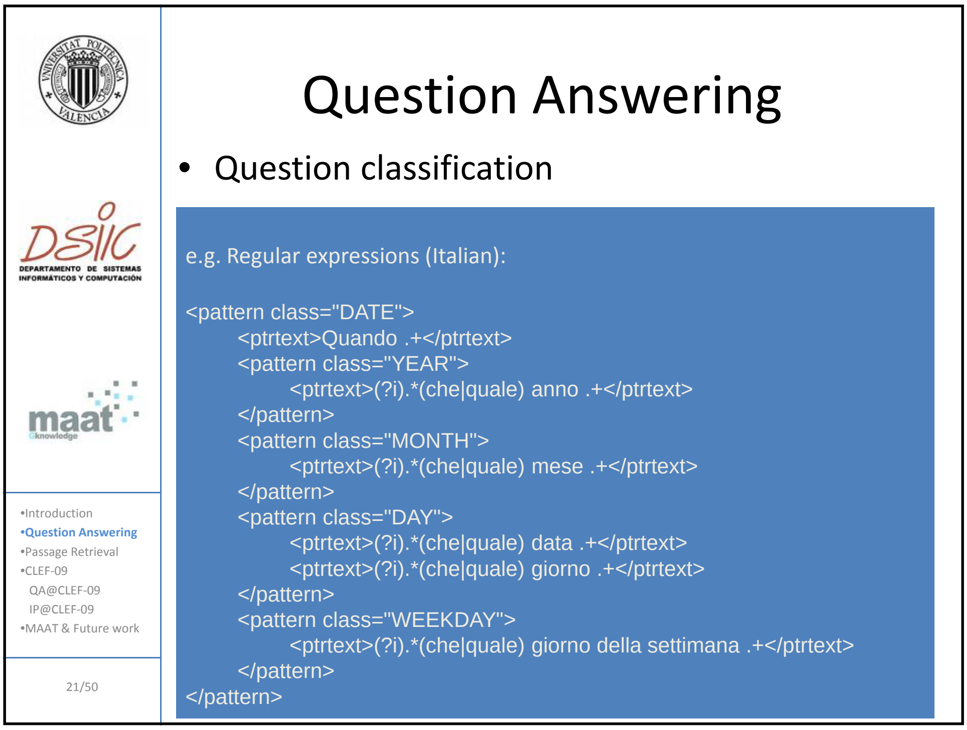


Question Answering

O4	O5	O6
QDPH	DFURQ\P SHUVRQ WIWOH ORFDWIRQ	FRXQWU\ FIW\ JHRJUDSKIFDO
GHIIQIWIRQ		
GDWH	GD\ PRQWK \HDU ZHHNGD\	
TXDQWIW\	PRQH\ GIPHQVIRQ DJH	

•Introduction
•**Question Answering**
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

20/50



Question Answering

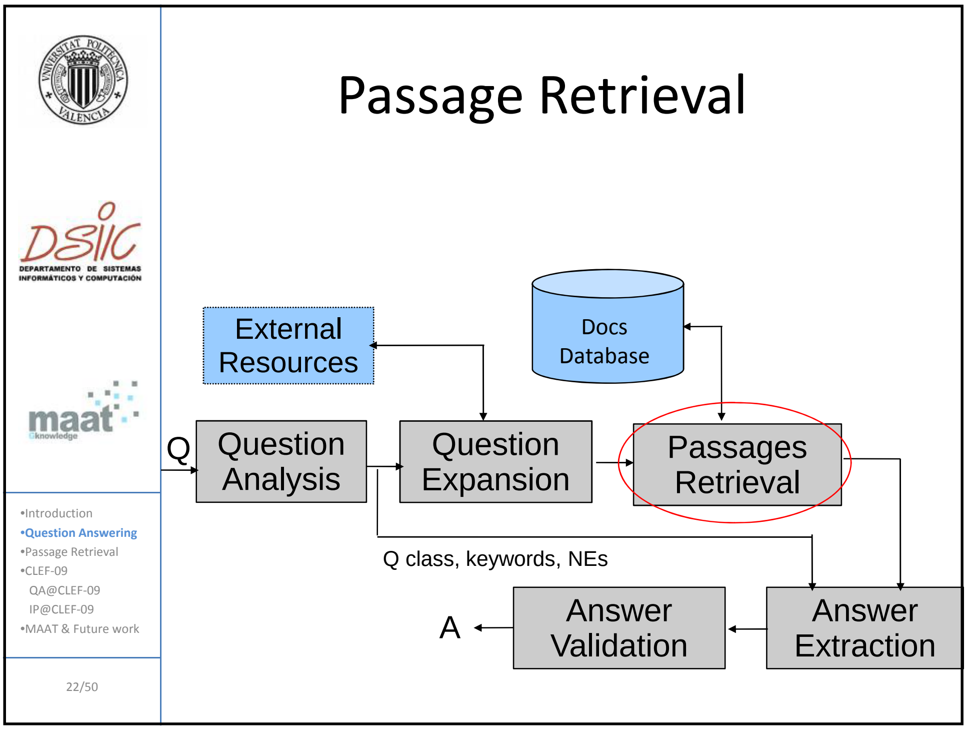
- Question classification

e.g. Regular expressions (Italian):

```
<pattern class="DATE">
        <ptrtext>Quando .+</ptrtext>
        <pattern class="YEAR">
                <ptrtext>(?i).*(che|quale) anno .+</ptrtext>
        </pattern>
        <pattern class="MONTH">
                <ptrtext>(?i).*(che|quale) mese .+</ptrtext>
        </pattern>
        <pattern class="DAY">
                <ptrtext>(?i).*(che|quale) data .+</ptrtext>
                <ptrtext>(?i).*(che|quale) giorno .+</ptrtext>
        </pattern>
        <pattern class="WEEKDAY">
                <ptrtext>(?i).*(che|quale) giorno della settimana .+</ptrtext>
        </pattern>
</pattern>
```

•Introduction
•**Question Answering**
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

21/50



Passage Retrieval

External Resources

Docs Database

Q

Question Analysis

Question Expansion

Passages Retrieval

Q class, keywords, NEs

A

Answer Validation

Answer Extraction

•Introduction
•**Question Answering**
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

22/50

DEPARTAMENTO DE SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

- Introduction
- Question Answering
- **Passage Retrieval**
- CLEF-09
  - QA@CLEF-09
  - IP@CLEF-09
- MAAT & Future work

23/50

Passage Retrieval

- What is Passage Retrieval?

A Passage Retrieval (PR) system is an IR system which, given a list of keywords (e.g.: "Electricity," "Motor", etc..) or a question such as:

e.g. Where is the Europol Drugs Unit?

A PR returns fragments of texts (passages) that are relevant to the user needs

JIRS is a open-source PR, developed in the UPV:

http://sourceforge.net/projects/jirs/

DEPARTAMENTO DE SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

- Introduction
- Question Answering
- **Passage Retrieval**
- CLEF-09
  - QA@CLEF-09
  - IP@CLEF-09
- MAAT & Future work

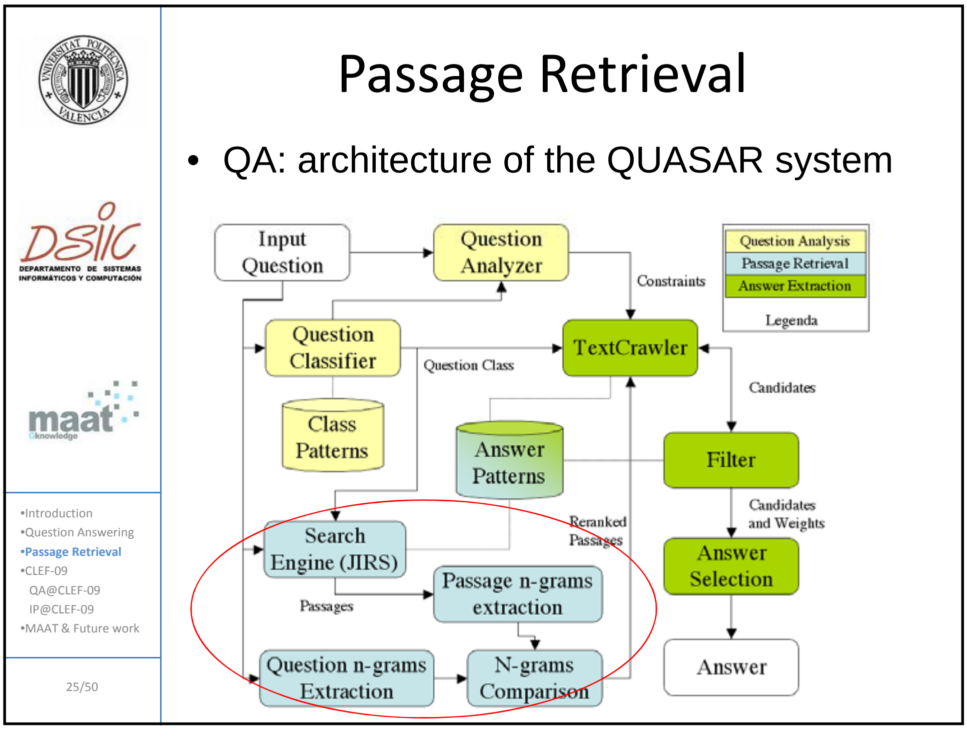
24/50

Passage retrieval system: JIRS

Most of nowadays PR systems are not oriented to the specific question answering problem, because they only take into account the keywords of the question in order to obtain the relevant passages.

JIRS is a PR engine based on n-grams

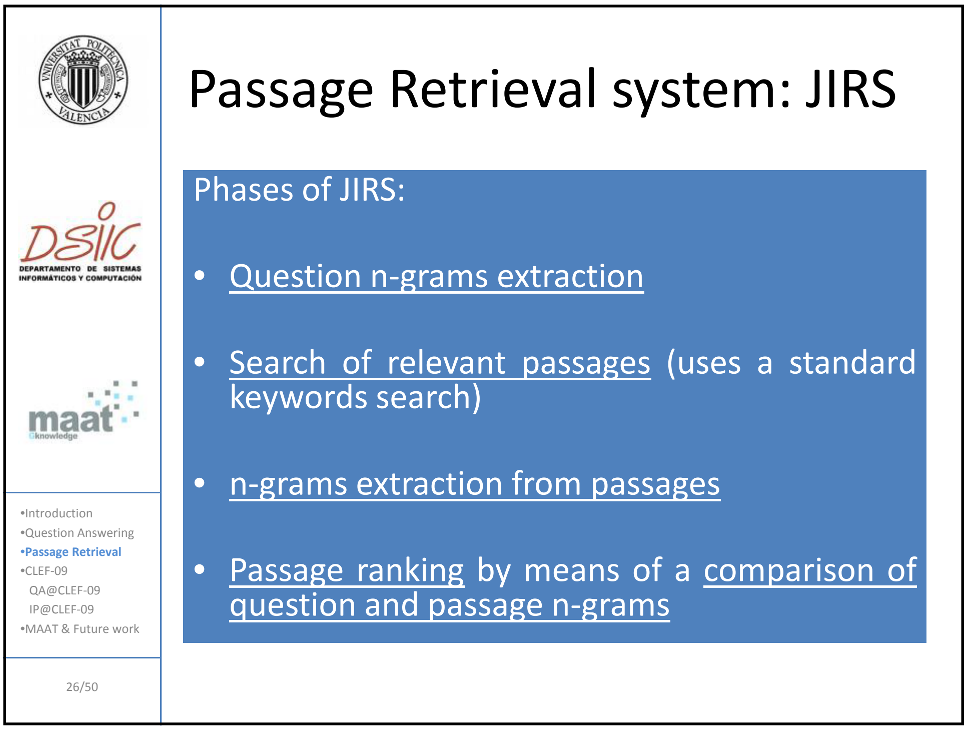
JIRS is based on the premise that in a large collection of documents, an n-gram associated with a question must be found in this collection at least once (redundancy)



Passage Retrieval

- QA: architecture of the QUASAR system

•Introduction
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work



Passage Retrieval system: JIRS

Phases of JIRS:

- Question n-grams extraction
- Search of relevant passages (uses a standard keywords search)
- n-grams extraction from passages
- Passage ranking by means of a comparison of question and passage n-grams

•Introduction
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

Passage Retrieval system: JIRS

JIRS example:

Let us suppose that we have a database of <u>publications of a newspaper</u>. Using the JIRS system we aim at finding in the document of the collection an answer to a question such as:

e.g. Who is the president of Colombia?"

For instance, the system could retrieve the following two passages:

"... Álvaro Uribe is the president of Colombia ..." and
"...Giorgio Napolitano is the president of Italy...".

Of course, the first passage should be given more importance because it contains the <u>5-gram</u> "is the president of Colombia", whereas the second passage contains only the <u>4-gram</u> "is the president of".

- Introduction
- Question Answering
- **Passage Retrieval**
- CLEF-09
  - QA@CLEF-09
  - IP@CLEF-09
- MAAT & Future work

27/50

JIRS: Weighting

The <u>weight of each term</u> is set to:

$$w_k = 1 - \frac{\log(n_k)}{1+\log(N)}$$

Where n_k is the number of passages in which the term appears and N is the total number of passages

The target is to establish a <u>measure of similarity between a passage</u> (*d*) and <u>a text</u> (*q*).

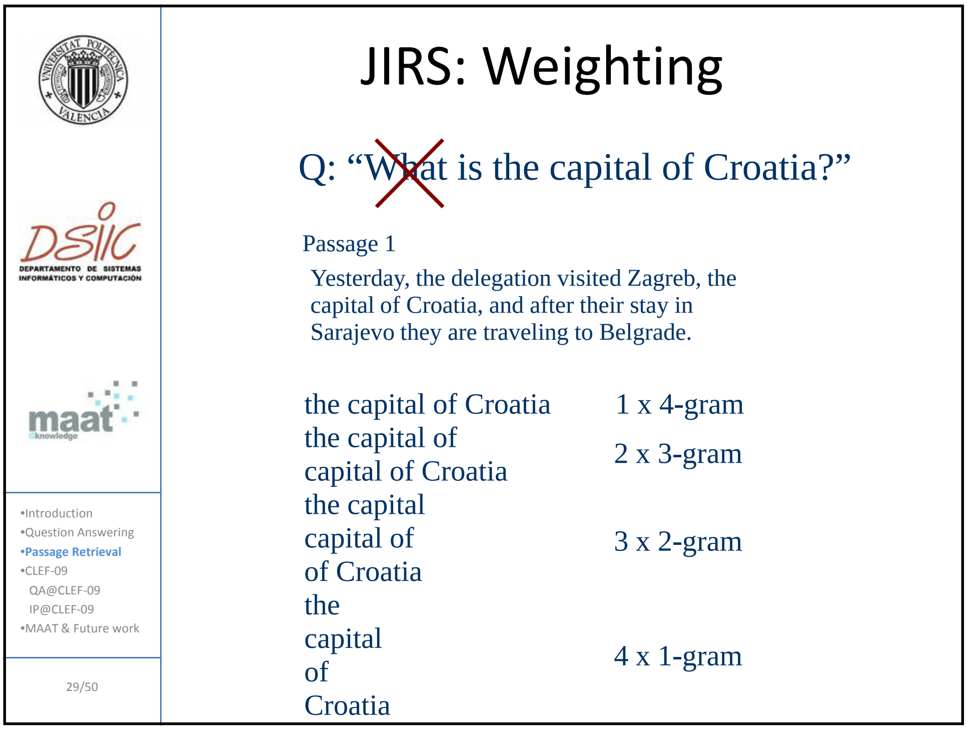
$$sim(d,q) = \frac{\sum_{j=1}^{n} \sum_{x \in Q} h(x, D_j)}{\sum_{j=1}^{n} \sum_{x \in Q} h(x, Q_j)}$$

The function *h(x, D_j)*, returns a weight for the <u>j-gram</u> *x* with respect to the <u>set of j-grams</u> D_j in the passage and it is defined as:

$$h(x, D_j) = \begin{cases} \sum_{k=1}^{|x|} W_x & if\ x \in D_j \\ 0 & otherwise \end{cases}$$

- Introduction
- Question Answering
- **Passage Retrieval**
- CLEF-09
  - QA@CLEF-09
  - IP@CLEF-09
- MAAT & Future work

28/50



JIRS: Weighting

Q: “~~What~~ is the capital of Croatia?”

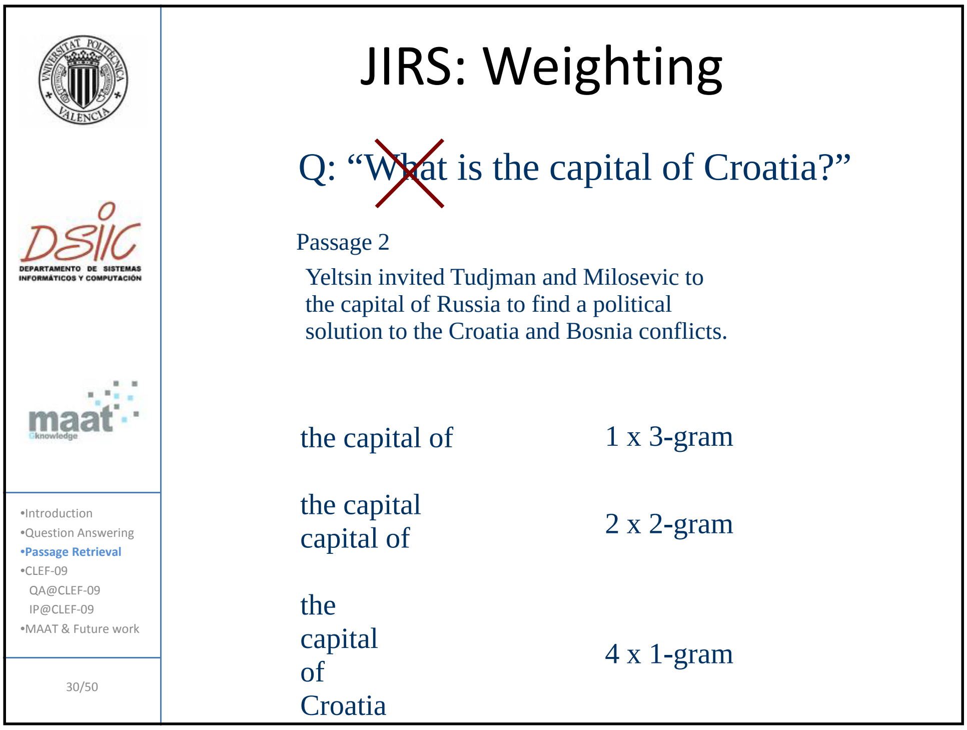
Passage 1

Yesterday, the delegation visited Zagreb, the capital of Croatia, and after their stay in Sarajevo they are traveling to Belgrade.

the capital of Croatia	1 x 4-gram
the capital of capital of Croatia	2 x 3-gram
the capital capital of of Croatia	3 x 2-gram
the capital of Croatia	4 x 1-gram

- •Introduction
- •Question Answering
- •Passage Retrieval
- •CLEF-09
  - QA@CLEF-09
  - IP@CLEF-09
- •MAAT & Future work

29/50



JIRS: Weighting

Q: “~~What~~ is the capital of Croatia?”

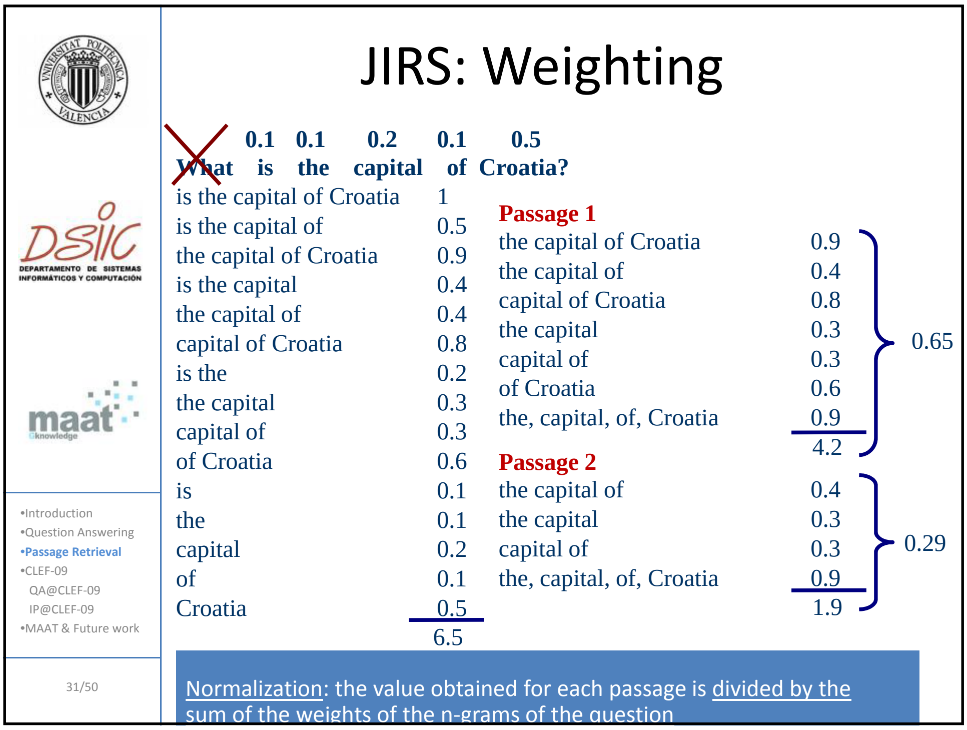
Passage 2

Yeltsin invited Tudjman and Milosevic to the capital of Russia to find a political solution to the Croatia and Bosnia conflicts.

the capital of	1 x 3-gram
the capital capital of	2 x 2-gram
the capital of Croatia	4 x 1-gram

- •Introduction
- •Question Answering
- •Passage Retrieval
- •CLEF-09
  - QA@CLEF-09
  - IP@CLEF-09
- •MAAT & Future work

30/50



JIRS: Weighting

~~What~~	is	the	capital	of	Croatia?
	0.1	0.1	0.2	0.1	0.5

is the capital of Croatia	1
is the capital of	0.5
the capital of Croatia	0.9
is the capital	0.4
the capital of	0.4
capital of Croatia	0.8
is the	0.2
the capital	0.3
capital of	0.3
of Croatia	0.6
is	0.1
the	0.1
capital	0.2
of	0.1
Croatia	0.5
	6.5

Passage 1

the capital of Croatia	0.9	0.65
the capital of	0.4	
capital of Croatia	0.8	
the capital	0.3	
capital of	0.3	
of Croatia	0.6	
the, capital, of, Croatia	0.9	
	4.2	

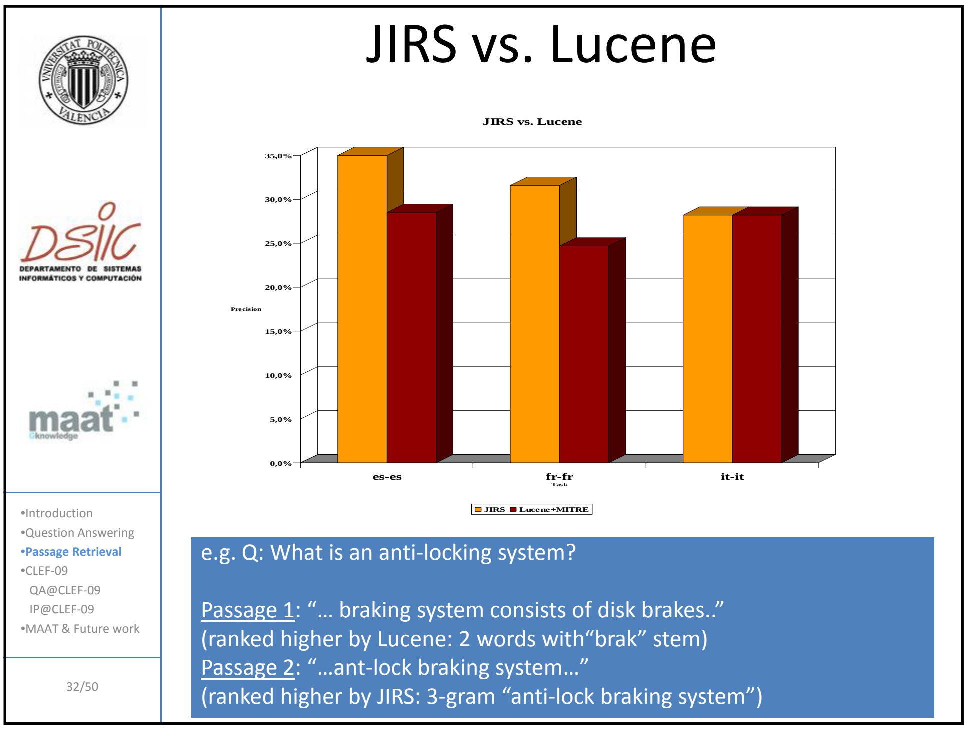
Passage 2

the capital of	0.4	0.29
the capital	0.3	
capital of	0.3	
the, capital, of, Croatia	0.9	
	1.9	

Normalization: the value obtained for each passage is divided by the sum of the weights of the n-grams of the question

•Introduction
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

31/50



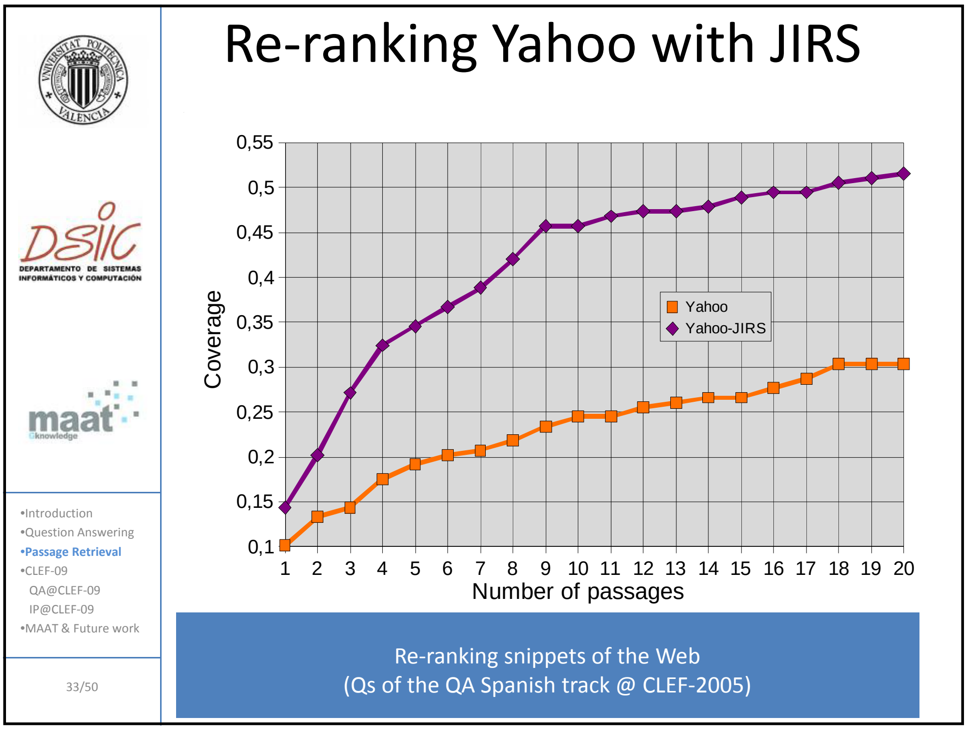
JIRS vs. Lucene

e.g. Q: What is an anti-locking system?

Passage 1: "… braking system consists of disk brakes.."
(ranked higher by Lucene: 2 words with"brak" stem)
Passage 2: "…ant-lock braking system…"
(ranked higher by JIRS: 3-gram "anti-lock braking system")

•Introduction
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

32/50

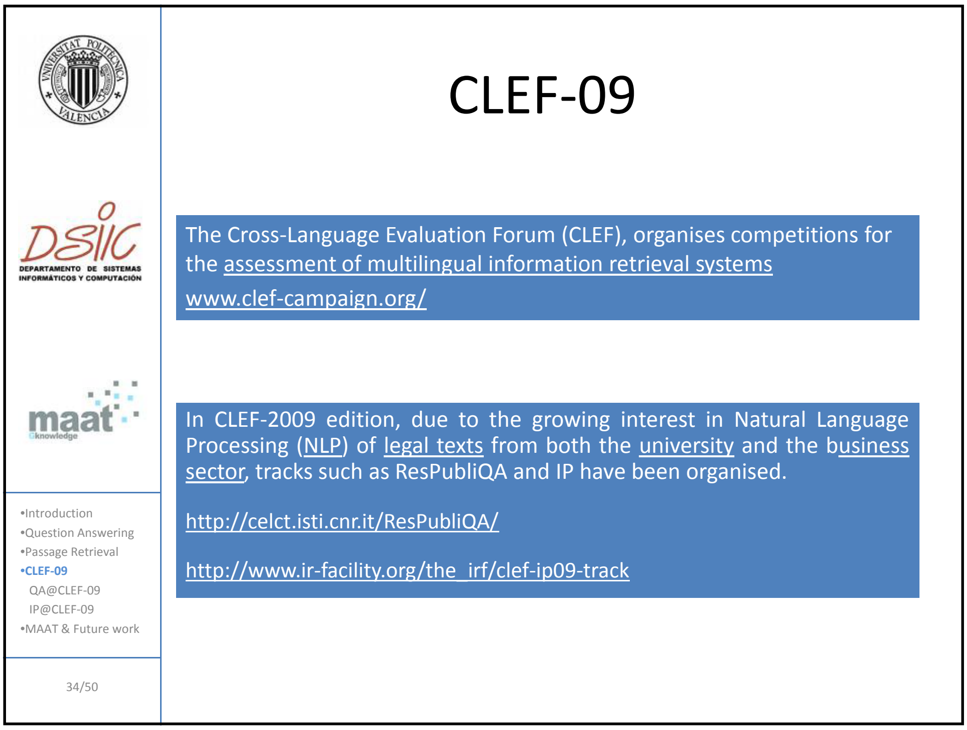


Re-ranking Yahoo with JIRS

Re-ranking snippets of the Web
(Qs of the QA Spanish track @ CLEF-2005)

- Introduction
- Question Answering
- **Passage Retrieval**
- CLEF-09
  - QA@CLEF-09
  - IP@CLEF-09
- MAAT & Future work

33/50



CLEF-09

The Cross-Language Evaluation Forum (CLEF), organises competitions for the assessment of multilingual information retrieval systems

www.clef-campaign.org/

In CLEF-2009 edition, due to the growing interest in Natural Language Processing (NLP) of legal texts from both the university and the business sector, tracks such as ResPubliQA and IP have been organised.

http://celct.isti.cnr.it/ResPubliQA/

http://www.ir-facility.org/the_irf/clef-ip09-track

- Introduction
- Question Answering
- Passage Retrieval
- **CLEF-09**
  - QA@CLEF-09
  - IP@CLEF-09
- MAAT & Future work

34/50

QA Track

ResPubliQA@CLEF-2009 competition address the problem of question answering in the restricted domain of legal texts (previous editions: open domain)

Given a pool of 500 independent natural language questions: e.g. Where is the Europol Drugs Unit?

Each system must return the passage (not the exact answer)

JRC-Acquis collection of EU legal traits (aligned documents): inter-language QA systems comparison

- Introduction
- Question Answering
- Passage Retrieval
- **CLEF-09**
  - QA@CLEF-09
  - IP@CLEF-09
- MAAT & Future work

IP Track

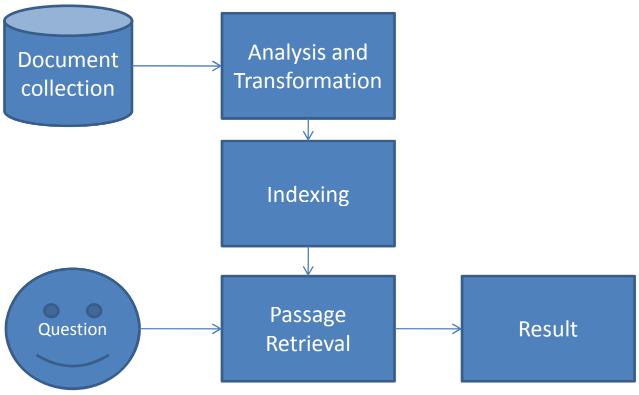
The CLEF IP track is coordinated by Information Retrieval Facility (IRF) and Matrixware . Its aim is to investigate IR techniques for patent retrieval in order to search for the prior state-of-the-art of a patent on a certain topic in order to determine whether or not a certain degree of plagiarism of ideas occurred.

The track provided a collection of more than 1M patent documents, mainly derived from European Patent Office sources, in three languages: English French and German.

A total of 500 patents are analysed using the supplied corpus to determine their prior state-of-the art; for each one of them the systems must return a list of 1000 documents with their score ranking.

- Introduction
- Question Answering
- Passage Retrieval
- **CLEF-09**
  - QA@CLEF-09
  - IP@CLEF-09
- MAAT & Future work

QA@CLEF-09: Approach

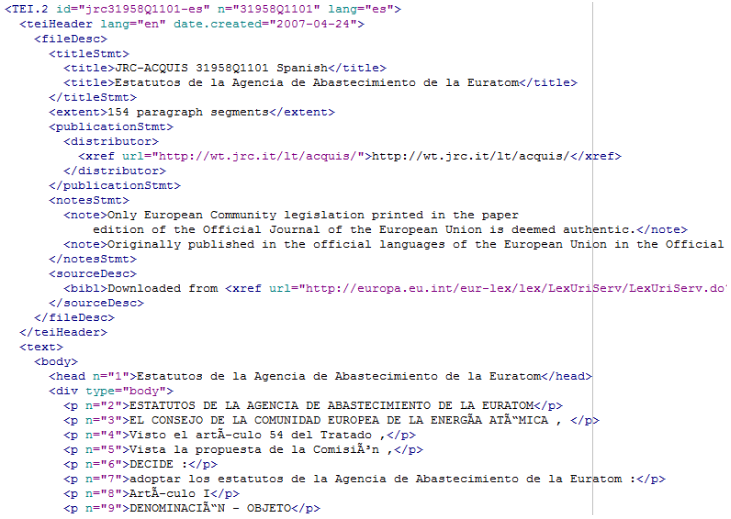


No Question Analysis module was used, just the PR one (JIRS)

•Introduction
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

- Document collection: JRC-Acquis collection of EU documentation, dealing with the related legislation, including written texts between the years 1950 to 2006 (in total 10,700 documents).

Example (original data set):

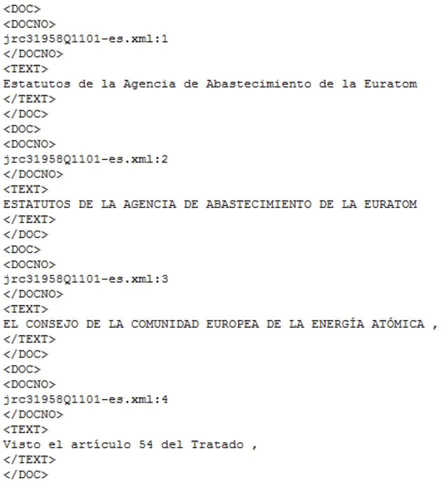


```
<TEI.2 id="jrc31958Q1101-es" n="31958Q1101" lang="es">
  <teiHeader lang="en" date.created="2007-04-24">
    <fileDesc>
      <titleStmt>
        <title>JRC-ACQUIS 31958Q1101 Spanish</title>
        <title>Estatutos de la Agencia de Abastecimiento de la Euratom</title>
      </titleStmt>
      <extent>154 paragraph segments</extent>
      <publicationStmt>
        <distributor>
          <xref url="http://wt.jrc.it/lt/acquis/">http://wt.jrc.it/lt/acquis/</xref>
        </distributor>
      </publicationStmt>
      <notesStmt>
        <note>Only European Community legislation printed in the paper
            edition of the Official Journal of the European Union is deemed authentic.</note>
        <note>Originally published in the official languages of the European Union in the Official
      </notesStmt>
      <sourceDesc>
        <bibl>Downloaded from <xref url="http://europa.eu.int/eur-lex/lex/LexUriServ/LexUriServ.do'
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
    <body>
      <head n="1">Estatutos de la Agencia de Abastecimiento de la Euratom</head>
      <div type="body">
        <p n="2">ESTATUTOS DE LA AGENCIA DE ABASTECIMIENTO DE LA EURATOM</p>
        <p n="3">EL CONSEJO DE LA COMUNIDAD EUROPEA DE LA ENERGÃA ATÃ“MICA , </p>
        <p n="4">Visto el artÃ-culo 54 del Tratado ,</p>
        <p n="5">Vista la propuesta de la ComisiÃ³n ,</p>
        <p n="6">DECIDE :</p>
        <p n="7">adoptar los estatutos de la Agencia de Abastecimiento de la Euratom :</p>
        <p n="8">ArtÃ-culo I</p>
        <p n="9">DENOMINACIÃ“N - OBJETO</p>
```

•Introduction
•Question Answering
•Passage Retrieval
•CLEF
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

- Pre-processing and analysis of documents: The collection of the competition is made of documents in XML format, each one divided into paragraphs delimited by the tag <p>. Therefore, each paragraph has been defined as a document, tagged with the name of the document where it is contained and the paragraph number that corresponds to it.

Example (after pre-processing):



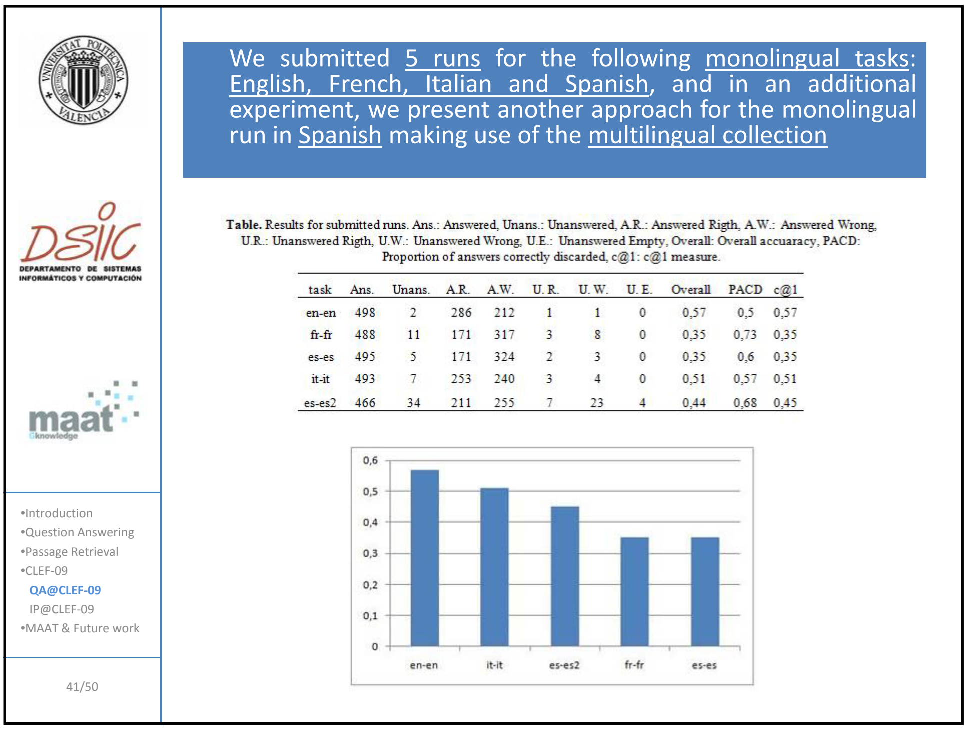
```
<DOC>
<DOCNO>
jrc31958Q1101-es.xml:1
</DOCNO>
<TEXT>
Estatutos de la Agencia de Abastecimiento de la Euratom
</TEXT>
</DOC>
<DOC>
<DOCNO>
jrc31958Q1101-es.xml:2
</DOCNO>
<TEXT>
ESTATUTOS DE LA AGENCIA DE ABASTECIMIENTO DE LA EURATOM
</TEXT>
</DOC>
<DOC>
<DOCNO>
jrc31958Q1101-es.xml:3
</DOCNO>
<TEXT>
EL CONSEJO DE LA COMUNIDAD EUROPEA DE LA ENERGÍA ATÓMICA ,
</TEXT>
</DOC>
<DOC>
<DOCNO>
jrc31958Q1101-es.xml:4
</DOCNO>
<TEXT>
Visto el artículo 54 del Tratado ,
</TEXT>
</DOC>
```

•Introduction
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

- Indexing: Once all the documents have been extracted from the collection, they have been indexed in JIRS according to the language that has been analysed

- Passage Retrieval: searched for the answer to each question of the track

•Introduction
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work



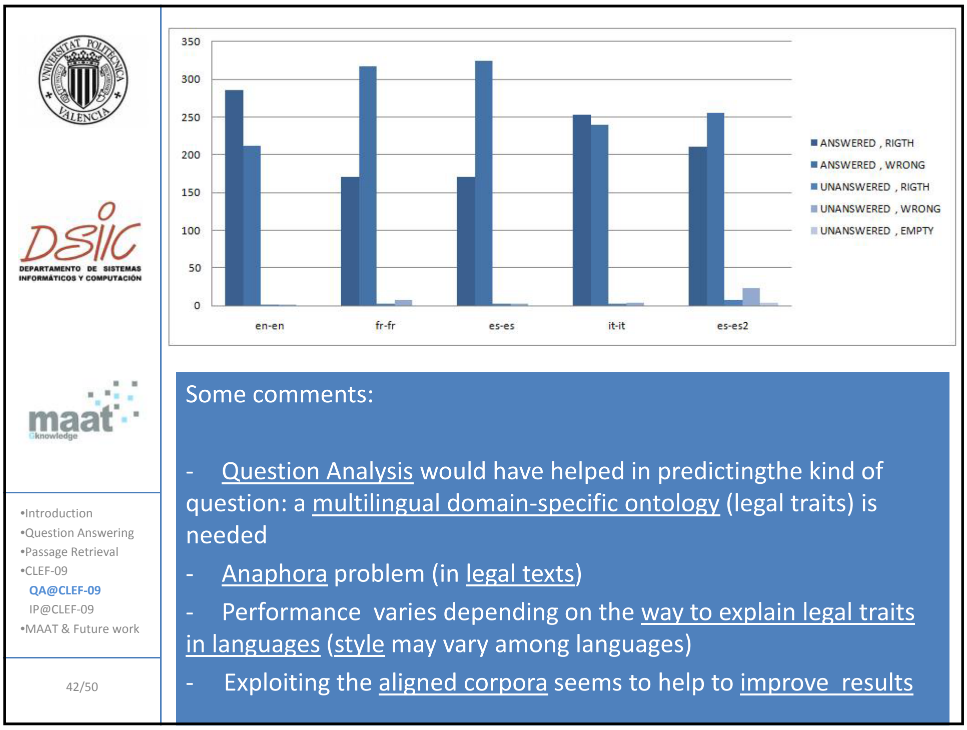
We submitted 5 runs for the following monolingual tasks: English, French, Italian and Spanish, and in an additional experiment, we present another approach for the monolingual run in Spanish making use of the multilingual collection

Table. Results for submitted runs. Ans.: Answered, Unans.: Unanswered, A.R.: Answered Rigth, A.W.: Answered Wrong, U.R.: Unanswered Rigth, U.W.: Unanswered Wrong, U.E.: Unanswered Empty, Overall: Overall accuaracy, PACD: Proportion of answers correctly discarded, c@1: c@1 measure.

task	Ans.	Unans.	A.R.	A.W.	U. R.	U. W.	U. E.	Overall	PACD	c@1
en-en	498	2	286	212	1	1	0	0,57	0,5	0,57
fr-fr	488	11	171	317	3	8	0	0,35	0,73	0,35
es-es	495	5	171	324	2	3	0	0,35	0,6	0,35
it-it	493	7	253	240	3	4	0	0,51	0,57	0,51
es-es2	466	34	211	255	7	23	4	0,44	0,68	0,45

•Introduction
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

41/50



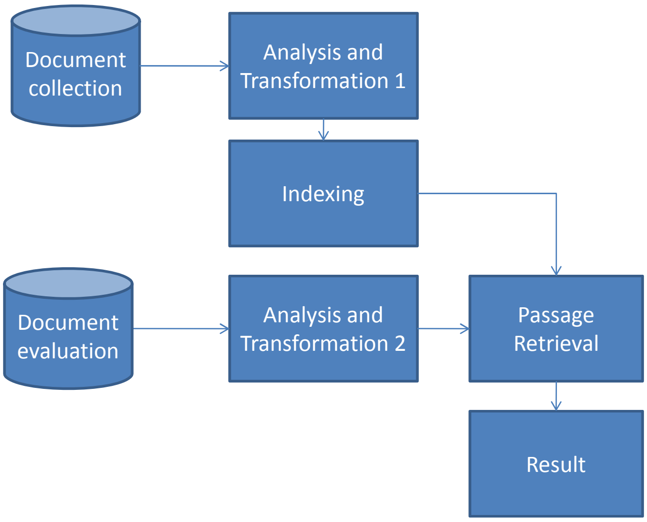
Some comments:

- Question Analysis would have helped in predictingthe kind of question: a multilingual domain-specific ontology (legal traits) is needed
- Anaphora problem (in legal texts)
- Performance varies depending on the way to explain legal traits in languages (style may vary among languages)
- Exploiting the aligned corpora seems to help to improve results

•Introduction
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

42/50

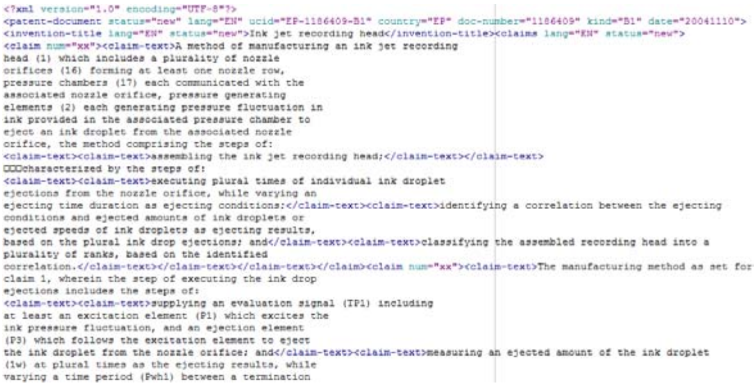
IP@CLEF-09: Approach



- •Introduction
- •Question Answering
- •Passage Retrieval
- •CLEF-09
  - QA@CLEF-09
  - **IP@CLEF-09**
- •MAAT & Future work

- Document collection by the European Patent Organization, a total of 1,958,955 patent documents relating to 1.022.388 patents.

- Example (original data set):

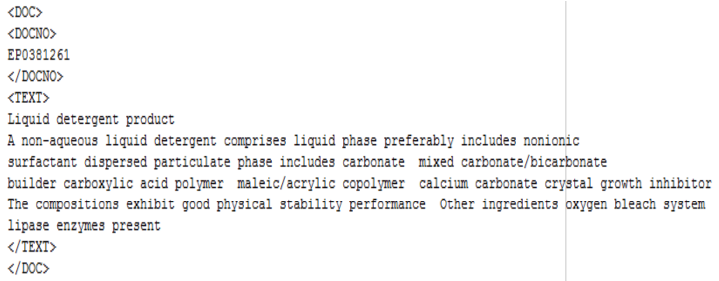


```
<?xml version="1.0" encoding="UTF-8"?>
<patent-document status="new" lang="EN" ucid="EP-1186409-B1" country="EP" doc-number="1186409" kind="B1" date="20041110">
<invention-title lang="EN" status="new">Ink jet recording head</invention-title><claims lang="EN" status="new">
<claim num="xx"><claim-text>A method of manufacturing an ink jet recording
head (1) which includes a plurality of nozzle
orifices (16) forming at least one nozzle row,
pressure chambers (17) each communicated with the
associated nozzle orifice, pressure generating
elements (2) each generating pressure fluctuation in
ink provided in the associated pressure chamber to
eject an ink droplet from the associated nozzle
orifice, the method comprising the steps of:
<claim-text><claim-text>assembling the ink jet recording head;</claim-text></claim-text>
□□□characterized by the steps of:
<claim-text><claim-text>executing plural times of individual ink droplet
ejections from the nozzle orifice, while varying an
ejecting time duration as ejecting conditions:</claim-text><claim-text>identifying a correlation between the ejecting
conditions and ejected amounts of ink droplets or
ejected speeds of ink droplets as ejecting results,
based on the plural ink drop ejections; and</claim-text><claim-text>classifying the assembled recording head into a
plurality of ranks, based on the identified
correlation.</claim-text></claim-text></claim-text></claim><claim num="xx"><claim-text>The manufacturing method as set for
claim 1, wherein the step of executing the ink drop
ejections includes the steps of:
<claim-text><claim-text>supplying an evaluation signal (TP1) including
at least an excitation element (P1) which excites the
ink pressure fluctuation, and an ejection element
(P3) which follows the excitation element to eject
the ink droplet from the nozzle orifice; and</claim-text><claim-text>measuring an ejected amount of the ink droplet
(1w) at plural times as the ejecting results, while
varying a time period (Pwh1) between a termination
```

- •Introduction
- •Question Answering
- •Passage Retrieval
- •CLEF-09
  - QA@CLEF-09
  - **IP@CLEF-09**
- •MAAT & Future work

- Pre-processing and analysis of documents (1): We decided to eliminate all the irrelevant information (just title and description)

Example (after pre-processing):



```
<DOC>
<DOCNO>
EP0381261
</DOCNO>
<TEXT>
Liquid detergent product
A non-aqueous liquid detergent comprises liquid phase preferably includes nonionic
surfactant dispersed particulate phase includes carbonate  mixed carbonate/bicarbonate
builder carboxylic acid polymer  maleic/acrylic copolymer  calcium carbonate crystal growth inhibitor
The compositions exhibit good physical stability performance  Other ingredients oxygen bleach system
lipase enzymes present
</TEXT>
</DOC>
```

- Indexing: Once all the documents have been extracted from the collection, they have been indexed with JIRS according to the language that has been analysed.

DEPARTAMENTO DE SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

•Introduction
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

45/50

- Document evaluation: A total of 500 patents

- Pre-processing and analysis of documents (2): the query is composed by the title of the patent followed by the most relevant n-grams composed by the heaviest terms , according to the weights assigned using the random walks method (method for summarisation)

Example (after preprocessing):

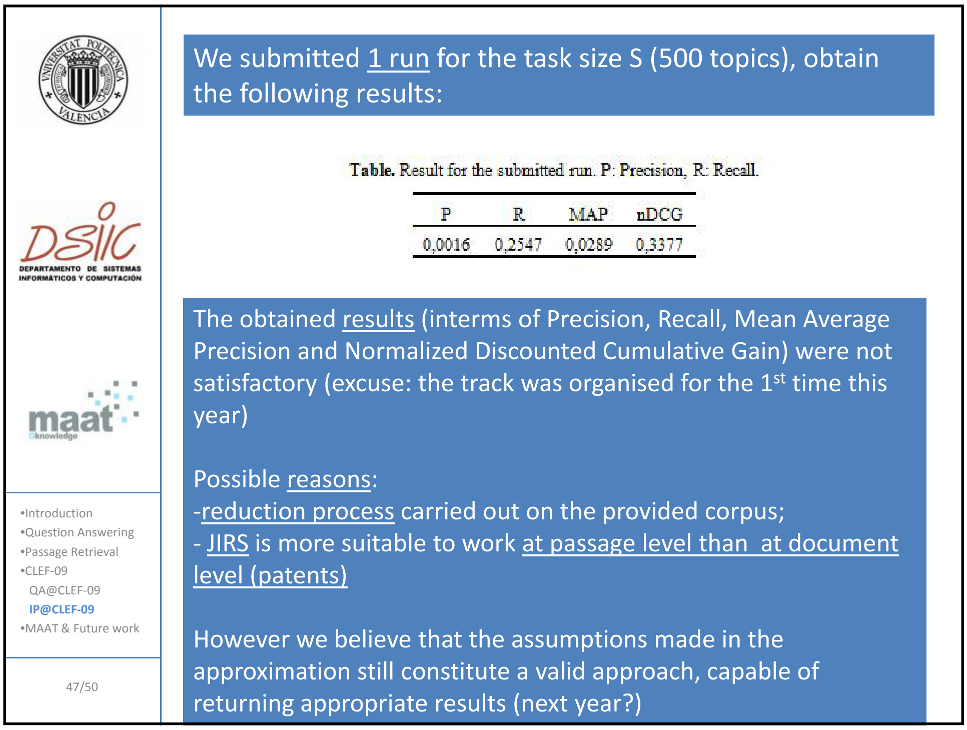
Topic: EP1445166
Name: Foldable baby carriage
Words extracted with random-walks: surface, seating
Question to JIRS: Foldable baby carriage, surface seating

- Passage Retrieval: searched for the rank of the patents in the data base relevamt to each "question" of the track

DEPARTAMENTO DE SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

•Introduction
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

46/50



We submitted 1 run for the task size S (500 topics), obtain the following results:

Table. Result for the submitted run. P: Precision, R: Recall.

P	R	MAP	nDCG
0,0016	0,2547	0,0289	0,3377

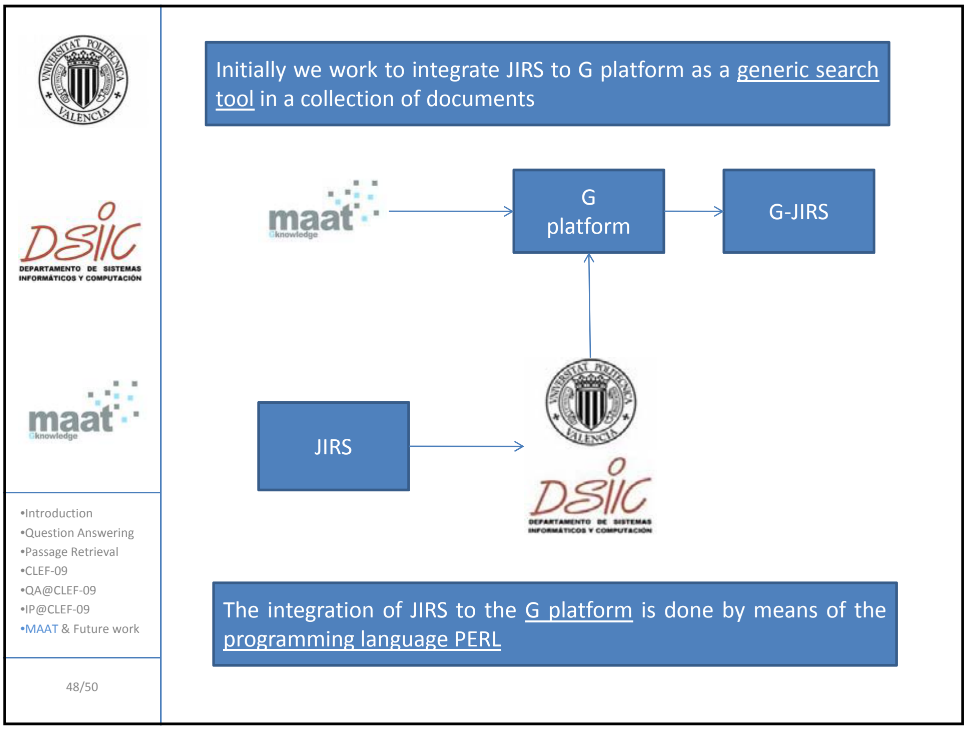
The obtained results (interms of Precision, Recall, Mean Average Precision and Normalized Discounted Cumulative Gain) were not satisfactory (excuse: the track was organised for the 1st time this year)

Possible reasons:
-reduction process carried out on the provided corpus;
- JIRS is more suitable to work at passage level than at document level (patents)

However we believe that the assumptions made in the approximation still constitute a valid approach, capable of returning appropriate results (next year?)

•Introduction
•Question Answering
•Passage Retrieval
•CLEF-09
QA@CLEF-09
IP@CLEF-09
•MAAT & Future work

47/50



Initially we work to integrate JIRS to G platform as a generic search tool in a collection of documents

The integration of JIRS to the G platform is done by means of the programming language PERL

•Introduction
•Question Answering
•Passage Retrieval
•CLEF-09
•QA@CLEF-09
•IP@CLEF-09
•MAAT & Future work

48/50

Future Work

Due to the growing interest in NLP-based approaches for the analysis of legal texts and patents (tacks of CLEF, TREC, NTCIR) from both the university (e.g. Technical University of Valencia) and the business sector (e.g. Maat Knowledge), we plan to employ JIRS in other commercial applications.

- •Introduction
- •Question Answering
- •Passage Retrieval
- •CLEF
- •QA@CLEF-09
- •IP@CLEF-09
- •MAAT & Future work

49/50

Thanks / Gracias / Grazie

Paolo Rosso
Natural Language Engineering Lab.
Universidad Politécnica Valencia, Spain

prosso@dsic.upv.es

http://users.dsic.upv.es/grupos/nle/

50/50!