

Question Answering and Multi-Dimensional Markup

Retrieving Content and Structure

Jaap Kamps, Maarten Marx, Maarten de Rijke

ISLA

University of Amsterdam, The Netherlands

ESLLI, August 2005

<http://ilps.science.uva.nl>

A bit on NLP

- 'Natural language processing'
 - Computational modeling and processing of human language, often toward understanding the language
- Main challenge: ambiguity
 - 'Visiting aunts can be a nuisance'
 - 'She boarded the airplane with two suitcases' vs 'She boarded the airplane with two engines'
- Major levels of linguistic analysis (for our purposes)
 - Phonology: the sound patterns of a language
 - Morphology: the structure of words
 - Syntax: how words combine into larger units
 - Semantics: the meaning of words and sentences

Tasks, tasks, tasks

- Nr 1 online app of NLP: document retrieval
 - Increasing sophistication in indexing, identification and presentation of relevant text
- Tagging
 - Assigning labels or types to words or word groups
- Document classification
 - Assigning docs to classes, based on content
- Information extraction
 - Specific info targets in a doc/set of docs

Noun phrases and name recognizers

- Go beyond POS tagging
- Noun ('George') → name ('George Bush')
- Noun phrase parsers
 - 'Shallow' or 'partial' parsers
 - Only identify certain constituents
 - Collocations
- More generally, named entity recognizers identify and classify proper names in docs
 - **Italy's** business world was rocked by the announcement **last Thursday** that **Mr. Verdi** would leave his job as vice-president of **Music Masters of Milan, Inc** to become director of **Arthur Anders**
- Some NERC tools use hand-crafted rules, others learn rules from training data or build statistical models such as HMMs

Today's menu

- **A bit on natural language processing**
- Question answering
 - Background
 - Anatomy of a question
 - History
 - The canonical architecture
 - QA at TREC
 - Back to retrieving content and structure
 - Wrapping up

NLP and linguistics

- Major levels of linguistic analysis (for our purposes)
 - Phonology: the sound patterns of a language
 - Morphology: the structure of words
 - Syntax: how words combine into larger units
 - Semantics: the meaning of words and sentences
- Two views on NLP
 - Symbolic
 - Rules for manipulating symbols
 - Empirical
 - Derive language data from large text corpora
 - Distinction
 - Methodological: top-down vs bottom-up
 - Handling ambiguity: propose additional rules vs associate probabilities with alternative analysis
 - NLP rarely pure, never simple

Linguistic tools

- Layers of linguistic analysis
 - Doc → paragraph → sentence → word
 - Words are **tagged**, prior to sentence being **parsed**
 - Not all apps need all layers
- Sentence delimiters, tokenizers
 - Sentence boundary detection not trivial
 - Tokenization (word segmentation) based on white spaces?
 - data-base, 'pomme de terre', no white spaces between words in Chinese, compound formation in Dutch
- Stemmers, taggers
 - Stemmers: associate variant of same term with root form
 - Porter stemmer: rules for removing suffix -ed, -ing, -ation, -ational
 - Part of speech tagger labels words with POS (noun, verb, adjective, etc)
 - Visiting/ADJ aunts/N-Pl can /AUX be/V-inf-be a/DET-Indef nuisance/N-Sg
 - Visiting/V-prog aunts/N-Pl can /AUX be/V-inf-be a/DET-Indef nuisance/N-Sg
 - Rule-based taggers (Brill), stochastic taggers (TrT)

Parsers and grammars

- Parsing done wrt a grammar: set of rules saying which combinations of which POS generate well-formed phrase and sentence structures
- Linguistic engineering by writing grammars is very labor-intensive
- Induce grammars from annotated corpora
 - Penn Treebank project at University of Pennsylvania
- Syntactic structure mostly annotated using brackets to produce embedded lists
 - (S: (NP: Green ideas) (VP: sleep furiously))
- Trees and embedded lists interchangeable

Natural language understanding is a much bigger field

- Semantic interpretation
- Knowledge representation
 - Logic, frames, ...
- Inference
- Discourse structure
- Natural language generation

What's next?

- A bit on natural language processing
- Question answering
 - **Background**
 - Anatomy of a question
 - History
 - The canonical architecture
 - QA at TREC
 - Back to retrieving content and structure
 - Wrapping up

NLP and question answering

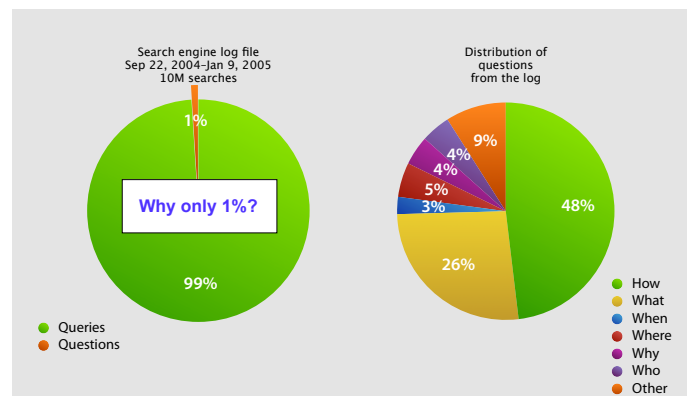
- IR typically retrieves or works with documents
 - Find *documents* that are relevant
 - Group *documents* on the same topic
- People often want a sentence fragment or phrase as the answer to their question
 - *Who was the first man to set foot on the moon?*
 - *What is the moon made of?*
 - *How many members are in the U.S. Congress?*
- Move IR from document retrieval to answer retrieval
 - Document retrieval is still valuable
 - Extends breadth of active IR research

Types of (web) information needs

- Navigational: try to reach a particular site
 - compaq
 - Probable target <http://www.compaq.com>.
 - Don Knuth
 - Probable target <http://www-cs-faculty.stanford.edu/~knuth/>
- Informational: acquire some information assumed to be present on one or more web pages
 - San Francisco
 - Zipf Law
- Informational needs can be broad or focused
 - Focused Informational need: user has a specific question in mind "where is the oldest windmill in holland?" → windmill netherlands oldest location

What if we could actually ask questions?

People do ask questions...



... what do they ask?

FACT(oid)S

- how tall is christina aquilera
- what does mig welding stand for
- where does sean hannity live?
- how long does ritalin stay in your bloodstream?
- where does moss grow

PROCEDURES

- how to repair scratches in leather
- how do you transfer money from one bank to another?
- how to speed up xp
- how to cook a sweet potatoe
- how to report a fraudulent bankruptcy claim
- jet engine how it works

DEFINITION(oid)S

- what is pink noise
- what is a catapult
- what is a pictograph
- define social justice
- what is a rational number?

...

- how to understand women
- how to stop your dog from pooping on your stairs
- almost everyone sees me without noticing me, for what is beyond is what he or she seeks. what am i?
- how smart are blonds

What's next?

- A bit on natural language processing
- Question answering
 - Background
 - **Anatomy of a question**
 - History
 - The canonical architecture
 - QA at TREC
 - Back to retrieving content and structure
 - Wrapping up

Anatomy of a question

- Question type
- Answer type
- Question focus
- Question topic

Question type

- Idiomatic categorization of questions for purposes of distinguishing between different processing strategies and/or answer formats
- TREC 2003
 - FACTOID: "How far is it from Earth to Mars?"
 - LIST: "List names of chewing gums"
 - DEFINITION: "Who is Vlad the Impaler?"
- Other possibilities:
 - RELATIONSHIP: "What is the connection between Valentina Tereshkova and Sally Ride?"
 - SUPERLATIVE: "What is the largest city on Earth?"
 - YES-NO: "Is Bin Laden alive?"
 - OPINION: "What do most Americans think of gun control?"
 - CAUSE&EFFECT: "Why did Iraq invade Kuwait?"
 - ...

Answer type

- The class of object (or rhetorical type of sentence) sought by the question:
 - PERSON (from "Who ...")
 - PLACE (from "Where ...")
 - DATE (from "When ...")
 - NUMBER (from "How many ...")
 - ...
- but also
 - EXPLANATION (from "Why ...")
 - METHOD (from "How ...")
 - ...

Question focus and topic

- **Question focus:** The property or entity that is being sought by the question.
 - Examples:
 - "In what **state** is the Grand Canyon?"
 - "What is the **population** of Bulgaria?"
 - "What **color** is a pomegranate?"
- **Question topic:** The object (person, place, ...) or event that the question is about
- The question might be about a property of the topic, which will be the question focus
 - "What is the **height** of **Mt. Everest**?"
 - **height** is the focus
 - **Mt. Everest** is the topic

What's next?

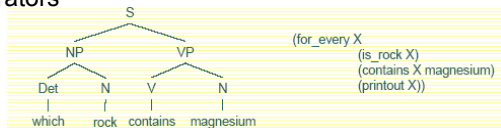
- A bit on natural language processing
- Question answering
 - Background
 - Anatomy of a question
 - **History**
 - The canonical architecture
 - QA at TREC
 - Back to retrieving content and structure
 - Wrapping up

Historical QA approaches

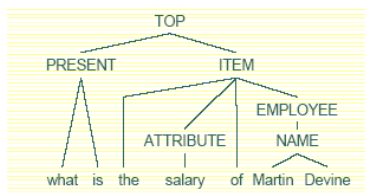
- Front-end to expert systems, databases (1970s–)
- Encyclopedias (1990s–)
- The Web (2000–)
- QA from databases—examples
 - BASEBALL – baseball statistics
 - Who did the Red Sox lose to on July 5?
 - On how many days in July did eight teams play?
 - LUNAR – analysis of lunar rocks
 - What is the average concentration of aluminum in high alkali rocks?
 - How many Brescias contain Olivine?

QA from databases – approaches

- Mapping rules between sentence structure and relational operators



"Question grammar"



QA from databases – cont.

- Works, with the following constraints:
 - Domain is narrow
 - Users have knowledge of it
 - Lots of manual work
- Will this work as a general QA approach?
 - Open domain natural language understanding researched for >30 years, still with little success

QA from encyclopedias

- MURAX (1993)
 - First open-domain QA system
 - Corpus is free-text
- Linguistic methods for
 - Question analysis
 - Answer extraction

QA in the web era

- (Almost) unlimited corpus of free text
- Caused systems to move to IR-based approaches
 - AskJeeves: map incoming questions to "known questions"
 - Not really a QA system
 - TREC
 - Evaluate QA in an open-domain, large corpus environment
 - Reason for majority of QA work today
 - More later

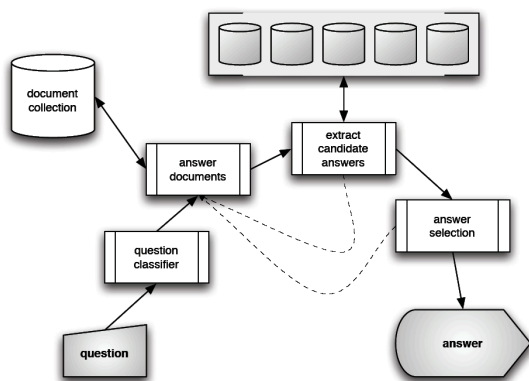
What's next?

- A bit on natural language processing
- Question answering
 - Background
 - Anatomy of a question
 - History
 - **The canonical architecture**
 - QA at TREC
 - Back to retrieving content and structure
 - Wrapping up

The canonical architecture

- Basic algorithm
 - Create query from question
 - Search a collection
 - Extract answers from retrieved documents
- Big challenge: lexical gap
 - "Where did the first atom bomb explode?"
 - "On August 6 the first nuclear bomb fell on Hiroshima"
 - **What can be done?**
- Lexical gap issue in all language processing tasks
 - Omnipresent in QA
 - Is this where NLP helps?

Typical QA system architecture



Question classification

- The process of extracting the answer type from the question
 - Additional possible by-products: topic, focus
- Used for
 - Query formulation
 - Answer extraction

How to classify?

- Regular expressions / patterns
 - "where ...": **location**
- But:
 - What tourist attractions are there in Reims?
 - What are the names of tourist attractions in Reims?
 - What do most tourists visit in Reims?
 - What attracts most tourists in Reims?
 - What is worth seeing in Reims?
- Manual classification requires lots of work, does not scale
 - Still, used effectively in many systems

How to classify? – cont.

- Linguistic approaches
 - Use parsing to identify "important" phrases in the question
 - Use knowledge such as WordNet to determine class
 - E.g. "Name a **female figure skater**"
- Machine learning approaches
 - Features: words, n-grams, POS from question
 - Classified into a pre-determined taxonomy of questions
 - Possibly hierarchically

Named entity recognition

- Classes of question usually closely related to the types of entities a system can identify
 - named entities
- Typical NEs
 - PERSON, LOCATION, ORGANIZATION, ...
- Additional in the QA setting
 - ABBREVIATION, QUANTITY, MANNER-OF-DEATH, ...

Named Entity Recognition – cont.

- How are entities recognized?
- Hidden Markov Model
 - BBN Identifier
- FSMs
 - IBM Textract
- Authority Files
- Probabilistic tagging
 - Rule out some possibilities, then tag with all the rest
 - E.g. "Beverly Hills" → possibly place, possibly person

PLACE	contemporary sporting sense was born when a young Genevese scientist, on a first visit to Chamonia in 1760, viewed Mont Blanc (at 15,771 feet)
REGION	Europe) and determined he would climb to the top of it or be responsible for prize money for the first ascent of Mont Blanc, but it was not until 1786, more than 20 years later, that his money was claimed by a Chamonia doctor, Michel-Gabriel Paccard, and his porter, Jacques Balmat. A year later, de Saussure himself climbed to the summit of Mont Blanc.
CONTINENT	After 1850 groups of British climbers with Swiss, Italian, or French guides scaled one after another of the high peaks of
COUNTRY	Switzerland.
COUNTRYPA	A landmark climb in the growth of the sport was the spectacular first ascent of the Matterhorn (14,692 feet) on July 14, 1865, by a party led by an English alpinist Edward Whymper. In the mid-19th century the Swiss developed a code of guides whose leadership helped make mountaineering a distinguished sport as they set the way to peak after peak throughout central Europe.
STATE	By 1870 all of the principal Alpine summits had been scaled, and climbers began to seek new and more difficult routes on peaks that had already been ascended. As the few remaining minor peaks of the Alps were overcome, by the end of the 19th century climbers turned their attention to the Andes of South America, the North American Rockies, the Caucasus, Africa's peaks, and finally the Himalayan vastness.
CITY	Aconcagua (22,831 feet), the highest peak of the Andes, was first climbed in 1852, and the Grand Teton (13,747 feet) in North America's Rocky Mountains was ascended in 1896. The Italian duke of the Abruzzi in 1897 made the first ascent of Mount St. Elias (18,009 feet), which stands at the international boundary of Alaska and Canada, and in 1906 successfully climbed Margherita in the Rwenzori Group (16,795 feet) in East Africa. In 1913 an American, Hudson Stuck, ascended Mount McKinley (20,320 feet) in Alaska, the highest peak in North America. The way was opening for greater conquests, but it would be midcentury before the final bastion, Mount Everest, was ascended.
CAPITAL	As the 20th century wore on, the truly international character of mountaineering began to reveal itself. Increasingly Austrians, Chinese, English, French, Germans, Indians, Italians, Japanese, and
NATIONAL	Russians turned their attention to opportunities inherent in the largest mountain landmass of the planet, the Himalayas. After World War I the British made Everest their particular goal. Meanwhile, climbers from other countries were making spectacularly successful climbs of other great Himalayan peaks. A Soviet team climbed Stalin Peak (24,590 feet), later renamed Communism Peak, in 1933; a German party
LANGUAGE	
MISC	
COMPOSITE	
MEDICAL	
COLLEGE	
SPORTS	
NAME	
NICKNAME	
BIO UNIAM	
ORG	
ROLE	
PERSON	
PRESIDENT	
ROYALTY	
DATE	

Document tagged with the IBM NE recognizer

How to formulate a query?

- Naïve approach: use question
 - Possibly stopped
- Recall problems → expansions:
 - Synonyms, hypernyms, ...
 - Units (dollar, pound, euro)
 - How? WordNet, thesaurus, lists
- Precision problems → restrictions
 - Question-to-answer reformulations
 - "Where is X?" → "X is in", "X is found in"
 - "When was X born?" → "X (DATE—DATE)"
 - How? Regular expressions, learning paraphrases

How to extract answers?

- Arguably, hardest part of the QA process
 - The biggest difference from standard IR
- Some methods:
 - Identify entities matching the answer type in retrieved documents/passages
 - Then: get highest frequency entity
 - Or: match patterns ("X was born on DATE")
 - Compare semantic structure of question and candidate sentence
 - Use logic to prove answer
 - ...

Why answer extraction is hard

- For the computer, a document looks like this

In a study that questions the industry's basic business model, economics professor Arthur De Vany's equations and economic modeling show that the combination of a big budget, top stars and a R rating may be the worst investment a Hollywood studio can make. In fact, the study concludes that not only are R-rated films less than half as likely as PG releases to gross \$25 million domestically, but that G, PG and PG-13 movies all generate better revenues and profits than R-rated films while keeping costs down. Yet, according to De Vany's study, more than half of the films released in the last decade were rated R, less than 3 percent were rated G and the remainder were split about evenly between PG and PG-13 movies.

- What are the relations between words?
- What do words mean?

Ranking candidate answers

- Global context
 - The relevance of the passage from which the candidate answer is extracted to the question
 - IR-engine score
 - Keyword overlap
 - IsFirstPassage, ...
- Local context
 - The likelihood that the answer fills in the gap in the question
 - Proximity of answer to question keywords
- Semantic type
 - The semantic type of a candidate answer should either be the same as or a subtype of the answer type identified by the question analysis component.
- Redundancy
 - How often the answer occurs in retrieved passages

What's next?

- A bit on natural language processing
- Question answering
 - Background
 - Anatomy of a question
 - History
 - The canonical architecture
 - QA at TREC
 - Back to retrieving content and structure
 - Wrapping up

Evaluation

- Question Answering is essentially a known-item-search task
 - Although there may be more than one correct answer
- Common measures
 - MRR of top N candidates given by a system, N=5,3,1
 - Precision@1
 - CWS: Confidence-Weighted Score
 - Used to check how "sure" the system is
 - First, system sorts its responses by confidence
 - Then,

$$CWS = \frac{1}{N} \sum_{i=1}^N \frac{\#correct \text{ up to rank } i}{i}$$

Question Answering at TREC

- Started in TREC-8 (1999)
 - Answers can be 50 or 250 bytes long
 - Systems return up to 5 answers
 - Answers had to be justified
 - Supply a "supporting document"
 - Scored by MRR
- TREC-9 (2000): similar
- TREC-10 (2001)
 - New: No-answer questions (NIL questions)

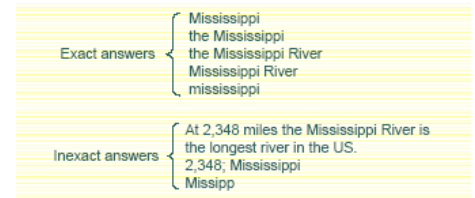
Questions with no answer

- Subtle difference between
 - This question has no answer (within the available resources),
 - This question has no answer (at all), and
 - I don't know the answer
- TREC-QA tests #1 (“NIL questions”), but systems typically answer as if #3
- Strategies used:
 - When allowed top N answers (with confidences)
 - Always put NIL in position X (X in {2,3,4,5})
 - If some criterion succeeds, put NIL in position X (X in {1,2,3,4,5})
 - Determine some threshold T, and insert NIL at corresponding position in confidence ranking (1-5, or not)
 - When single answer
 - Determine some threshold T, and insert NIL if answer confidence < T

Question Answering at TREC – cont.

■ TREC-11 (2002)

- Only 1 answer
- Exact answers only
- Scored by CWS



■ TREC-12 (2003)

- Definition questions, list questions

■ TREC-13 (2004)

- Scenario-based

■ TREC-14 (2005)

- Scenario-based

Definitions/“Others”

- Definition
 - Who is Aaron Copland?
 - What is a quasar?
- “Other” organized around a target
 - Target: **Americorps**
 - Factoid: **How many volunteers work for it?**
 - List: **What activities are its volunteers involved in?**
 - Other: **Tell me other interesting things about this target I didn't know enough to ask directly?**
- Targets
 - 2004: people, organizations
 - 2005: people, organizations, events

Issues with answering definitionoids

- Answer types are clear for many factoid questions
 - Who is . . . ? < person >
 - How long . . . ? < duration > or < length >
 - When did . . . ? < time >
- Answer type for definitionoids not easily characterised, not even with a very detailed question-typology
- No answer pattern produces 100% reliable results
 - <NP> is a (~25%)
 - Spelunking is a . . . fancy term for exploring caves (OK)/worthwhile pastime in Laos (NO)/popular activity (NO)
 - <NP> is (~18%)
 - Spelunking is . . . another name for the popular activity of caving (OK)/the touristic visitation of wild caves (OK)/my “caving pack” (NO)/\$68.00 (NO)
 - <http://www.isi.edu/natural-language/projects/webclopedia/Taxonomy/definitions.patterns>

Answering definitionoids

- Consult offline resources
 - E.g., “Answering Agents” (Chu-Carroll et al, 2002)
- For general terms
 - general lexical resources such as WordNet, with their glosses
 - e.g., cave, spelunk – (explore natural caves)
- For specialized terms, e.g., biomedical,
 - glosses in MeSH (Medical Subject Headings) (<http://www.nlm.nih.gov/mesh/meshhome.html>)
 - the NCI Metathesaurus <http://ncimeta.nci.nih.gov/indexMetaphrase.html>
- Alternative, indirect use of on-line resources to boost . . .
 - . . . retrieval status values (Prager et al, 2001a,b)
 - . . . importance of text snippet potentially answering a definitionoid or otheroid (Ahn et al, 2005)

Otheroids, offline of course

- Remember, TREC 2004 QA track organized around “targets”
 - Target: **Americorps**
 - Factoid: **How many volunteers work for it?**
 - List: **What activities are its volunteers involved in?**
 - Other: **Tell me other interesting things about this target I didn't know enough to ask directly?**
- Requirements on text snippets returned for “other” questions
 - Relevant
 - New (no duplicates, no overlap with answers returned under “factoid” or “list”)
 - Important
 - **Bill Clinton** boarded a plane for Mexico City last Monday
 - **Bill Clinton** was president of the US

Important snippets

- Estimate importance of “facts” found in a given (newspaper) collection by comparison with facts extracted from external “reference” corpus
 - E.g., Wikipedia
- Extract **importance models** rather than important facts
 - Will help with targets not covered in reference corpus
 - Will allow for novel facts from dynamic news sites
- Importance modeling for people
 - “Personalia”
 - Education
 - Career moves
 - “Known for”
 - Authors: book titles, Actors: movies, Politicians: party, etc.

What's next?

- A bit on natural language processing
- Question answering
 - Background
 - Anatomy of a question
 - History
 - The canonical architecture
 - QA at TREC
 - **Back to retrieving content and structure**
 - Wrapping up

Many helpful kinds of markup

- Document structure: paragraphs, sentences,...
- Named entities: locations, names, dates,...
- Part-of-speech: nouns, verbs, adjectives,...
- Parse trees: subjects, objects, predicates,...
- Semantic roles: agent, manner, topic, victim,...
- Anaphoric links: he, she, they,...
- Extracted knowledge: birth dates, manners of death,...
- External resources: semistructured encyclopedias,...
- External resources: structured databases (e.g., IMDB)
- ...

Many, many views

- Proposal
 - View annotated corpora, off-line tables, knowledge bases, etc, etc as semistructured data
 - Implement QA as retrieval of elements from the semistructured data
 - Map incoming question to (series of) queries against ssd
 - Combine query results
 - Select answer(s) and present
- Why?
 - Tight integration between question analysis and retrieval, and between retrieval and answer extraction/generation
 - Increased online performance
- Issues (warning: not all fully resolved yet)
 - Mapping questions to queries
 - Organizing the data – overlapping annotations
 - Querying the data, output format

Example: Who killed Abraham Lincoln?

- Sentence boundaries, paragraphs
 - //p//s[contains(killed) AND contains("Abraham Lincoln")]

<p> ... <s>John Wilkes Booth 5 Botten, spieren en gewrichten 47 Osteoporose Inleiding

- Can match simple templates of
- Document layout
 - Initial paragraph in Wikipedia
 - Initial paragraph in (many) new
 - Resources such as Merck Manual
 - Pages about diseased typical
 - types
 - symptoms
 - diagnosis
 - prevention and treatment

Symptomen

De botdichtheid neemt langzaam af, vooral bij patiënten met seniele osteoporose; in het te waarneembaar. Sommige mensen vertonen zelfs nooit symptomen.

Wanneer de botdichtheid zoveel afneemt dat beenderen bezwijken of breken, ontstaan er wervels inzakken (crushfractuur van de wervels) kan er chronische rugpijn optreden. De na licht letsel inzakken. Gewoonlijk begint de pijn plotseling, beperkt zich tot een bepaald wanneer iemand staat of loopt. Het gebied kan bij aanraking pijnlijk zijn, maar gewoonlijk paar weken of maanden. Als er verschillende wervels breken, kan er een afwijkende krom ontstaan, die pijnlijke en gespannen spieren veroorzaakt.

Er kunnen ook andere beenderen breken, vaak door een geringe belasting of een val. Een heupfractuur, een belangrijke oorzaak van invaliditeit en verlies van zelfstandigheid bij ou (radius) aan de kant van de pols, fractuur van Colles genaamd, komt ook regelmatig voor ouderen met osteoporose over het algemeen langzaam.

Diagnose

Bij patiënten met een botbreuk wordt de diagnose osteoporose gebaseerd op een combinatie van röntgenfoto's van de beenderen. Verder onderzoek kan noodzakelijk zijn om osteoporose leidende aandoeningen uit te sluiten.

Nog voor er een breuk optreedt, kan de diagnose osteoporose worden gesteld met onder meest accurate onderzoek is 'dual-energy x-ray'-absorptiometrie (DEXA). Dit onderzoek in minuten worden uitgevoerd. Het is zinvol voor vrouwen met een verhoogd risico van oste diagnose onzeker is, of bij patiënten bij wie de behandelresultaten goed moeten worden in

Preventie en behandeling

Preventie is succesvoller dan behandeling. De preventie bestaat uit het handhaven van voldoende calciuminname, door lichaamsbeweging waarbij de botten worden belast en, bij van geneesmiddelen.

Named entities

- Person names, locations, organizations, etc
 - //s[contains(killed) AND ./person="Abraham Lincoln"]
- <s>
 - <person>John Wilkes Booth</person>
 - killed
 - <person>Abraham Lincoln</person>
 - </s>
- If expected answer type is a named entity, can enforce its presence
- Can't ask for "killed" next to "<person>Abraham Lincoln</person>"

Part-of-speech

- //s[contains(./VBD, killed) AND ./person="Abraham Lincoln"]
- But all words would be marked up:
 - <s>
 - <person><NNP>John</NNP><NNP>Wilkes</NNP><NNP>Booth</NNP></person>
 - <VBD>killed</VBD>
 - <person><NNP>Abraham</NNP><NNP>Lincoln</NNP></person>
 - </s>
- //s[contains(./VBD, killed) AND ./person[./NNP[1]="Abraham" AND ./NNP[2]=last()]="Lincoln"]
- Starting to get ugly

Parse trees

- Multiple parses cannot be encoded in same hierarchy
 - Multiple correct parses
 - Weighted output of a parser
- What's the correct order for nesting named entities in the tree?
 - <s>
 - <NP><person><NNP>John</NNP> <NNP>Wilkes</NNP>
 - <NNP>Booth</NNP></person></NP>
 - <VP><VBD>killed</VBD>
 - <NP><person><NNP>Abraham</NNP> <NNP>Lincoln</NNP>
 - </person></NP>
 - </VP>
 - </s>
- //s[contains(./VBD, killed) AND ./person[./NNP[1]="Abraham" AND ./NNP[2]=last()]="Lincoln"]

Extracted knowledge

- <s>
- <NP><person><NNP>John</NNP> <NNP>Wilkes</NNP>
- <NNP>Booth</NNP></person></NP>
- <VP><VBD>killed</VBD>
- <NP><person><NNP>Abraham</NNP>
- <NNP>Lincoln</NNP></person></NP>
- </VP>
- </s>
- <opv-triple>
- <object>John Wilkes Booth</object>
- <property>killed</property>
- <value>Abraham Lincoln</value>
- </opv-triple>
- Can we always keep our hierarchies nicely nested?

When can annotations overlap?

- Annotations done automatically
 - Tools make mistakes
 - Analysis can be ambiguous
- Inconsistencies between tools: words, punctuation,...
- Hugo de Vrieslaan
- Intrinsically "overlapping" syntactic/semantic structures
 - "Arlington police shot and killed Kenneth Harris"

QA as SSIR: challenges

- Rankings
- Term weighting
- Order and proximity
- Multiple hierarchies... overlap
- Expressing relations between hierarchies
- Approximate matching of hierarchy relationships

Rankings

- Crucial to effective retrieval
- Results should be ordered on probability of containing answers
 - Contrast with: unordered set resulting from Boolean queries
 - Same as relevance ranking examined at INEX?
- Answer profiles?
- QA: returning a ranked list of answers
 - Think of "Where's the Rijksmuseum located?"

Term weighting

- Give feedback to the system on which terms are important
 - Rankings should gracefully decline when query terms aren't matched
- Provide alternatives
 - 1 President Lincoln
 - 1 Abraham Lincoln
 - 0.3 Honest Abe
 - 0.1 Lincoln
- Can be fixed in query language

Order and proximity

- Some challenges already met
- Would like to ask term "killed" adjacent to a "person" component
- Could easily be addressed in query language
- Should not have much impact on indexes

Multiple hierarchies

Offset annotation

1. John
2. Wilkes
3. Booth
4. killed
5. Abraham
6. Lincoln

Hang on ...

```

<named-entities>
  <person beg="1" len="3"/>
  <person beg="5" len="2"/>
</named-entities>
<parse-trees>
  <S weight="1" beg="1" len="6">
    <NP beg="1" len="3"/>
    <VP beg="4" len="3">
      <NP beg="5" len="2">
        </VP>
      </S>
    </parse-trees>
  <opv-triples>
    <opv-triple>
      <object beg="1" len="3"/>
      <property beg="4" len="1"/>
      <value beg="5" len="2"/>
    </opv-triple>
  </opv-triples>

```

Offset annotation

- Allows multiple hierarchies to be cleanly represented
- Requires new indexing methods
 - But should not be too cumbersome
- Querying
 - Refer to the text in a tag easily
 - Should be able to express desired "descendants" across hierarchies
 - A person in a sentence
 - A person in an object of extracted knowledge
 - Requires new query constructs or redefinition of existing query languages

XML and Multiple Hierarchies

- (Durand, 1996): *...breaking of strict hierarchies is the rule rather than the exception...*
 - Different kinds of analysis, different "views" on data
 - Possible errors in case of automatic annotation
 - Discontinuous and overlapping phenomena
- What if our annotation requires overlapping tags?
 - Switch from our beloved XML to something more powerful? (LMNL)
 - Try to encode all we need in "plain" XML?
 - Design XML extensions?
- Overlapping Markup SIG (TEI-OM-SIG)
- But first, what kind overlap we are talking about?

Bible: a structured view

- The most read, studied, queried and quoted collection
- Two obvious (independent) hierarchies:
 - book / chapter / verse
 - book / story / paragraph / sentence
- But scholars want much more! (Jer.2):

(1) Moreover the word of the LORD came to me, saying,

(2) Go and cry in the hearing of Jerusalem, saying,

Thus says the LORD:

I remember you,

The kindness of your youth, The love of your betrothal, When you went after Me in the wilderness, In a land not sown.

(3) Israel [was] holiness to the LORD,...

Overlapping markup: approaches

- Fragmentation of elements
 - Splitting secondary components
 - Joins
- Simultaneous annotations
 - SGML CONCUR
 - Milestones
 - Just-In-Time-Trees (JITTs)
- Stand-off annotations
 - Bottom Up Virtual Hierarchies (BUVH)
 - Distributed XML: GODDAG

- No standards yet, no perfect solutions, but let's have a look

Fragmentation of elements

- Choose one "primary" hierarchy
- Split all other problematic elements

```
<verse>Moreover the word of the LORD came to me, saying,</verse>
  <verse><q splitID="1">Go and cry in the hearing of Jerusalem, saying,
    <q splitID="2">Thus says the LORD:
      <q splitID="3">I remember you,
        The kindness of your youth, The love of your
        betrothal, When you went after Me in the
        wilderness, In a land not sown.</q></q></verse>
  <verse><q splitID="1"><q splitID="2"><q splitID="3">
    Israel [was] holiness to the LORD,...
```

- Do you want to add paragraphs and sentences?

Assembling fragments

- TEI join element:


```
...<q id="x1">...</q> ... <q id="x2">...</q> ... <q id="x3">...</q>...
<join targets="x1 x2 x3" result="qs" scope="root"/>
```
- Supports almost any conceivable structure (even discontinuous)
- Complex modifications in case of changes

Co-existing annotations: SGML CONCUR

- A part of SGML standard


```
<(DTD1)p>And the Lord said,
<(DTD2)q>Read my lips: Do not murder.</(DTD1)p>

<(DTD1)p>Be nice to each other instead.</(DTD2)q>
And the people said "Amen."</(DTD1)p>
```
- No way to constrain relationships between DTDs
- No self-overlap (elements from one DTD)
- Virtually not support by SGML software

Milestones and Trojan milestones

- Insert empty elements


```
<p> text <q> quotation </p> quotation goes on </q>

<p> text <start gi="q"/ id="x1"> quotation </p>
  quotation goes on <end gi="q" coid="x1"/>

<p> text <q sID="x1"/> quotation </p> quotation goes on <q eID="x1"/>
```
- Schema languages cannot enforce well-formedness
- Becomes CLIX and is related to LMNL

LMNL and JITTs

- Layered Markup Annotation Language:
 - Non-XML syntax: [tag] ... {tag}
 - Arbitrary overlap
 - Disambiguation possible (but not obligatory) using IDs
 - But isn't not XML...Query languages?
- Just-In-Time-Trees
 - Similar to CONCUR and LMNL
 - XML-like syntax
 - Arbitrary overlap
 - Tags are filtered only when the file is processed
 - Again, no self-overlap, weak validation capabilities

Bottom Up Virtual Hierarchies

- Use XPath to "invert" multiple hierarchies

<pre><pages> <page id="p1"> <line id="l1"> <w id="w1">This</w> <w id="w2">is</w> </line> </page></pre>	<pre><text> <para id="par1"> <sent id="s1"> <w id="w1">This</w> <w id="w2">is</w> <w id="w3">text</w>.</pre>
--	--

```
<baseFile>
  <w id="w1"
  pages="/pages/page[1][@id='p1']/line[1][@id='l1']/w[1]"
  sents="/text/para[1][@id='par1']/sent[1][@id='s1']/w[1]"
  This</w>
  <w id="w2"
  pages="/pages/page[1][@id='p1']/line[1][@id='l1']/w[2]"
  sents="/text/para[1][@id='par1']/sent[1][@id='s1']/w[2]"
  is</w>
  ...
```

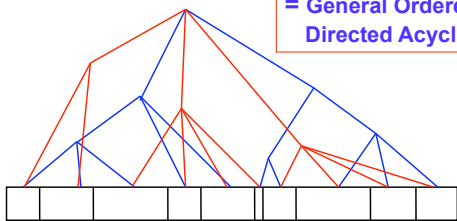
Bottom Up Virtual Hierarchies (2)

- Allows cross-hierarchy querying/XSLT by string concatenation
- Completely unreadable
- Dropped by Society of Biblical Literature in favour of JITTs

Going “away”: stand-off annotation

- Keep different annotations separate
- A distributed XML document is GODDAG
- Hierarchies joined at the root and text nodes
- Natural interpretation: *multiple views of the same data*

= General Ordered-Descendant
Directed Acyclic Graph



Going “away”: stand-off annotation

- Querying Concurrent Markup Hierarchies: EXPath
- All hierarchies visible via new axes:
 - xdescendant: all nodes with “smaller” text ranges
 - xancestor: all nodes with “larger” text ranges
 - xfollowing: all nodes whose range “follows” range of the current node
 - xpreceding: all nodes whose range “precedes” range of the current node
 - preceding-overlapping: nodes whose range contains start tag but not end tag
 - following-overlapping: nodes whose range contains end tag but not start tag
- Ionut Emil Iacob: PhD in progress
- Matches ILPS’s current QA-as-SSIR efforts
 - Now implemented in an ad-hoc way
- Does it work?
 - CLEF 2005, TREC 2005

What’s next?

- A bit on natural language processing
- Question answering
 - Background
 - Anatomy of a question
 - History
 - The canonical architecture
 - QA at TREC
 - Back to retrieving content and structure
 - **Wrapping up**

Wrapping up

- Negation in answer passages
 - Q: Who invented the electric guitar?
 - A: While Mr. Fender did not invent the electric guitar, he did revolutionize and perfect it.
- Negation in questions
 - Q: Name a US state where cars are not manufactured
 - Answer unlikely to be presented explicitly
- Name an astronaut who nearly made it to the moon

And more issues: richer question types

- Simple factoids (When was Madonna born?)
 - approaching “solved problem” status
- More complex factoids
 - How did Chicago get its name?
 - What Arthur Miller play recounts his marriage to Marilyn Monroe?
- Subjective
 - What is the general opinion of Gun Control in the U.S.?
 - What is the best restaurant in Hawaii?
- Non-factoid
 - Why did Iraq invade Kuwait?
 - How do I install a dual-boot PC?
- Is there a God?

Recap

- QA raises interesting challenges for document retrieval
- NLP for QA raises interesting semistructured document retrieval issues
 - Some easily addressed in query language extensions
 - Some need different representations and indexing
- QA could benefit from structured retrieval
- Additional interesting research issues on the interface of QA and SSIR
 - Mapping questions to (sets of) queries
 - INEX’s NLP track?

Upshot

- NLP helps to create richly annotated documents
 - Becoming more common
 - Question answering
 - Biotext
 - Semantic Web
- These structured documents will pose challenges to XML retrieval
- Explore these problems and tasks
 - To optimize indexing
 - To have appropriate query languages
 - To be able to represent structures needed

Sources

- James Allan
 - *NLP for IR*, NAACL/ANLP 2000, Seattle
- Boris Katz, Jimmy Lin
 - *QA Techniques for the World Wide Web*, EACL 2003, Budapest
- Maarten de Rijke, Bonnie Webber
 - *Question Answering*, ESSLLI 2003, Vienna
- John Prager
 - *Tutorial on Question Answering*, RANLP 2004, Borovets
- Paul Ogilvie
 - *Retrieval Using Structure for Question Answering*. In: *Proceedings of the First Twente Data Management Workshop (TDM’04)*, 2004
- Valentin Jijkoun, Gilad Mishne, Maarten de Rijke
 - *From Question Answering to Semistructured Retrieval*, SIKS Spring School on XML, Vught, 2005