Information Retrieval

Question Answering

Gintarė Grigonytė

gintare@ling.su.se
Department of Linguistics and Philology
Uppsala University
Slides based on previous IR course given by Jörg Tiedemann 2013-15

Information Retrieval

- Search relevant documents
- Given a query (usually some keywords)
- Return documents that contain the requested information
- Clustering & classification can be used to organize the collection

Now, something completely different: Question Answering!

"All right," said Deep Thought.
"The Answer to the Great Question..."
"Yes..!"
"Of Life, the Universe and Everything..." said Deep Thought.
"Yes...!"
"Is..." said Deep Thought, and paused.
"Yes...!"
"Is..."
"Yes...!!!...?"

Now, something completely different: Question Answering!

"All right," said Deep Thought.
"The Answer to the Great Question..."
"Yes..!"
"Of Life, the Universe and Everything..." said Deep Thought.
"Yes...!"
"Is..." said Deep Thought, and paused.
"Yes...!"
"Is..."
"Yes...!!!...?"

"Forty-two," said Deep Thought, with infinite majesty and calm."

Question Answering

What is the task?

- automatically find answers
- to a natural language question
- in pre-structured data (databases) and/or
- unstructured data (news texts, Wikipedia, ...)

Question Answering

Question Answering (QA) is answering a question posed in natural language and has to deal with a wide range of question types including: *fact, list, definition, how, why, hypothetical, semantically constrained, and cross-lingual questions*.

Question Answering

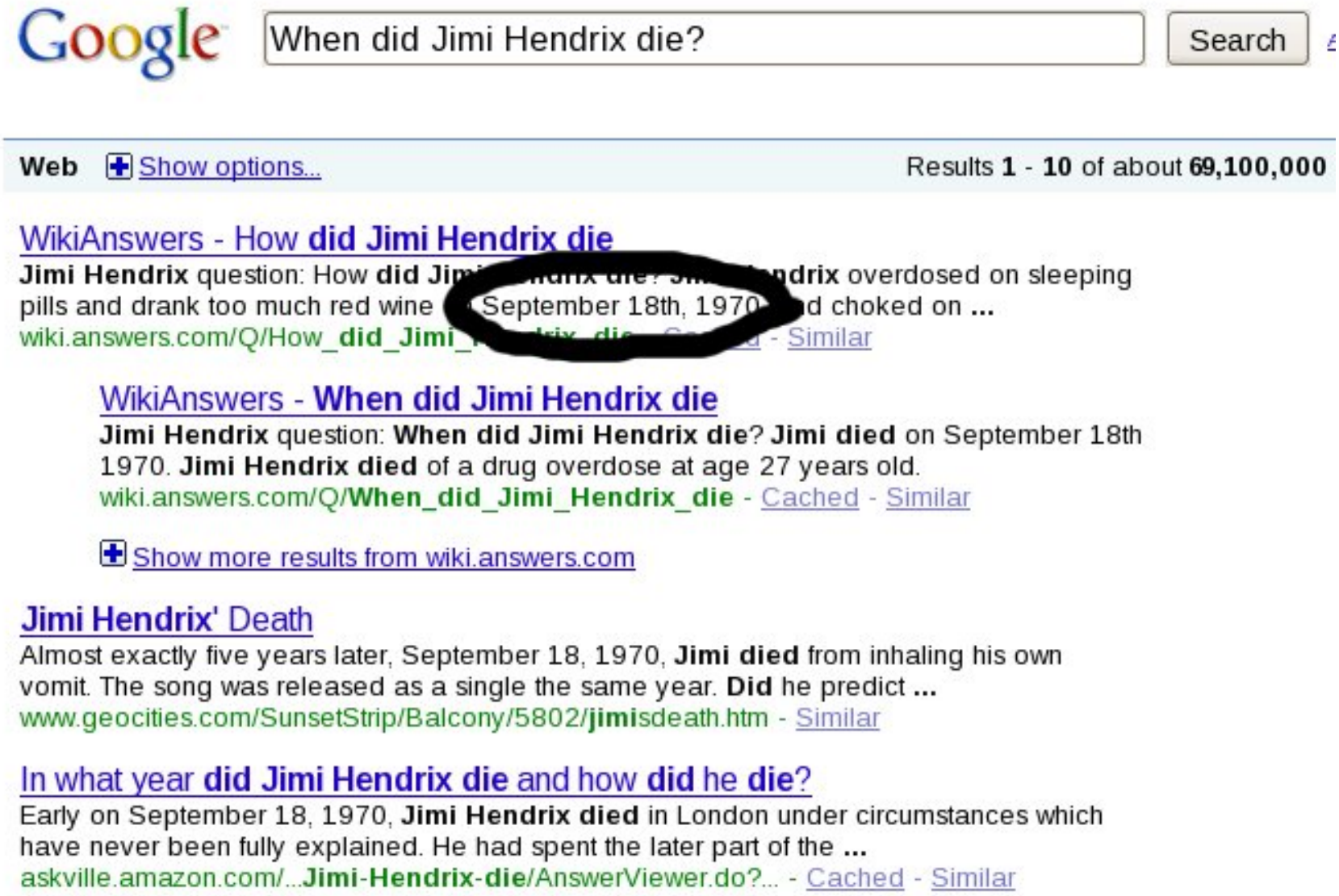
Question Answering (QA) is answering a question posed in natural language and has to deal with a wide range of question types including: *fact, list, definition, how, why, hypothetical, semantically constrained, and cross-lingual questions*.

Motivation:

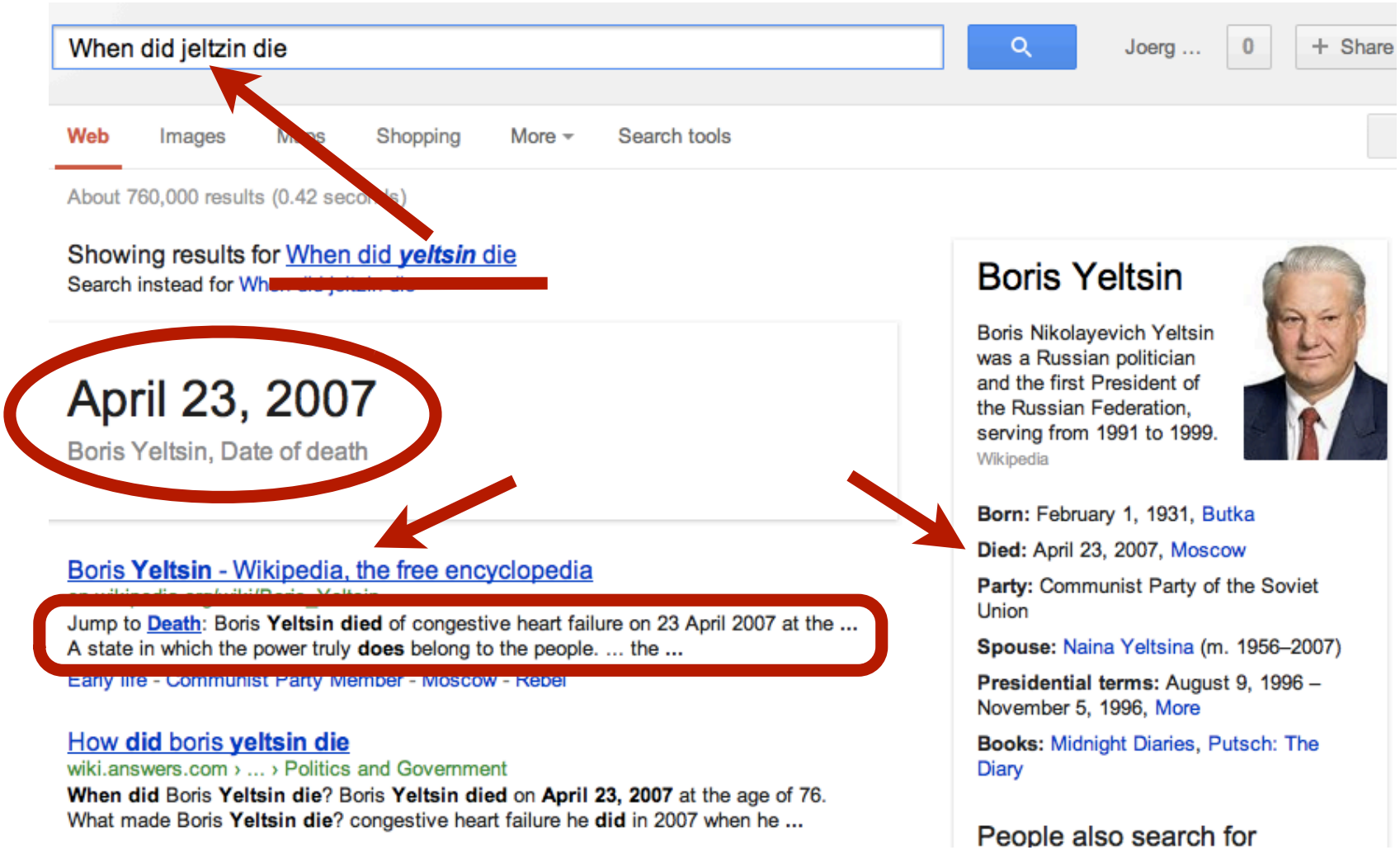
- retrieve specific information (general or domain-specific)
- do it in a more natural (human) way
- possibly integrate this in a dialogue system
  - follow-up questions (with coreferences)
  - interactive topic specific dialogues
  - multi-modal, speech recognition & synthesis

$\rightarrow$ Very ambitious!

Answering Questions with Text Snippets



When did Boris Yeltsin die?



Question Answering from Unstructured Data

- When was the Ebola virus first encountered?
  - `244 persons died of the Ebola virus, that was first found in Zaire in 1976`
- How did Jimi Hendrix die?
  - `...and when on September 18, 1970, Jimi Hendrix died of an overdose, her reaction...`
- What is the capital of Russia?
  - `The riders had a tour of Moscow this morning. Tomorrow morning they are leaving the Russian capital..`

Question Answering: Why is this difficult?

- When was the unification of Germany?
  - *Already in* **1961** *he predicted the unification of Germany.*

Question Answering: Why is this difficult?

- When was the unification of Germany?
  - *Already in* **1961** *he predicted the unification of Germany.*
- What is the capital of Ireland?
  - *It is a common joke in Ireland to call* **Cork** *the "real capital of Ireland"*
  - *In the middle ages* **Kilkenny** *was the capital of Ireland*

Question Answering: Why is this difficult?

- When was the unification of Germany?
  - *Already in* **1961** *he predicted the unification of Germany.*
- What is the capital of Ireland?
  - *It is a common joke in Ireland to call* **Cork** *the "real capital of Ireland"*
  - *In the middle ages* **Kilkenny** *was the capital of Ireland*
- What is RSI?
  - Website with "Common misconceptions about RSI" .... *RSI is the same as a mouse arm.*

Information Extraction

Information Extraction (IE) is extracting structured information from unstructured machine-readable documents by means of natural language processing (NLP).

Information Extraction

Information Extraction (IE) is extracting structured information from unstructured machine-readable documents by means of natural language processing (NLP).

Motivation:

- unstructured diverse data collections are full of information
- extract & store world knowledge from those collections
- structured collections (databases) for many puposes
  - searchable fact databases
  - question answering

$\rightarrow$ for example, turn Wikipedia into a well-structured fact-DB

Information Extraction

Sub-tasks:

- named entity recognition
- coreference resoultion
- terminology extraction
- relationship extraction

Find patterns like

- PERSON works for ORGANIZATION
- PERSON lives in LOCATION

Precision is usually more important than Recall!

Information Extraction

IE often requires heavy NLP and semantic inference

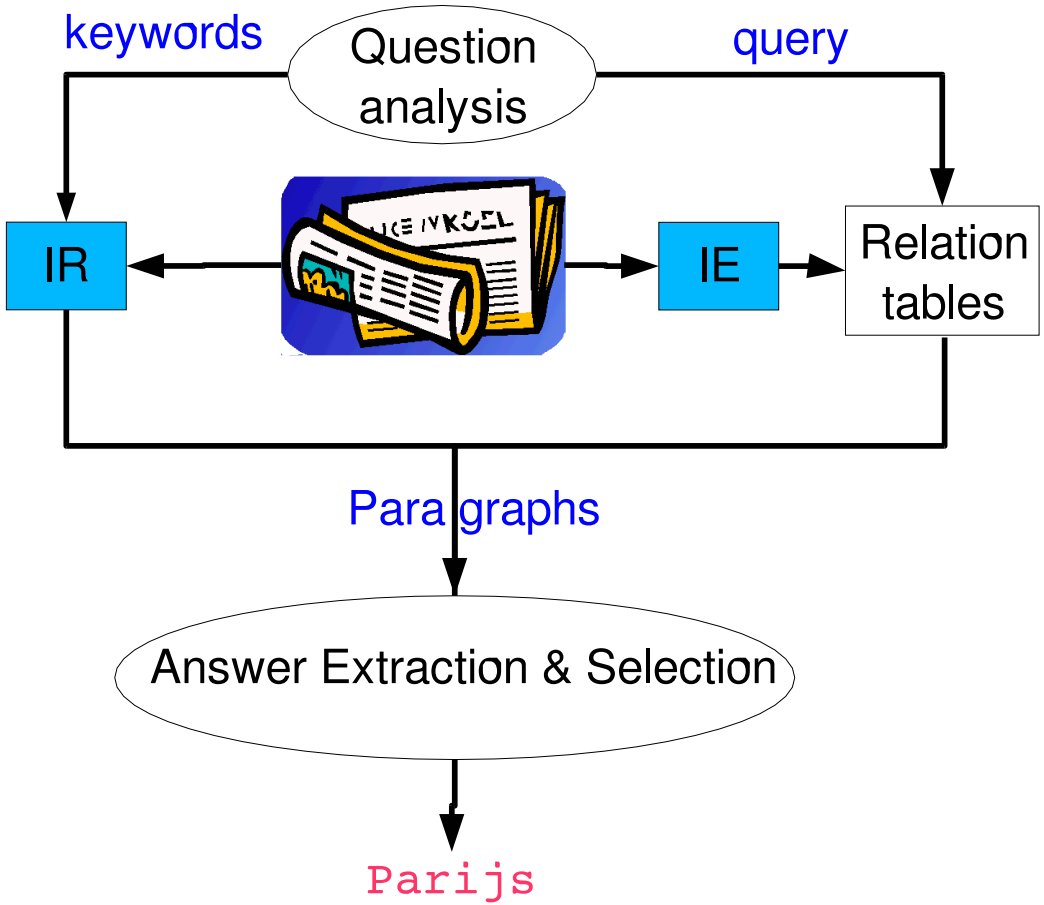
- extract dates of people's death
  - `...and when on September 18, 1970, Jimi Hendrix died of an overdose, her reaction...`
- extract capitals of countries in the world
  - `The riders had a tour of Moscow this morning. Tomorrow morning they are leaving the Russian capital..`

Simple cases exist (e.g. infoboxes at Wikipedia)

QA Systems: Do IE on-line with arbitrary questions

QA Systems: Do IE on-line with arbitrary questions

Wat is de hoofdstad van Frankrijk?



Challenges

- Natural Language Understanding
- Precision and Efficiency!
- Natural Interface, Dialog, Domain Flexibility

Evaluation in QA

- Evaluation with mean reciprocal ranks:

$$MRR_{QA} = \frac{1}{N}\sum_{1}^{N}\frac{1}{rank(\text{first\_correct\_answer})}$$

  (considering first 5 answers per question)
- accuracy: compute precision of first answer
- often: distinguish between
  - correct
  - inexact
  - unsupported (correct answer but not in its context)

Step 1: Question Analysis

What type of Question?

- question type $\rightarrow$ predict type of answer
- define question typology
- often quite fine-grained
  - location, date, capital, currency, founder, definition, born_date, abbreviation, ...
- even more fine-grained: types may take arguments
  - inhabitants(Sweden)
  - function(president, USA)

Step 1: Question Analysis

Patterns for question analysis

- could use machine learning
  - requires annotated training data
  - difficult to define features to be used
- hand-crafted patterns
  - very effective
  - easy to extend and to tune
  - easy to add new Q-types

Step 1: Question Analysis

Use syntactic information (e.g. dependency relations)

- When was the Rome Treaty signed?
- ⟨when, wh, Verb⟩, ⟨Verb, su, Event⟩
  → event_date(Rome Treaty)
- In which city did the G7 take place?
- ⟨in, obj, Geotype⟩, ⟨Geotype, det, which⟩,
  ⟨in, wh, Verb⟩, ⟨Verb, su, Event⟩
  → location(G7,city)

Next steps

- Given a question type,
- Search through relevant documents for sentences containing a phrase that is a potential answer.
  - keywords and keyphrases from question
  - question analysis $\rightarrow$ create appropriate query
- Rank potential answers & select top *n* answers

Step 2: Passage Retrieval

- Information Retrieval: Search relevant documents
- QA: Need the text snippet that contains an answer!

Step 2: Passage Retrieval

- Information Retrieval: Search relevant documents
- QA: Need the text snippet that contains an answer!

Passage retrieval is used as filtering component:

- answer extraction is expensive ($\rightarrow$ heavy NLP)
- narrow down search space
  $\rightarrow$ smaller segments (e.g. paragraphs)
  $\rightarrow$ only a few but relevant matches (including answer)

Step 2: Evaluating Passage Retrieval

Mean reciprocal ranks: The mean of the reciprocal rank of the first passage retrieved that contains a correct answer.

$$MRR_{IR} = \frac{1}{N}\sum_{1}^{N}\frac{1}{rank(\text{first\_relevant\_passage})}$$

Step 2: Evaluating Passage Retrieval

Mean reciprocal ranks: The mean of the reciprocal rank of the first passage retrieved that contains a correct answer.

$$MRR_{IR} = \frac{1}{N}\sum_{1}^{N}\frac{1}{rank(\text{first\_relevant\_passage})}$$

Coverage: Percentage of questions for which at least one passage is retrieved that contains a correct answer.

Redundancy: The average number of passages retrieved per question that contain a correct answer

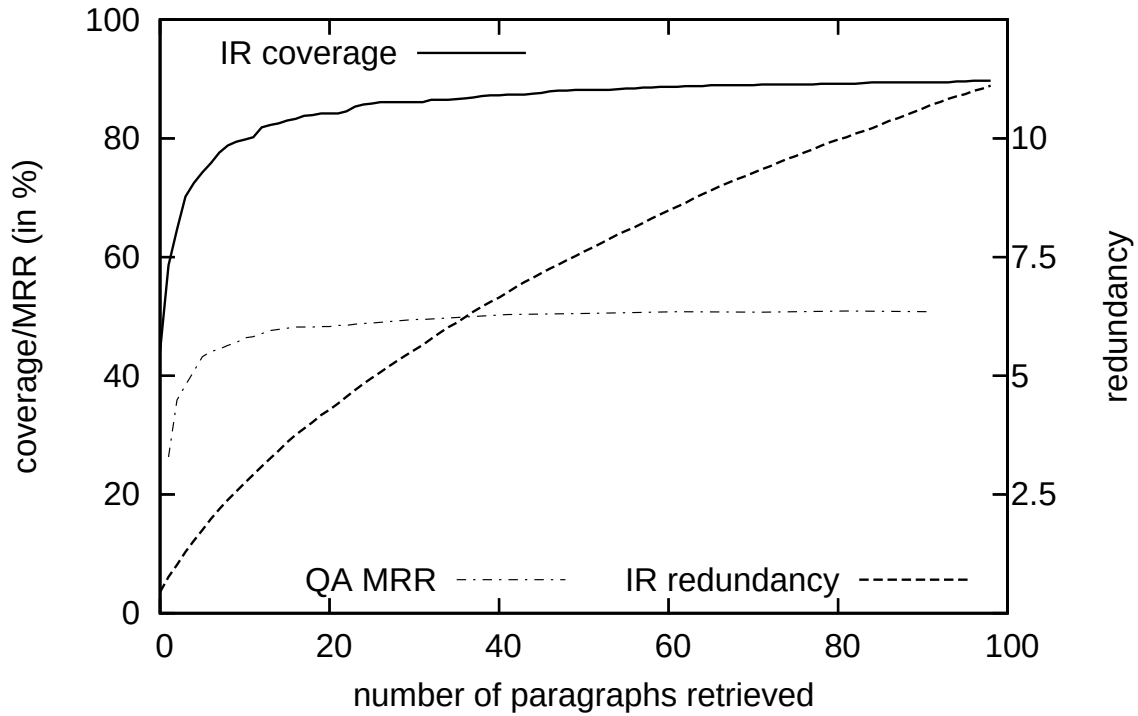
Evaluation Example: Retrieve 3 Paragraphs

Question: Wie is de leider van Bosnië ?
Accepted Answers: Ratko Mladic, Radovan Karadzic

rank 1, IR-score: 12.754168
Kroatië voldoet aan het verzoek van Bosnië om steun in de noordwestelijke enclave Bihac , die dreigt te worden ingenomen door Serviërs uit Bosnië en Kroatië , bijgestaan door afvallige milities van moslim-leider Abdic .
rank 2, IR-score: 12.581567
De leiders van de Bosnische en de Kroatische Serviërs deden gisteren een beroep op de Servische regering te hulp te komen bij het pareren van het Kroatische offensief in West-Bosnië en de Krajina ." De Serviërs in Bosnië en Kroatië vechten met gebonden handen en hebben dringend steun nodig van Servië en Joegoslavië " , aldus de Bosnisch-Servische leider Radovan Karadzic en de Kroatisch-Servische leider Milan Martic .
rank 3, IR-score: 12.418680
In een vraaggesprek met de Griekse krant Ta Nea heeft de Bosnisch-Servische leider Radovan Karadzic gezegd dat hij geen nieuw ultimatum van de Navo zal accepteren . Karadzic wil spoedig een bezoek brengen aan Griekenland om Athene te vragen de kwestie Bosnië op de agenda van de Europese Unie te plaatsen . De Servische leider zei niet gelukkig te zijn met de Amerikaanse bemoeienis met Bosnië omdat de VS de zaak uit handen dreigt te halen van de VN .

- coverage: 1
- redundancy: 2
- MRR_{IR}: 1/2 = 0.5

Coverage & Redundancy: Experiments with Joost



→ strong correlation between coverage and QA MRR

→ coverage is more important than redundancy

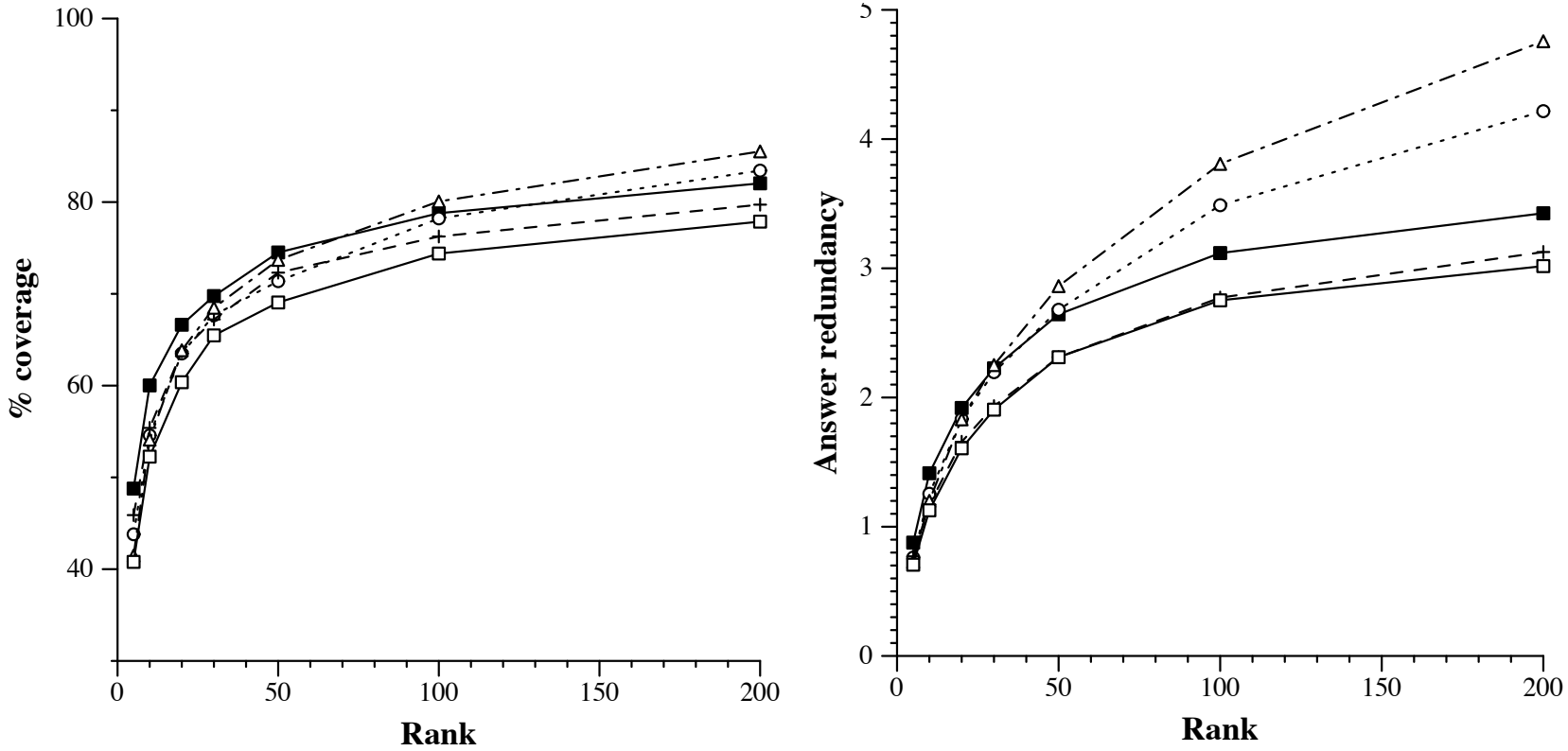
Step 2: Two Types of Passage retrieval

- retrieve relevant passages for given query
- can use IR techniques
  - index passages instead of documents
  - rank passages according to standard tf-idf or similar
    $\rightarrow$ index-time passaging

Step 2: Two Types of Passage retrieval

- retrieve relevant passages for given query
- can use IR techniques
  - index passages instead of documents
  - rank passages according to standard tf-idf or similar
    → index-time passaging
- two-step procedure:
  1. standard document retrieval
  2. segmentation + passage selection
     → seach-time passaging

Index-time versus Search-time Passaging



dotted line = index-time passaging ((Roberts & Gaizauskas, 2004))

$\rightarrow$ not much to gain with two-step procedure

Step 2: Passage Retrieval: What is a Passage?

How much should we retrieve?

Step 2: Passage Retrieval: What is a Passage?

How much should we retrieve?

	#sent	cov	red	*MRR* IR	*MRR* QA	accuracy CLEF
sentences	16,737	0.784	2.95	0.490	**0.487**	**0.430**
paragraphs	80,046	0.842	4.17	0.565	**0.483**	**0.416**
documents	618,865	0.877	6.13	0.666	0.457	0.387

(CLEF Experiments with Joost, retrieve 20 units/question)

→ retrieving smaller units better than document retrieval

Step 2: Passage Retrieval: Text Segmentation

Different document segmentation approaches:

	# sentences	MRR_{IR}	MRR_{QA}
sentences	16,737	0.490	0.487
paragraphs	80,046	0.565	0.483
TextTiling	107,879	0.586	△ **0.503**
2 sentences	33468	0.545	△ **0.506**
3 sentences	50190	0.554	0.504
4 sentences	66800	0.581	△ **0.512**
2 sentences (sliding)	29095	0.548	△ **0.516**
3 sentences (sliding)	36415	0.549	0.484
4 sentences (sliding)	41565	0.546	0.476

Step 2: Improved Passage Retrieval

Augment passage retrieval with additional information from question analysis and from linguistic annotation:

Step 2: Improved Passage Retrieval

Augment passage retrieval with additional information from question analysis and from linguistic annotation:

- include index of Named Entity labels
  $\rightarrow$ match question type

Step 2: Improved Passage Retrieval

Augment passage retrieval with additional information from question analysis and from linguistic annotation:

- include index of Named Entity labels
  $\rightarrow$ match question type
- use syntactic information to include phrase queries
  $\rightarrow$ match dependency relation triples

Step 2: Improved Passage Retrieval

Augment passage retrieval with additional information from question analysis and from linguistic annotation:

- include index of Named Entity labels
  $\rightarrow$ match question type
- use syntactic information to include phrase queries
  $\rightarrow$ match dependency relation triples
- use weights to boost certain keyword types
  $\rightarrow$ named entities, nouns, entities in specific relations
  $\rightarrow$ include linguistic annotation in zone/tiered index

Step 2: Improved Passage Retrieval

Augment passage retrieval with additional information from question analysis and from linguistic annotation:

- include index of Named Entity labels
  $\rightarrow$ match question type
- use syntactic information to include phrase queries
  $\rightarrow$ match dependency relation triples
- use weights to boost certain keyword types
  $\rightarrow$ named entities, nouns, entities in specific relations
  $\rightarrow$ include linguistic annotation in zone/tiered index
- query expansion $\rightarrow$ increase recall

Step 2: Passage Retrieval: Summary

- use IR techniques, but
- ... focus on smaller units (sentences or passages)
- ... focus on coverage (but also ranking)
- ... include extra-information (q-type, annotation)

- retrieve less
  $\rightarrow$ lower risk for wrong answer exraction
  $\rightarrow$ increased efficiency

Step 3: Answer Extraction & Ranking

Now we have

- expected answer type ($\rightarrow$ question type)
- relevant passages
- retrieval scores (ranking)
- (syntactically) analysed question

$\rightarrow$ Need to extract answer candidates
$\rightarrow$ Rank candidates and return answers

Step 3a: Answer Extraction

Can use syntactic information again:

- Where did the meeting of the G7-countries take place?
  - `location(meeting,nil)`
  - `.. after a three-day meeting of the G7-countries in Napels.`
- in Napels is a potential answer for `location(meeting,nil)`
  - if NE class = LOC
  - if Answer syntactically related to Event
  - if modifiers of Event in Q and A overlap

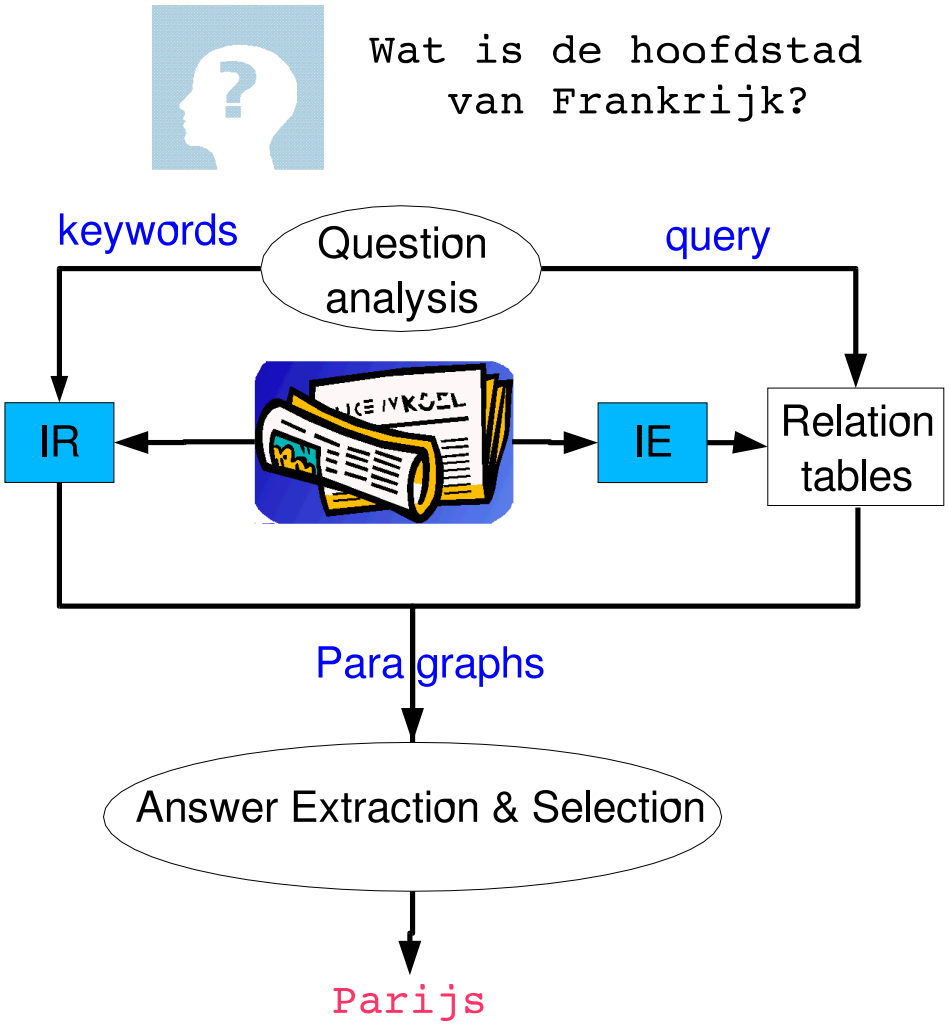
Step 3b: Answer Ranking

Combine various knowledge sources: for example

- Final score of an answer is weighted sum of
  - TypeScore
  - Syntactic Similarity of Q and A sentence
  - Overlap in Names, Nouns, Adjectives between Q and A sentence and preceding sentence
  - IR score
  - Frequency of A

$\rightarrow$ tune combination weights (experts eller machine learning)

Step 2b: Match fact databases



Question Answering vs. IR

Information Retrieval

- **input:** keywords (+ boolean operators, ...)
- **output:** links to relevant documents (maybe snippets)
- **techniques:** vector-space model, bag-of-words, tf-idf

Question Answering

- **input:** natural language question
- **output:** concrete answer (facts or text)
- **techniques:** shallow/deep NLP, passage retrieval (IR), information extraction

$\rightarrow$ IR is just a component of question answering!

Summary of Terminology

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

Information Extraction (IE) is extracting structured information from unstructured machine-readable documents by means of natural language processing (NLP).

Question Answering (QA) is answering a question posed in natural language and has to deal with a wide range of question types including: *fact, list, definition, how, why, hypothetical, semantically constrained, and cross-lingual questions*.