

Modelling and utilizing uncertainty in ontology based semantic search

Henrik Legind Larsen

Department of Electronic Systems

Aalborg University

Esbjerg Institute of Technology

Niels Bohrs Vej 8, DK-6700 Esbjerg

legind@aaue.dk

DanNet Symposium, 12 March 2009

Overview

- I. Application system example: Ontology-based spoken query answering for mobile information access
- II. Extracting useful ontological knowledge from various resources
- III. Remarks on some knowledge representation formalisms
 1. Description Logic
 2. Fuzzy Concept Algebra
- IV. Utilization of ontologies in information retrieval and query answering
 1. Recognition of query concepts in document text
 2. Heuristic semantic disambiguation of query terms
- V. Utilization of ontologies for semantic disambiguation in "connecting the dots"
- VI. Conclusion

I. Application system example: Ontology-based spoken query answering for mobile information access

Developed at the Software Intelligence and Security Research Center (SIS-RC) as part of the project *Perceived QoS of Services in Heterogeneous Networks* (POSH) at the Center for InfraStruktur (CTIF), Aalborg University.

Published in:

Tom Brøndsted, Henrik Legind Larsen, Lars Bo Larsen, Børge Lindberg, Daniel Ortiz-Arroyo, Zheng-Hua Tan, Haitian Xu: Mobile Information Access with Spoken Query Answering. *Proceedings of the Applied Spoken Language Interaction in Distributed Environments, November 10th and 11th 2005*, Aalborg University, Denmark.

- The system can answer spoken questions in a limited domain (here: soccer).
- It applies a soccer ontology (soccer concepts and individuals) for understanding questions.
- For retrieving the most relevant news articles, it further applies a term association relation.
- If the system can understand the question and answer it from the database:
 - it provides the direct answer, plus
 - additional information of typical interest, including video clips, and links to the most relevant articles
- *otherwise:*
 - the system retrieves the most relevant articles, with the screen window focussed on the subtext in the article that is most likely to contain the answer.

Continued

I. Application system example:
Ontology-based spoken query answering for mobile information access

Ontology creation, maintenance, and utilization:

- Soccer ontology: *manually created*
- Information on individuals/instances (clubs, players, coach, ect.):
automatically extracted from tabular information on relevant websites
- The term associations: *created by text mining of large set of sport articles*
- The soccer ontology (concepts and individuals) provides domain knowledge that is utilized for extraction of news events from soccer feeds and soccer news sites
- Extracted soccer news are is stored in a database used for question answering

Continued

I. Application system example:
Ontology-based spoken query answering for mobile information access

Q-A examples:

- **Who scored in the match between Brøndby and FC Copenhagen?**
 - *Direct answer:* Lists of the players who made the goals, the time (in the match) the goal was made, and who won.
 - *Supplementing information:* Link to a video stream (10 – 20 seconds) for each goal.
 - *Retrieved articles:* First article is an interview with the Brøndby trainer.

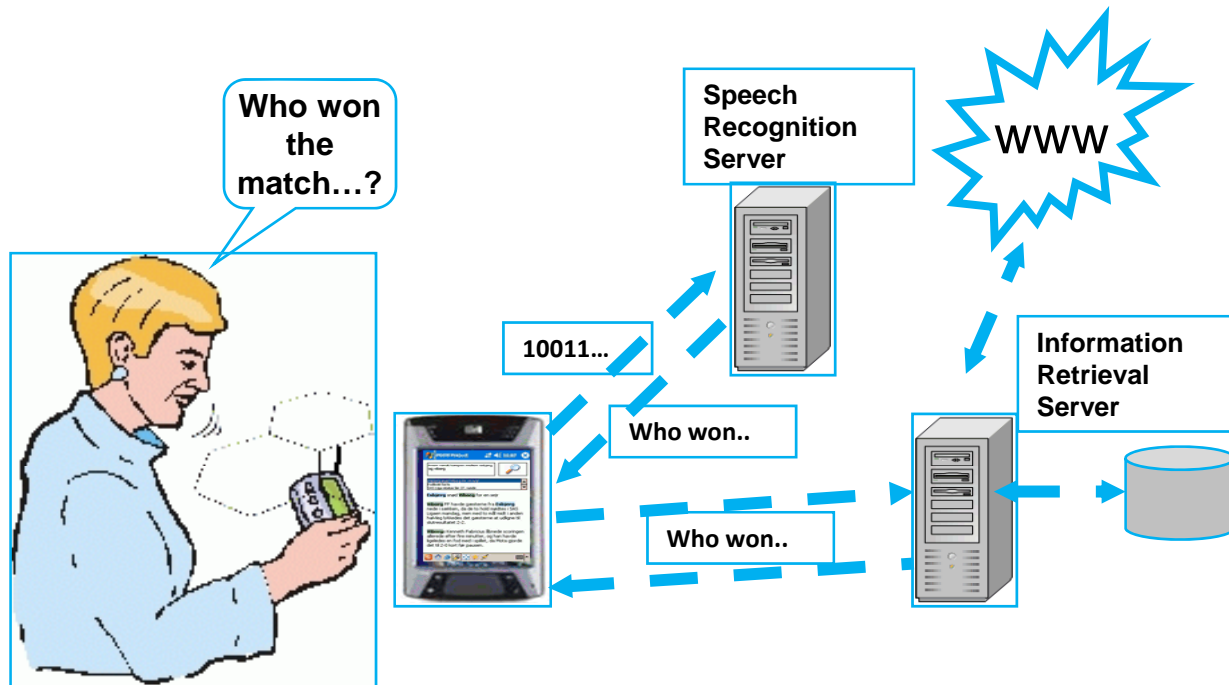
- **Who is Odense's next opponent?**
 - *Retrieved articles:* First article with window centred on the text:
“FC Nordsjælland is also OB's next opponent.”

Continued

I. Application system example:
Ontology-based spoken query answering for mobile information access

Question: Who won the match between Esbjerg and Viborg?

The answer displays the date, result, and the players who made goals. It further provides a list of the most relevant news articles.

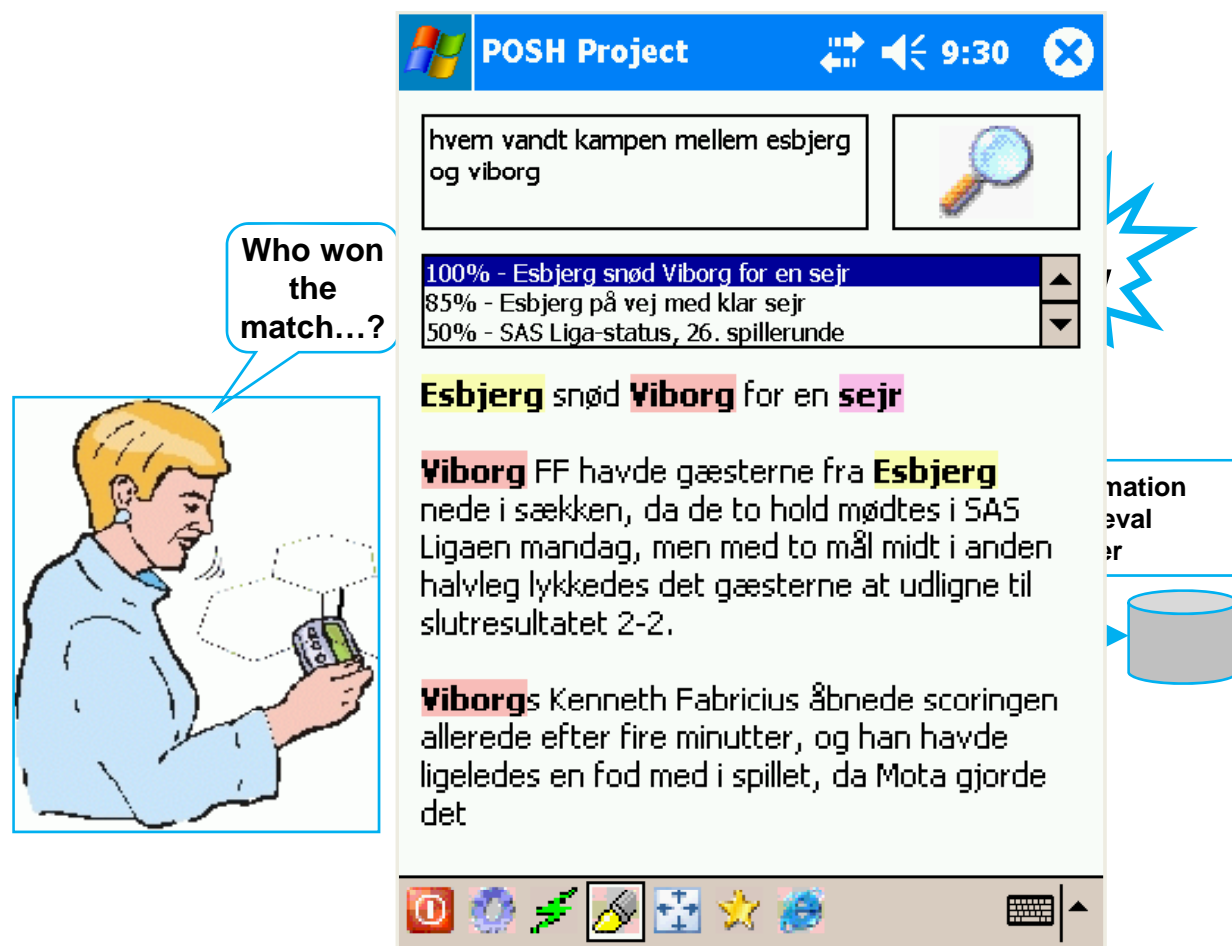


Continued

I. Application system example:

Ontology-based spoken query answering for mobile information access

In the list of relevant articles, the first three are shown, with the first displayed:

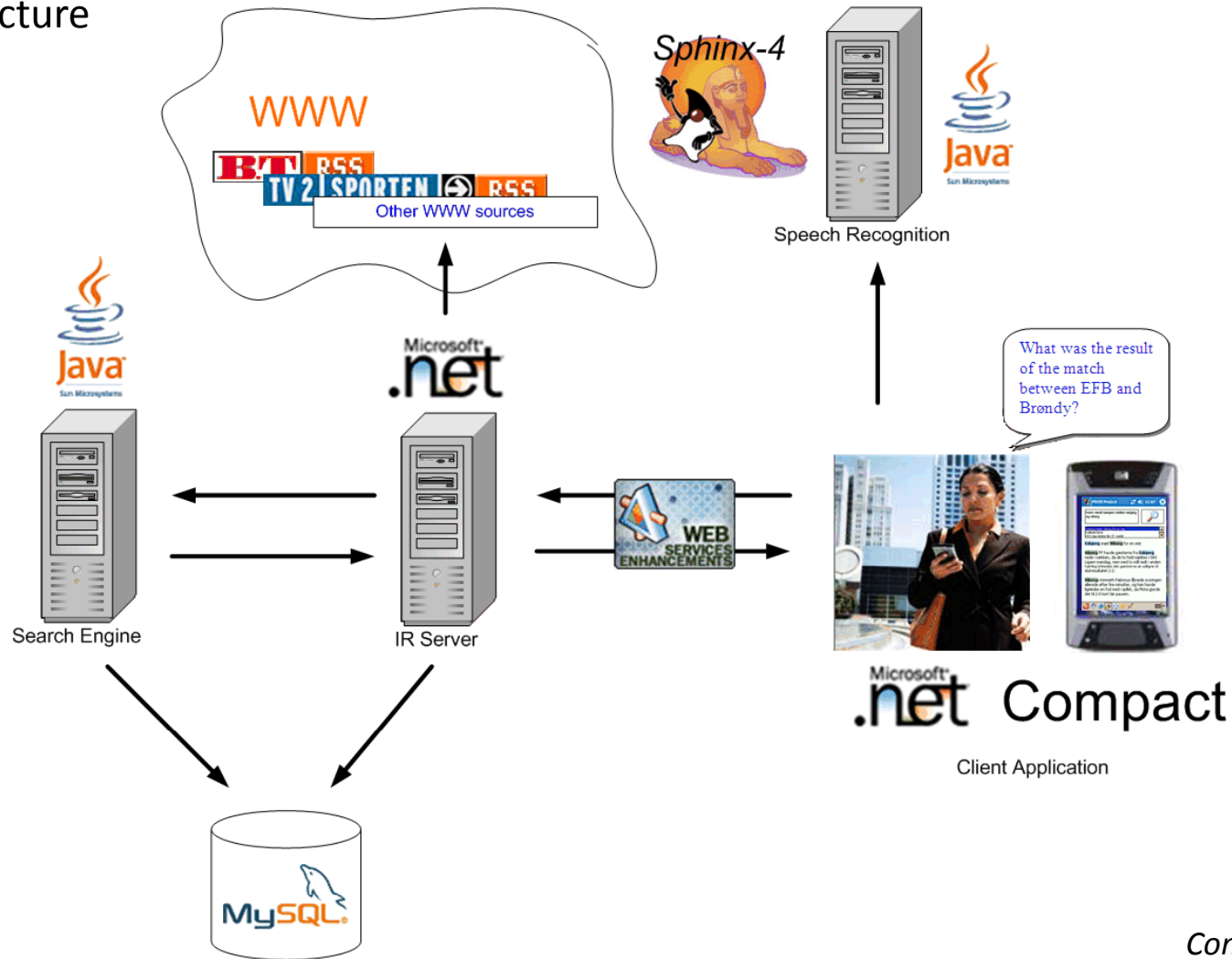


The image shows a woman on the left with a speech bubble that says "Who won the match...?". To her right is a screenshot of a mobile application window titled "POSH Project". The window has a blue header bar with a Windows logo, the title "POSH Project", and icons for navigation, volume, and time (9:30). Below the header is a search bar containing the text "hvem vandt kampen mellem esbjerg og viborg" and a magnifying glass icon. A list of search results is displayed below the search bar, with the first result selected and highlighted in blue: "100% - Esbjerg snød Viborg for en sejr". The other two results are "85% - Esbjerg på vej med klar sejr" and "50% - SAS Liga-status, 26. spillerunde". Below the list, the first article is displayed in full. The title is "Esbjerg snød Viborg for en sejr". The text of the article is: "Viborg FF havde gæsterne fra Esbjerg nede i sækken, da de to hold mødtes i SAS Ligaen mandag, men med to mål midt i anden halvleg lykkedes det gæsterne at udligne til slutresultatet 2-2." Below this, another paragraph starts with "Viborgs Kenneth Fabricius åbnede scoringen allerede efter fire minutter, og han havde ligeledes en fod med i spillet, da Mota gjorde det". At the bottom of the application window is a taskbar with several icons: a red stop button, a gear, a green arrow, a yellow pencil, a blue cross, a yellow star, a blue globe, and a keyboard icon. To the right of the application window, there is a blue lightning bolt icon and a box labeled "Information eval" with a cylinder icon below it.

Continued

I. Application system example: Ontology-based spoken query answering for mobile information access

Architecture



Continued

I. Application system example:

Ontology-based spoken query answering for mobile information access

The speech recognizer was trained (using a large corpus) to understand (speaker independent) questions in Danish on soccer.

The example has shown:

Domain ontologies can be utilized for:

- Information extraction from text
- Interpretation of questions and queries
- Question answering (NL question -> direct answer from database)
- Information retrieval (unstructured query -> rank list of relevant documents)

Term associations (derived by text mining) can supplement ontologies in information retrieval.

II. Extracting useful ontological knowledge from various resources

For ontological knowledge can be extracted from various sources, such as:

- Ontologies (wordnets, domain ontologies, etc.)
- Dictionaries

Useful associations can be extracted by mining of:

- Text collections (e.g., document collections and email sets)
- User search logs

For fast utilization, the extracted knowledge may be represented in a simple, but computationally efficient form, namely by so-called **fuzzy semantic-term nets (FSTN)**

FSTN is, in its simplest form, depicted by a weighted, directed graph:

- Nodes represent entities
- Edges represent direction and strengths of relationships between entities

Continued

II. Extracting useful ontological knowledge from various resources

The direction of a relationship between two entities is the *evidence endorsing* direction.

An entity e_i is related to an entity e_j to the degree s_{ij} that:

$$\exists x : e_i(x) \Rightarrow \exists y : e_j(y)$$

This is in the FSTN represented by the edge: $e_i \xrightarrow{s_{ij}} e_j$

In particular, for the semantic abstraction relations:

KO (Kind Of): e_i KO e_j (e_i is a kind of e_j)

IO (Instance Of): e_i IO e_j

FO (has Feature Object, or has part): e_i FO e_j

that are all represented by $e_i \xrightarrow{s_{ij}} e_j$ with $s_{ij} = 1$.

III. Remarks on some knowledge representation formalisms

1. Description Logic

Description Logic (DL)

DL is the formalism adopted by the semantic web.

Advantages:

- Semantically rich
- Well-defined semantics
- Terms: Entities (concepts and individuals) and relationships (roles)
- Computational aspects reasonably well-known

Disadvantages:

- No way of dealing with uncertainty (vague concepts, vague/uncertain relationships)
- Computationally expensive for large real-world ontologies
- Lacks a broader industrial experience and acceptance

Continued

III. Remarks on some knowledge representation formalisms

1. Description Logic

DL is aimed at representing and reasoning with **definitorial** (analytic) knowledge (*with truth determined only by the definition of the terms*).

Examples:

Parent := Human and atLeast(1, Children)

Blackbird :< Bird

DL lacks proper and efficient handling of **synthetic** knowledge (*with truth based on known or experienced relations between instances covered by the terms*)

Examples (\rightarrow reads "implies") :

Man \rightarrow Mortal (certain)

Smoke \rightarrow Fire (uncertain)

Continued

III. Remarks on some knowledge representation formalisms

1. Description Logic

The examples modelled in FSTN:

- Parent := Human and atLeast(1, Children)

$$\begin{array}{l} \text{Parent} \xrightarrow{1} \text{Human} \\ \text{and} \quad \text{Parent} \xrightarrow{1} \text{atLeast}(1, \text{Children}) \end{array}$$

- Blackbird <: Bird Blackbird $\xrightarrow{1}$ Bird
- Man \rightarrow Mortal Man $\xrightarrow{1}$ Mortal
- Smoke \rightarrow Fire Smoke $\xrightarrow{0.8}$ Fire

Inference models, example:

$$\text{From } c_i \xrightarrow{s_{ij}} c_j \xrightarrow{s_{jk}} c_k \quad \text{infer } c_i \xrightarrow{s_{ik}} c_k, \quad s_{ik} = s_{ij} * s_{jk}$$

Proposals for extending DL and other KR formalisms for handling uncertainty, see, e.g.:

Henrik Legind Larsen and Mai Gehrke (Eds.): Special Section on Concept Oriented Knowledge Representation under Uncertainty, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **14**(1): 1-59 (2006).

III. Remarks on some knowledge representation formalisms

2. Fuzzy Concept Algebra

Introduced in:

Larsen, H.L , Nilsson, J.F.: Fuzzy Querying in a Concept Object Algebraic Datamodel. In Christiansen, H., Andreasen, T., Larsen, H.L., Eds., *Flexible Query Answering Systems* , Kluwer Academic Publishers, 1997.

Fuzzy Concept Algebra (FCA) is an extension of the Concept Object Algebra introduced by J. Fischer Nilsson.

Based on three algebraic operators: \cdot (attribution), \times (product), $+$ (sum)

Knowledge (concepts and roles), data (instances), and queries with importance weighting are handled in the FCA framework.

Continued

III. Remarks on some knowledge representation formalisms

2. Fuzzy Concept Algebra

Example:

That the type of home 'flat' also is satisfied to some degree by 'terraced house' and 'bungalow' may in FCA be represented by:

$$\begin{aligned} \text{homeType(Flat)} = & 1/\text{homeType(Flat)} \\ & + 0.6/\text{homeType(TerracedHouse)} \\ & + 0.2/\text{homeType(Bungalow)} \end{aligned}$$

represented in a FSTN for homeType:

$$\begin{aligned} \text{TerracedHouse} & \xrightarrow{0.6} \text{Flat} \\ \text{Bungalow} & \xrightarrow{0.2} \text{Flat} \end{aligned}$$

IV. Utilization of ontologies in information retrieval and query answering

1. Recognition of query concepts in document text

$Q: t_1 t_2 \cdots t_m$ interpreted as $t_1 \tilde{\wedge} t_2 \tilde{\wedge} \cdots \tilde{\wedge} t_m$ ($\tilde{\wedge}$ is soft AND)

Each query term t_i is expanded into a soft OR of terms that associates to it:

$$E(t_i) = (t_i, 1) \tilde{\vee} (t_{i1}, s_{i1}) \tilde{\vee} \cdots \tilde{\vee} (t_{in_i}, s_{in_i})$$

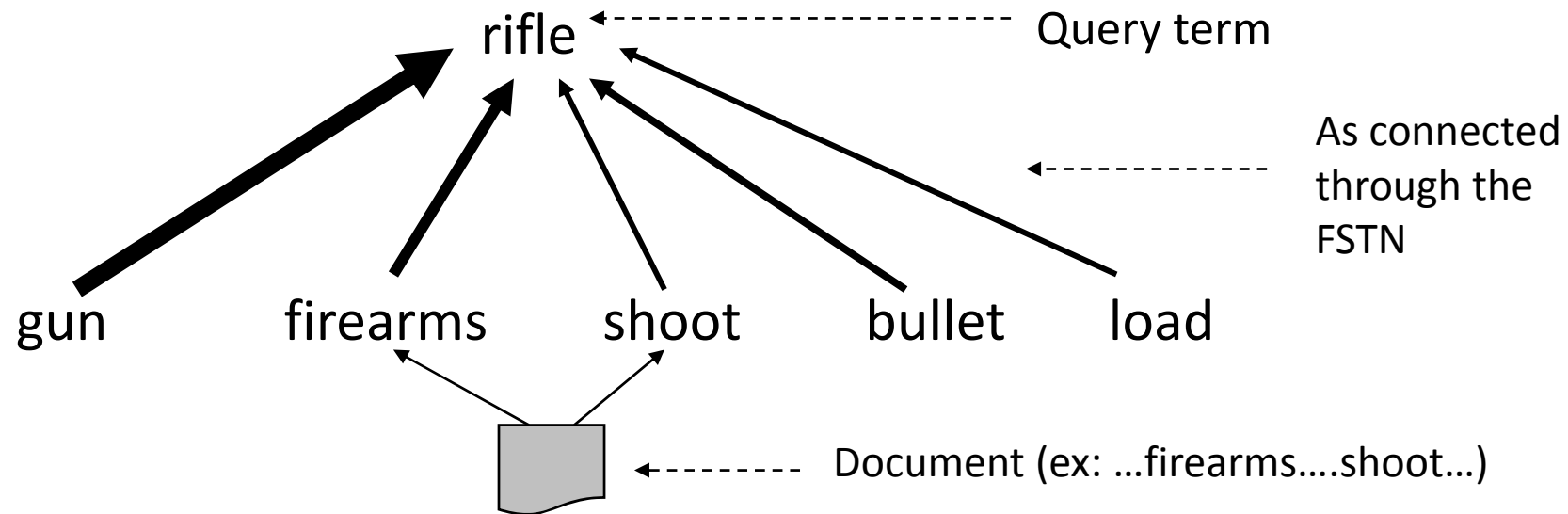
where s_{ij} is the strength of the association of term t_{ij} to term t_i

Continued

IV. Utilization of ontologies in information retrieval and query answering

1. Recognition of query concepts in document text

Illustrative example:



Fat arrow: strong indicator of 'rifle'; thinner arrow: weaker indication of 'rifle'

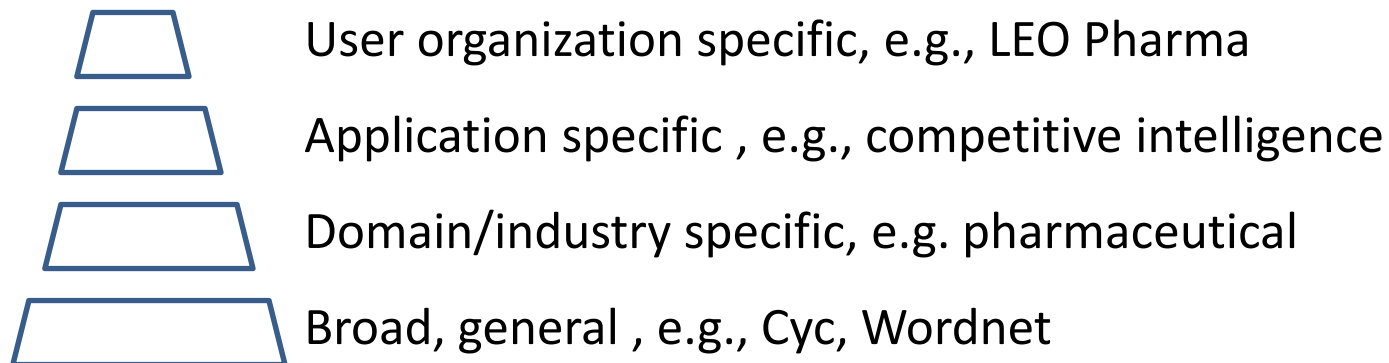
Multiple indicators (ex: 'firearms' and 'shoot') reinforces recognition

Continued

IV. Utilization of ontologies in information retrieval and query answering

1. Recognition of query concepts in document text

Preference layers of ontologies (for enterprise search engines):



If a term is defined at more than one level, the definition at the most specific level is most likely to apply for users in the organization.

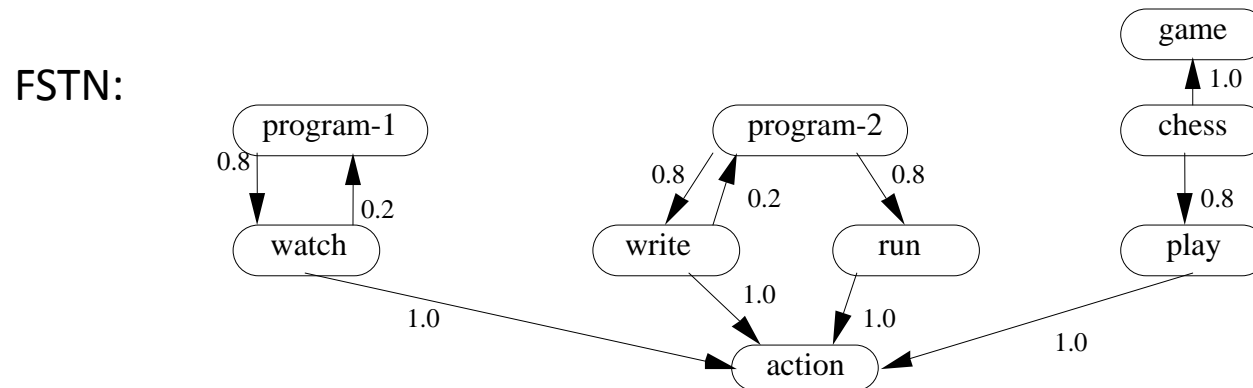
Hence, a query term is first looked up in the most specialized ontology; if not found here, we try the next level, and so on, until the term is found.

Different specific ontologies may be integrated at the same layer.

IV. Utilization of ontologies in information retrieval and query answering

2. Heuristic semantic disambiguation of query terms

Utilizes the proximity of (entity) terms in the FSTN and the context of the terms in the query.



Query: "I want to watch a program on chess"

Semantic ambiguity of 'program':

- program-1: television program
- program-2: computer program

Continued

IV. Utilization of ontologies in information retrieval and query answering

2. Heuristic semantic disambiguation of query terms

Heuristic criterion:

”Natural” query: few concepts, each described by a set of semantically related terms.

Procedure:

Evaluate the possible interpretations of the query, i.e., in this case:

- 1) watch, program-1, chess
- 2) watch, program-2, chess

For each interpretation, cluster the terms in the semantic proximity space induced by the FSTN.

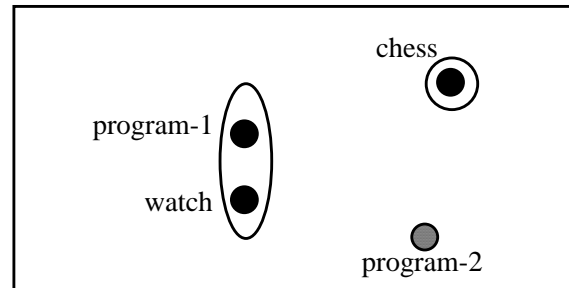
The interpretation whose clustering of terms best satisfies the heuristic Criterion, i.e., **few clusters, each containing only semantically closely related terms**, provides the most likely intended semantics.

Continued

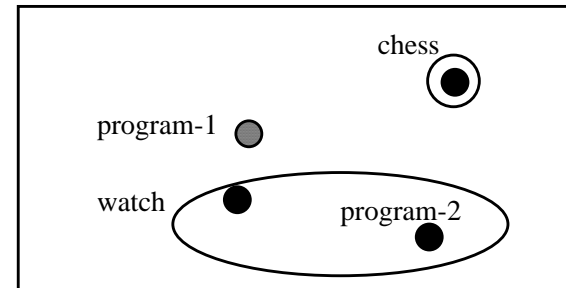
IV. Utilization of ontologies in information retrieval and query answering

2. Heuristic semantic disambiguation of query terms

Overall Best clustering



Example of worse clustering



Same number of clusters, but the terms are less close in one cluster

Key papers on FSTNs and their utilization in information retrieval and disambiguation:

- Larsen, H.L., and Yager, R.R.: The use of fuzzy relational thesauri for classificatory problem solving in information retrieval and expert systems. *IEEE J. on System, Man, and Cybernetics* **23**(1):31–41, 1993.
- Larsen, H.L.: Recognition of implicit concepts in unstructured queries. *Proc. VI IFSA World Congress, Sao Paulo, Brazil, 1995, Vol. 1*, pp. 233–236.
- Larsen, H.L., and Yager, R.R.: Query Fuzzification for Internet Information retrieval. In Dubois, D., Prade H., and Yager, R.R., Eds., *Fuzzy Set methods in Information Engineering: A Guided Tour of Applications*, John Wiley & Sons, pp. 291–310, 1997.

V. Utilization of ontologies for semantic disambiguation in "connecting the dots"

- "Connecting the dots" is central in investigative search.
- It comprises finding connections between seemingly unconnected pieces of information, in different documents independent of the source.
- Basically, two pieces of information are connected, if they deal with the same entity.
- Connections may be indirect, through a chain of documents
- **Key problem: *semantic disambiguation (of entity representations) in unstructured text***

Continued

V. Utilization of ontologies for semantic disambiguation in "connecting the dots"

- Two kinds of entities:
 1. Concept entities
 2. Instance entities
- Semantic disambiguation (or identity resolution, ambiguity resolution): Which entity does a representation refer to?
- Semantic disambiguation utilizes:
 - Ontologies
 - Entity databases (ex: Central Business Register)
 - Knowledge on constraints, e.g. cardinality (ex: a company can only have one CEO at a given time)(may be defined in the ontology)
 - Belief evaluation (source reliability, evidence for the identification,..)
 - Relevant context (in the text, in the user's task and domain, ...)

Continued

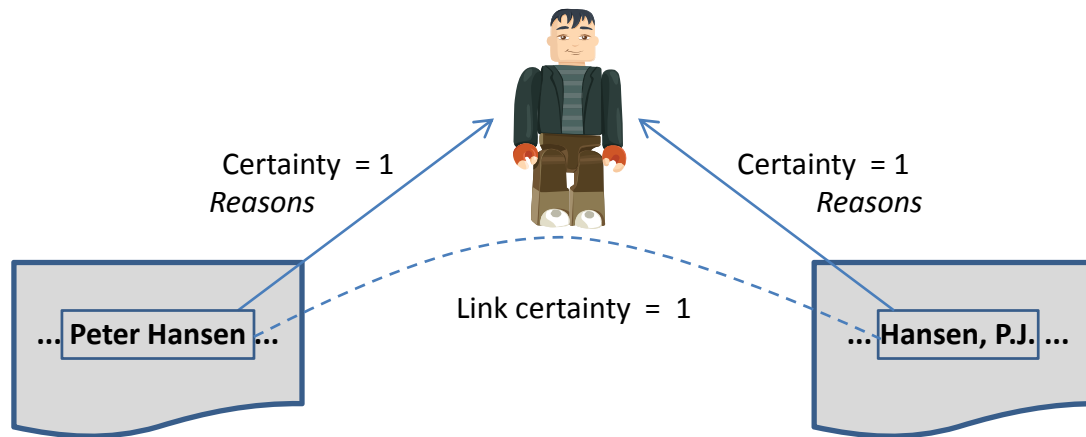
V. Utilization of ontologies for semantic disambiguation in "connecting the dots"

Causes of problems in entity identification:

- Inconsistent information
- Missing information
- More or less reliably source
- Spelling and typing errors
- Misinformation, intend or unintended
- Ambiguous (imprecise, under specified, ...) information: the same representation (name, value) refers, per se, to different entities
- Non-normalized data: different representations refer to the same entity; example:
 - Concept: car; automobile
 - Instance (e.g., person): Peter Hansen; P. Hansen; Hansen, P.; Social security No. 123456789

Continued

V. Utilization of ontologies for semantic disambiguation in "connecting the dots"



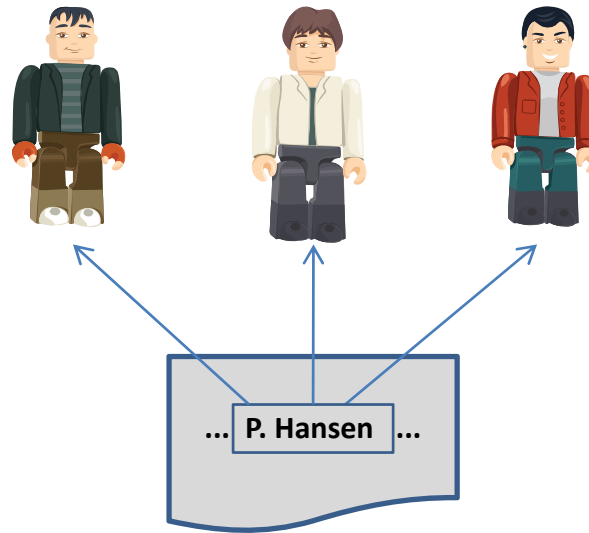
Certain unique identification, no ambiguity

The two documents are linked by the common reference to the same entity.

Reasons (for the certainty of the identification) are stored for documentation purpose.

Continued

V. Utilization of ontologies for semantic disambiguation in "connecting the dots"



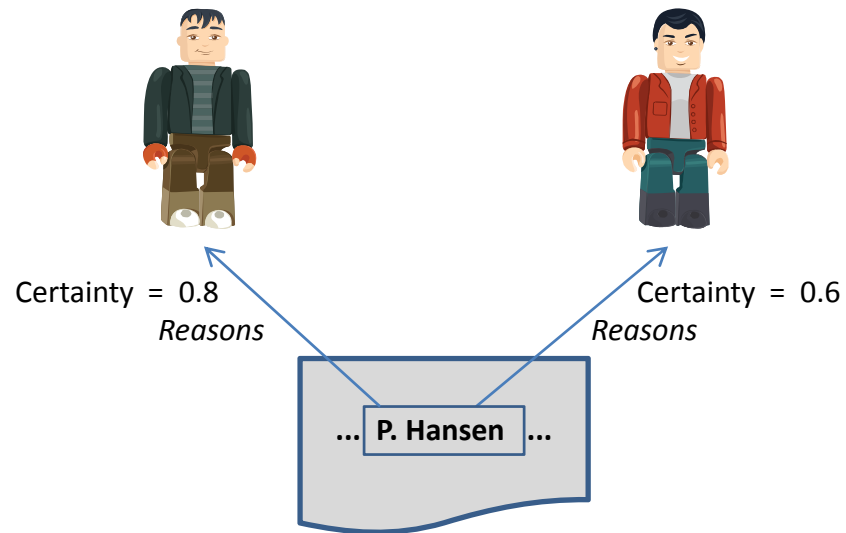
Ambiguity

Attempt to resolve using additional knowledge
(ontologies, entity databases, relevant context, etc.)

Continued

V. Utilization of ontologies for semantic disambiguation in "connecting the dots"

Assume that a unique identification, is not possible

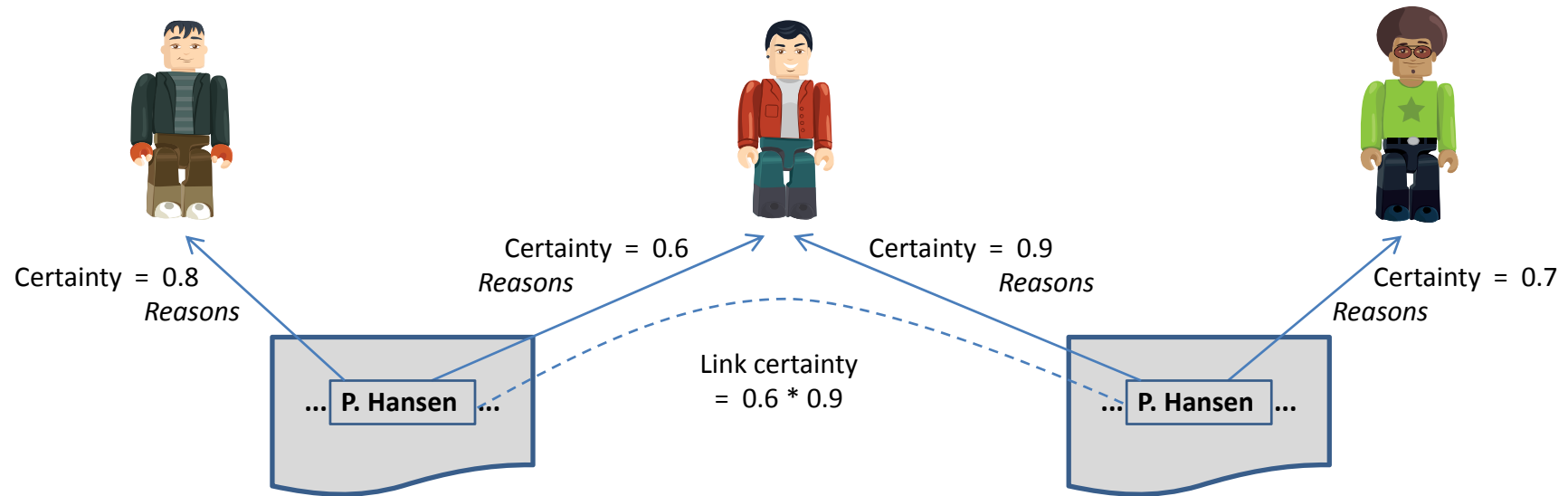


Ambiguity remaining

Both possibilities are stored, with their certainty and reasons.
The user can verify the reasons and make a decision
(user decisions are documented under reasons)

Continued

V. Utilization of ontologies for semantic disambiguation in "connecting the dots"



Certainty of possible link

A lower threshold for the certainty may be applied.

Maintaining possible links gives a high recall: No potential interesting link is missed, which is important in investigative use.

Formal frameworks: Fuzzy sets, possibility theory, and probability theory

VI. Conclusion

- Large potentials for utilizing ontologies in information access:
 - Information retrieval
 - Question answering
 - Investigative search (“connecting the dots”, etc.)
- Term relationships may be extracted from ontologies and other sources for efficient utilization through FSTNs
- Preference layers of ontologies are, in particular, useful for enterprise search engines (DanNet will fit into the broad NL ontology; *creation and test of a domain/industry specific extension of DanNet may be interesting*)
- “Connecting the dots” is central in Investigative Business Intelligence, where the key issues are:
 - semantic entity disambiguation, utilizing ontologies and relevant context;
 - if a unique identification is not possible, maintain possible identifications;
 - maintain reasons for documentary purpose