Search engines, Question Answering and Syntactic Analysis

Kaarel Kaljurand (kaarel@ut.ee)
Tartu University

Theory Days in Koke 2004, Koke, Estonia

Outline of the talk

- Search (information retrieval, information extraction, question answering)
- Problems with currently available search tools (e.g. Google)
- Currently available NLP tools and how they can be put to use: Question Answering system
- Closer look to syntactic analysis in Question Answering

The search problem

- Definition: provide an answer to a statement of user's information need
- How is this statement formulated?
- How is the answer formulated?
- What are the features of the knowledge source?
- How to process the knowledge source (= understand its meaning)?

The search problem (cont.)

- Knowledge source
  - Database (information is highly structured)
  - Web (natural language, redundancy)
  - Small text collection (e.g. technical manual)
- Information need
  - Summarization
  - ”List of the characters in Hamlet.”
  - ”What did the author want to say in this essay?”
  - ...

Keyword-based (web) search

- Keyword-based search: mapping a set of keywords to a set of documents
- Query as a Boolean formula (”pet” AND ”dog” AND-NOT ”cat”)
- Bag-of-words model to represent documents
- Ranking
- Small amount of NLP: lemmatization, stop-word lists

Problems with keyword-based search

- Documents are written in natural language: ambiguity (synonymy, polysemy) exists at every level of language
- User has to convert his question into a set of keywords, not very intuitive (”Find a document that contains the word ‘dog’”)
- Too many results usually retrieved
- Result unit is a file (which can be of any size), instead of a linguistic unit, e.g. a sentence or a paragraph

Overcoming the problems

- Phrase search, to overcome poor syntax modeling (probably works better with English where the word order is more fixed)
- Ranking (using meta-information like links), classification (teoma.com)
- Excerpts and highlighting (to overcome big text sizes)
- Location information, personalized results
- NLP: lemmatization, query expansion with synonyms (from e.g. WordNet)

NLP intensive search: Question Answering

- Maps a natural language question to natural language (short) answer
- As ambitious as Machine Translation, tries to understand the documents by applying analysis of all levels of language
- Interesting are NLP intensive methods, although QA can be attempted by simple pattern matching + wrapper for keyword-based search (e.g. askjeeves.com)

Levels of language analysis

- Morphology: dog = dogs, quick = quickly, koer = koerakeselikkusegagi
- Syntax: John gave Mary a book = A book was given to Mary by John
- Semantics:
  - John gave Mary a book = Mary got a book from John
  - John would have run = John runs
  - ‘vi’ edits texts = ‘vi’ is a text editor
  - John kills himself = John kills John
  - John kills Mary $\Rightarrow$ Mary is dead

- Pragmatics: John $\in$ Person, CEO $\in$ JobTitle

Components of languagecomputer.com

- Named Entity Recognition (names of companies, persons, locations etc.)
- Syntactic Analysis (noun and verb groups, PP attachments)
- Coreference Resolution (President Bush = Georg W. Bush)
- Meta-information extraction from WordNet glosses
- Logical Form Generation
- Theorem proving (with Otter)

Document representation example

Heavy selling of Standard & Poor's 500-stock index futures in Chicago relentlessly beat stocks downward.

heavy_JJ(x1) & selling_NN(x1) & of_IN(x1,x6) & Standard_NN(x2) & &_CC(x13,x2,x3) & Poor_NN(x3) & 's_POS(x6,x13) & 500-stock_JJ(x6) & index_NN(x4) & future_NN(x5) & nn_NNC(x6,x4,x5) & in_IN(x1,x8) & Chicago_NN(x8) & relentlessly_RB(e12) & beat_VB(e12,x1,x9) & stocks_NN(x9) & downward_RB(e12).

Question Answering screenshot

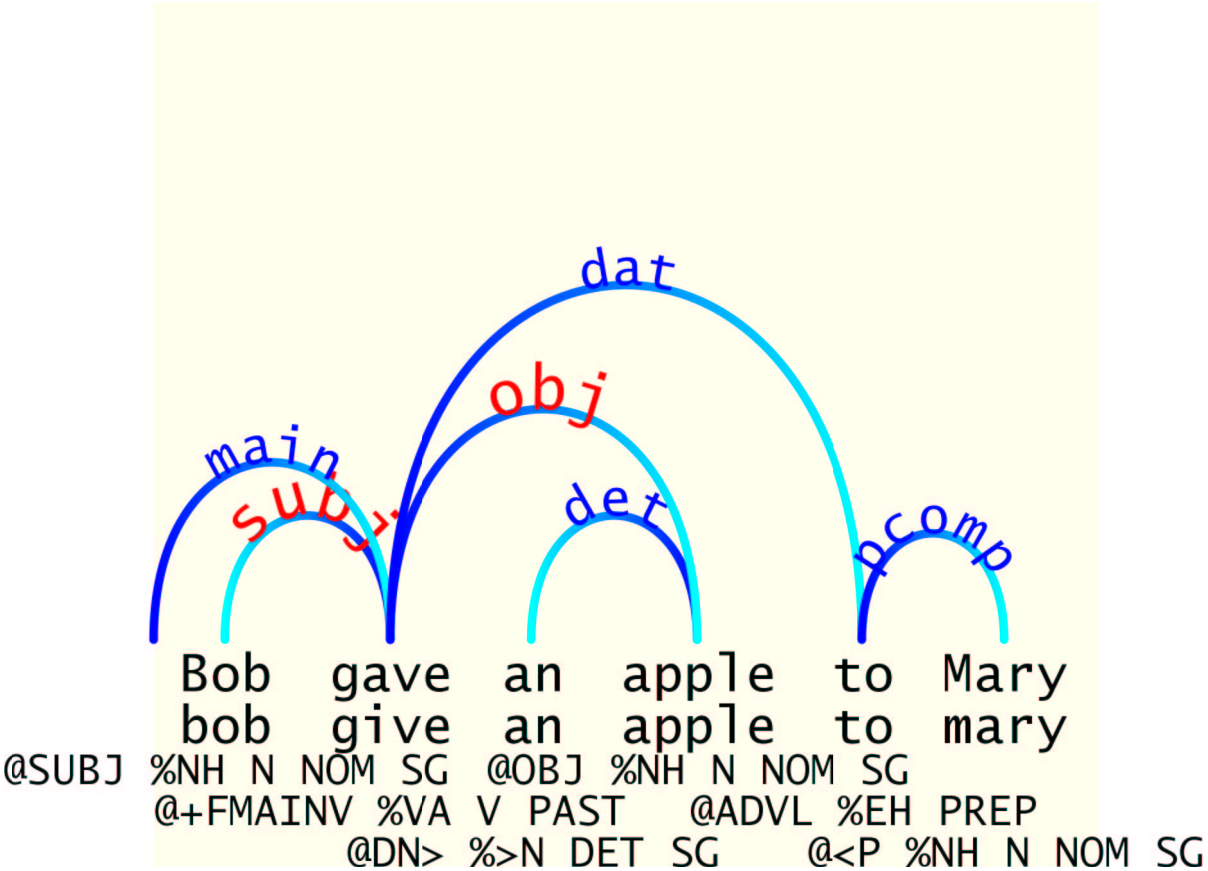
1. O [2] is the oxygen we breathe , and it makes up about 21 percent of the Earth 's air...
   earthobservatory.nasa.gov/Study/ProtonOzone/
2. Analyses of the gases in these bubbles show that the earth s atmosphere , 67 million years ago , contained nearly 35 percent oxygen compared to present levels of 21 percent...
   minerals.cr.usgs.gov/gips/na/0amber.htm
3. Ward , working with UW biologist Raymond Huey and UW radiologist Kevin Conley , believes that breathing system , still found in todays birds , made the Saurischian dinosaurs better equipped than mammals to survive the harsh conditions in which oxygen content of air at the Earths surface was only about half of todays 21 percent...
   www.innovations-report.de/...linaere_forschung/bericht-22933.html
4. The troposphere is the layer closest to the Earths surface and contains more than 90 percent (by weight) of all the gases in the atmosphere , It is composed of about 78 percent nitrogen , 21 percent oxygen , trace gases , water droplets , dust , and other particles...
   www.enviroliteracy.org/category.php/1.html
5. Background A significant portion of the earth is made up of water , more specifically , salt water , comprised mostly of the following elements : oxygen (89 percent) , hydrogen (six percent) , chlorine (two percent) , sodium (one percent) and bromine (one percent)...
   www.oxychem.com/...ts/chlorine/literature/elemental_chlorine.html
6. The Earth ' s atmosphere is composed of a mixture of gases , mostly nitrogen (78 percent) and oxygen (21 percent)...

Open domain QA: What percent of the Earth's air is oxygen?

Syntax formalisms

- Phrase Structure Grammar (Chomsky 1957)
  - Focuses on phrase structure
  - Analysis and generation
  - Sensitive to word order
- Dependency Grammar (Tesnière 1959, Mel’ĉuk 1987)
  - Focuses on binding words
  - Compatible with free word order languages
  - Structure is ”more semantic”
  - Less focus on grammatical correctness

Dependency Grammar example



Subject, object and indirect object

Closeness to semantics

- Syntactic relations map nicely to semantic ones:
  - subject $\mapsto$ actor
  - object $\mapsto$ patient
  - adjective modifier $\mapsto$ property

Levels of dependency analysis

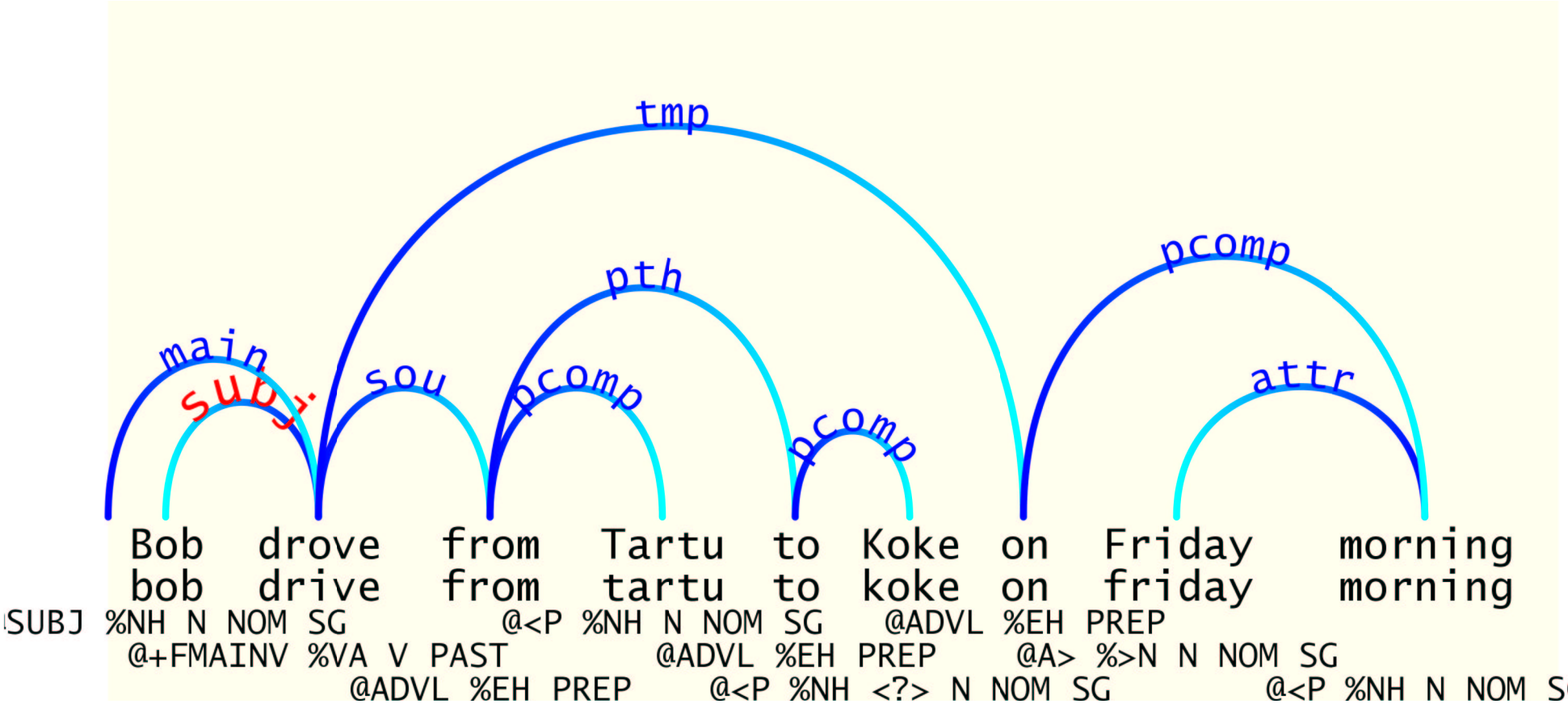
- Shallow
  - The nature of modification (e.g. subject) is specified, but not the target
  - Quite reliable (Constraint Grammar: $\sim$95% of reliability for English)
- Deep
  - The full relation is specified, e.g. `subject(run, dog)`
  - Subject and object relations detected correctly $\sim$90% of the times

- Difficult problems, e.g. PP-attachment (‘I saw a man with a hat’ vs. ‘I saw an ant with a microscope’)
- Existing systems: Connexor Machinese Syntax, MINIPAR, Link Parser etc

Deep Dependency Grammar rules

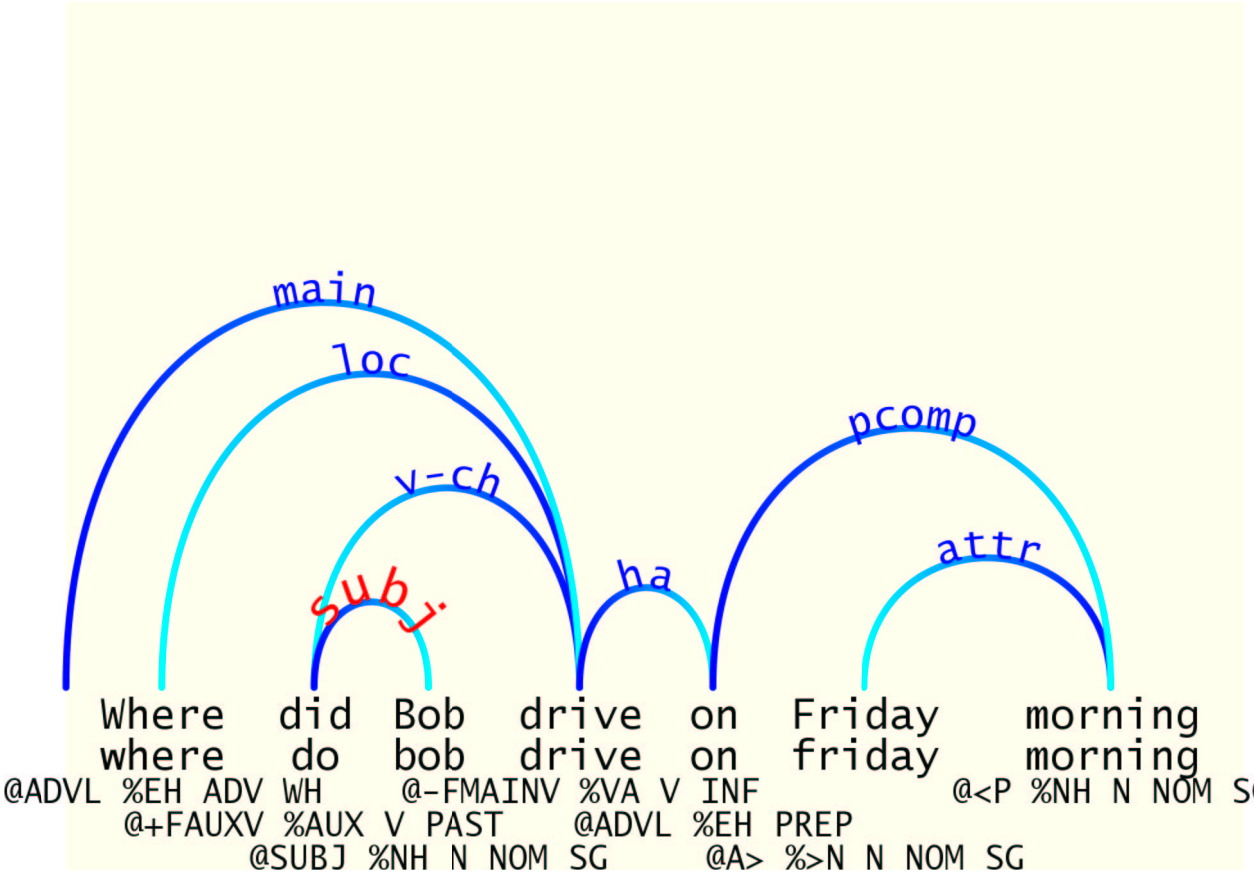
- Each word in the sentence modifies (is a dependent of) another word (so called ”head”)
- Each word can modify only one head
- Head-modifier relations have types (e.g. main verb, subject, object, attribute)
- The sentence structure is a tree (no modification cycles are allowed)

Example 1



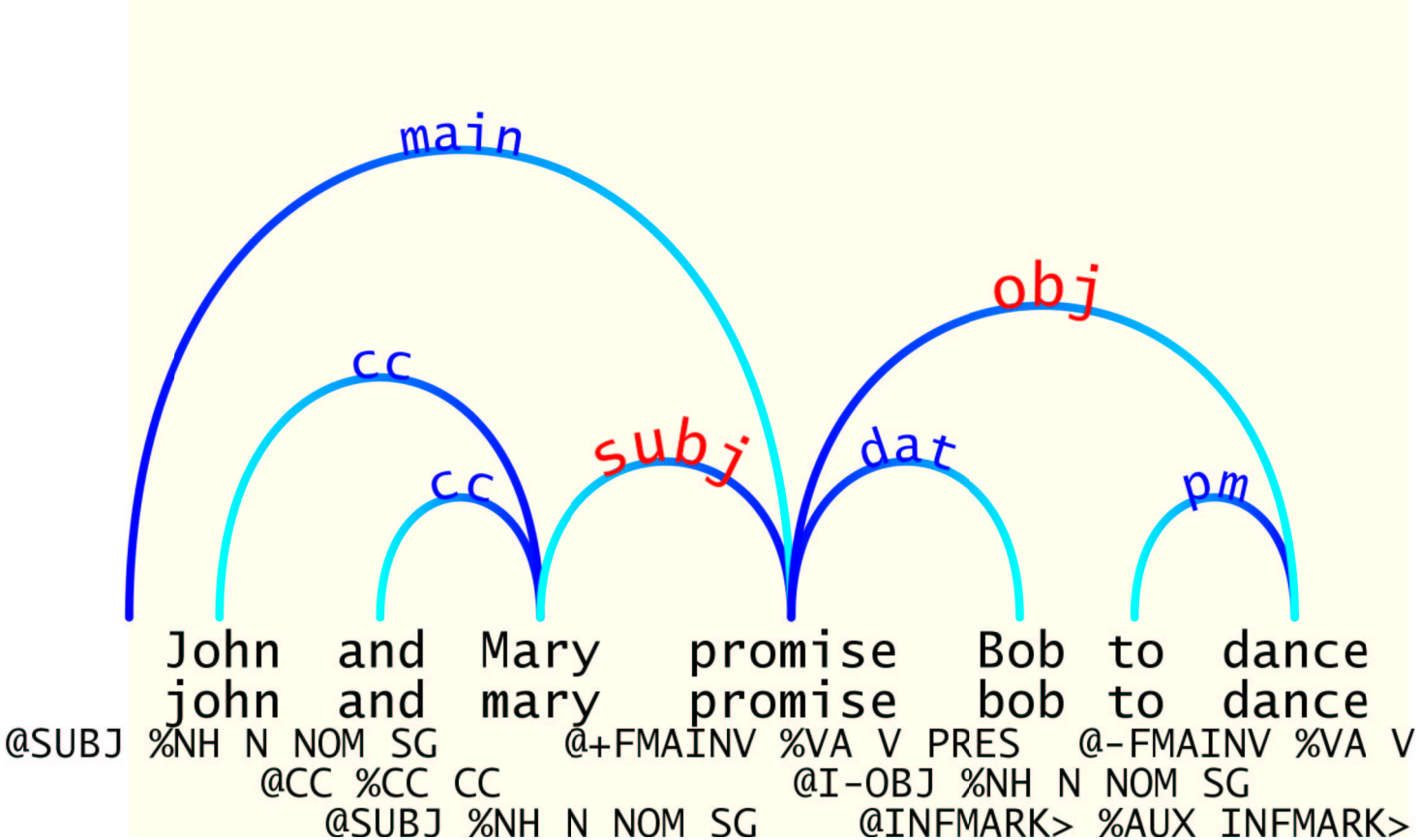
Classification of adverbs

Example 2



Question analysis

Example 3



Coordination, control structures: John and Mary are subjects of ‘promise’ and ‘dance’

Existing Estonian NLP tools

- Morphological analyzer
- A shallow dependency parser based on Constraint Grammar formalism
- WordNet semantic dictionary