Information Retrieval in Natural Language Texts[1]

Agata Savary

February 11, 2008

[1] Master in Information Systems and Decision Support, Faculty of Computer Science, Blois

Textual Information Retrieval

Information Retrieval (IR) = extracting relevant information from a set of documents

Textual IR = when the documents or their parts are expressed in a natural language

Most visible example of a textual IR: Web browsers

Examples of Applications in Textual IR

Intelligent Web browsers: "understand" the query and find "relevant" pages

Text summarization: produce a resume of a text

Question answering: ask a "wh-query" (*what/who/when/where*) in a natural language, find sentences containing the answer

Text classification: group together texts talking about the same subject

Automatic indexing: find terms in a text that best characterize its contents

Term acquisition: find the technical terminology appearing in a set of texts

Multilingual information extraction: ask a query in one language, get information in different languages

Retrieval techniques

Bag-of-words:

- a query is seen as a bag of words with no relations between them
- each word is searched for in a document, the weight of a document may or may not take co-occurrences of words into account

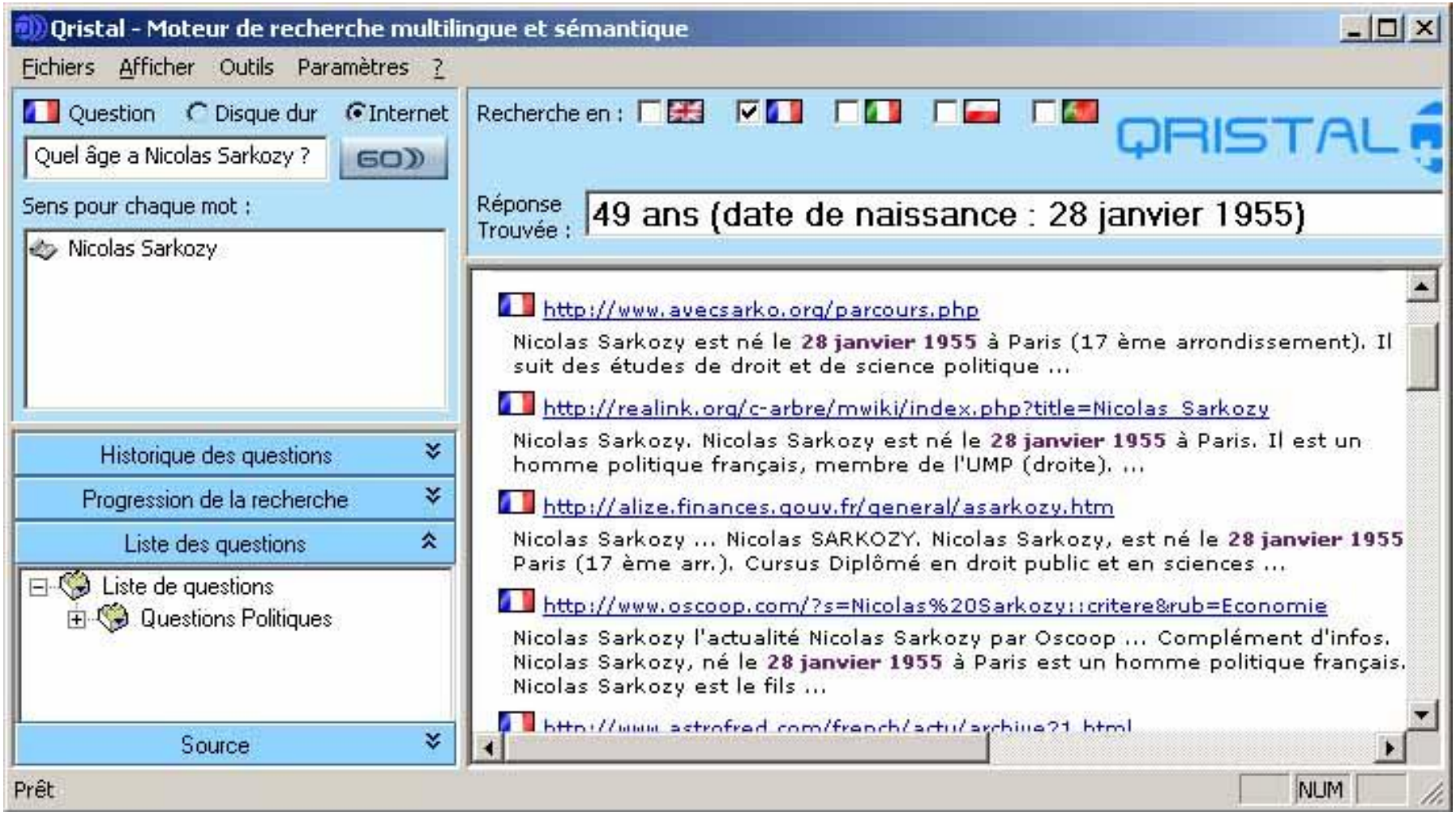
Linguistically motivated tools:

- a query is seen as a phrase where each word has its morphological and syntactic properties
- the linguistic properties of (simple or/and compound) words are taken into account in document search

Linguistic levels of word analysis

	Query word	Relevant text occurrence	Irrelevant text occurrence
spelling	flavour	flavor	favour
morphology	apply	applies	apples
syntax	birth date	date of birth date and place of birth	date cards and birth cards
semantics	automobile cleaning	car washing	machine washing

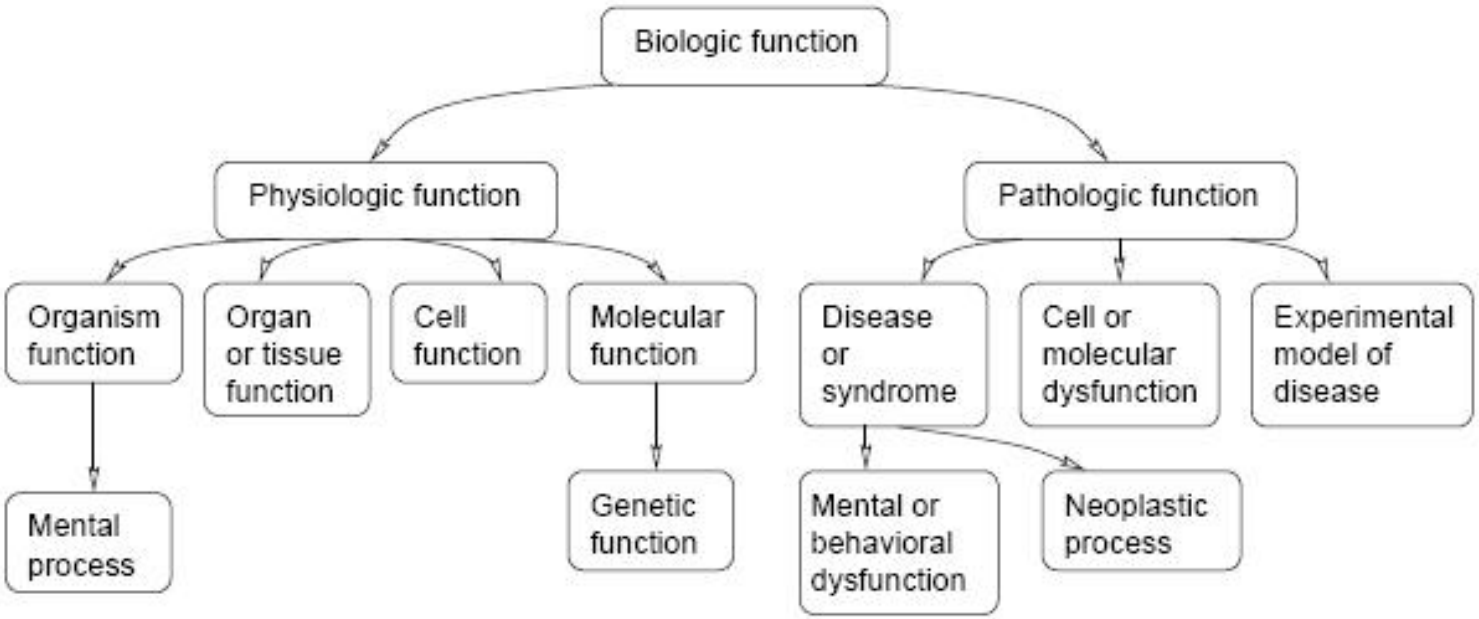
Application Example: Question Answering System



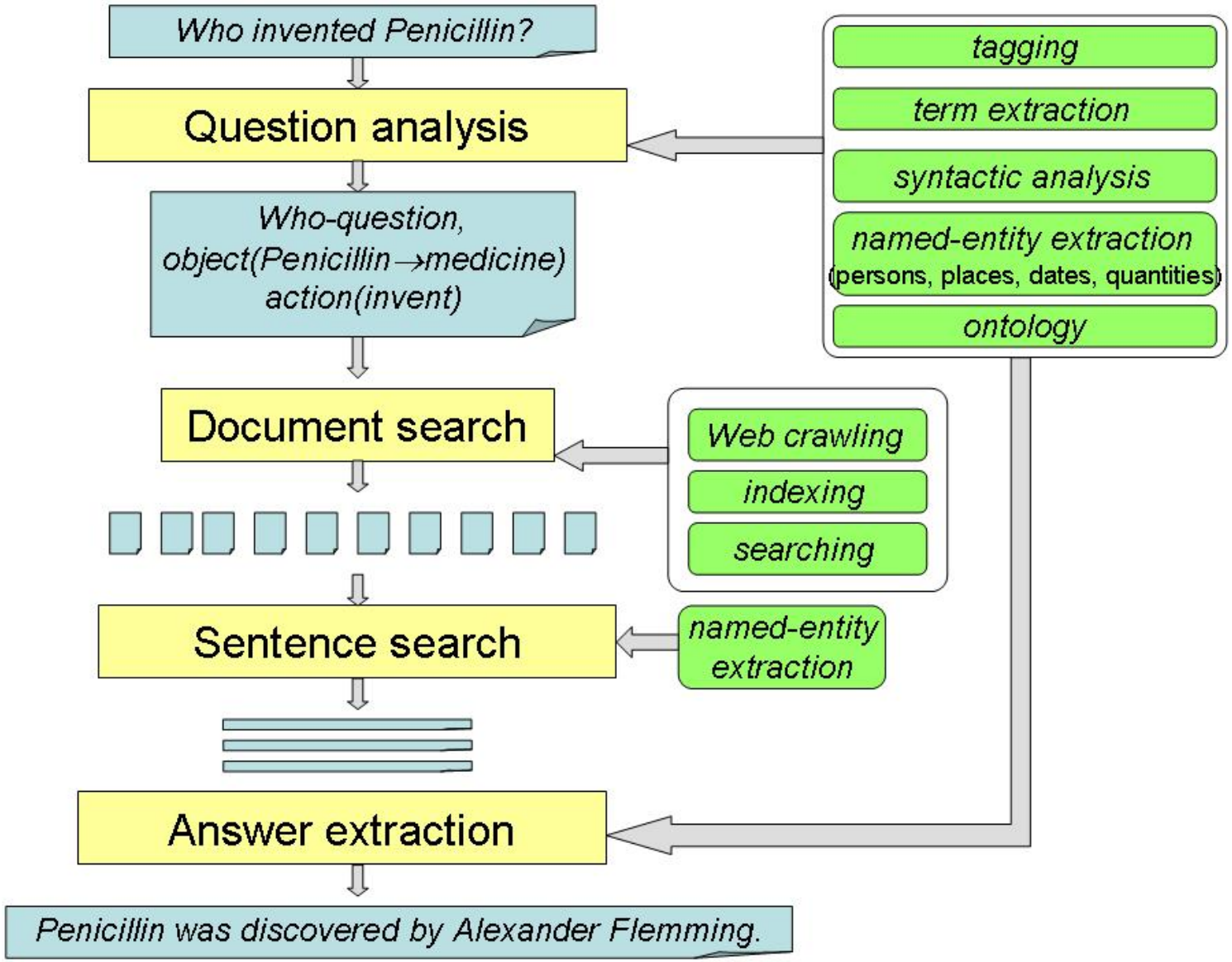
Question Answering System

Open-domain QA: only a general knowledge can be used (e.g. *Wordnet*)

Closed-domain QA: a domain-specific knowledge can be exploited, e.g. a medical ontology:



Question Answering: Possible Architecture



Future Question Answering Systems

Aggregation of results

Resolution of contradictions

Generation of a synthetic answer

Future QA Systems

Question: *When did the Secession War take place?*

Extracted answers:

- **From April 17 to May 21, 1861**, the states of Arkansas, North Carolina, Tenessee and ... It's the **beginning of the Secession War**...
- **April 12, 1861-April 9, 1865 - War of Secession**. Capitulation of Appamatox September 22, 1862...
- **May 26, 1865, end of the Secession War**. General Lee has capitulated on April the 9th, and Johnson on the 26th...
- The **Secession War has started on April 12, 1861** with the Confederate forces attacking Fort Sumter...
- the South declares **secession on July 21, 1861** - the beginning of the **War**, July 1st-3rd 1963: Gettysburg battle...

Desired (generated) answer:

- *The Secession War took place from April 12, 1861 to April 9 1865, or, less probably, from April 17, 1861 to April 9, 1865.*