Faculty of Humanities

DanNet – Teaching and Research Perspectives at CST

Patrizia Paggio

Centre for Language Technology
University of Copenhagen
paggio@hum.ku.dk

Outline

Previous and current research:

- Concept-based search: results from research projects and student papers

Future plans:

- Combining sense tagging with shallow processing
- Extraction of frame semantic structures from sense-tagged text
- Enriching anaphora resolution algorithms with semantic relations

Information Search: OntoQuery

Our interest for content-based search goes back to the OntoQuery project* (Andreasen, Jensen, Nilsson, Paggio, Pedersen and Thomsen 2004):

Ontology-based search methods

Domain concepts + general ontology from SIMPLE (Lenci et al. 2000), appr. 10,000 concepts

Domain of application: nutrition text from Danish on-line encyclopaedia

* Univ. Roskilde, Univ. Southern DK, Technical Univ. of DK., Univ. of Cph, CBS.

Information Search: OntoQuery (2)

Both queries and documents undergo:

- pos-tagging
- lemmatising
- chunking
- mapping from NPs onto concepts

Matching is between concepts

Results are weighted according to concepts' proximity in the ontology, e.g.:

sygdom (illness) matches **pellagra** by 0.8: second-level concept in the ontology.

Information Search: OntoQuery (3)

Compounds and nominal phrases are mapped onto (complex) concepts:

"vitaminmangel" and "mangel på vitamin" (vitamin deficiency) both mapped onto the descriptor **lack[WRT: vitamin]**

"tykke børn" (fat kids) and "børn med overvægt" (kids with obesity) both mapped onto **child[CHR: obesity]**

Assumption:

In domain-specific applications, the recall provided by keyword-based search is too low.

Information Search: VID*

Content-based search on a company's intranet.

Documents: patent descriptions.

Domain-ontology created semi-automatically.

Query-expansion:

- morphologically-related word forms ("patent" → "patenter")
- synonyms and hypernyms
- terms related by weaker semantic relations (near synonymy)
- semantically-related compounds ("patent" → "patentansøgning")

* Ankiro, Zacco A/S, Univ. of Cph

Information Search: VID (2)

Evaluation (Pedersen, Paggio and Navarretta 2007):

Results	Precision	Recall	F-score
Keywords-based search	99.7	54.2	76.9
Morphological expansion	98.8	79.2	89
Morphological and semantic expansion	89.2	95.1	92.5

Information Search: teaching

Elective course on Information Search

The students are introduced to standard IR methods.

Document indexing based on:

- stemming/lemmatising
- stop words lists
- normalisation (capital letters, mwu's)
- frequency-based schemes (tf*idf)

They also experiment with content-based query-expansion using DanNet.

Information Search: teaching

Example from nutrition domain

Queries:

brød (bread)
agurk (cucumber)
fødevarer børn (food children)
grøntsager (vegetables)

Access to a fragment of DanNet:

rdfs:label="**{føde,1_1; kost,2_1; mad_1; æde,1_1_1; ædelse_1}**"> (food)

Appr. 1500 synsets

Information Search: teaching

DanNet gives access to:

- Hyponyms: **brød** (bread) ← **boller** (rolls)
- Hypernyms: **agurk** (cucumber) → **grøntsager** (vegetables)
- Synonyms: **fødevarer** ↔ **madvarer** (foods)

Results:

Use of synonyms gives best results (f2-score: 0.73)

However, the students created their own gold standard.

Building on top of shallow processing

We have developed several tools for shallow processing of Danish:

- **tokeniser**
- **name recogniser**
- **POS-tagger**
- **lemmatiser**
- **word splitter**
- **NP-recogniser**
- **repetitiveness checker**
- **n-gram frequencies**
- **keywords**
- **multiple word terms**

Goal: to add word sense tagging, and create a semantic gold standard with the Danish PAROLE corpus (300,000 tokens) as starting point.

Word sense tagging

Granularity of sense distinction

- Semantic distinctions are often too fine-grained even for human annotators (Edmonds and Kilgariff, 1998).
- Senses can be clustered by mapping the WordNet inventory against a list of senses from another lexical source (Navigli, 2006)
- Better inter-annotator agreement (from 67-73% to 86-93%) achieved with the obtained coarse distinctions (SemEval 2007, task 7)

Word sense tagging

Granularity: ontological type

Synsets in DanNet are ontologically-typed. We expect this information to be useful for clustering.

Example:

nål (needle)

Important distinction between Plant-part and Instrument.

Less important to distinguish different types of sewing needle.

Development of Semantic Frames for Danish

Semantic frames as defined in the FrameNet project.

Frames account for the range of semantic and syntactic combinatory possibilities (valences) of each word in each of its senses.

We want to focus on words related to physical change, cognition and communication, spanning from physical to mental properties.

The three domains group appr. 2000 verb senses in DanNet.

Automatic Extraction of Semantic Frames

A task at SemEval 2007:

Automatically label frame-evoking words with appropriate frames (similar to WSD) and their dependents with frame element names.

Best result (f-score 60-68%) obtained by Johansson and Nugues (2007):

- dependency parser detects potential frames
- SVM-base classifier trained on FrameNet resource assigns FE
- lexical coverage is extended by adding to the FrameNet lemmas related words from WordNet

DanNet and anaphora resolution

Bridge anaphora, e.g. hidden discourse entities:

When I got into the room I saw a strange screen saver on the big monitor. The other computer was off.

Computer has-as-part **Monitor**

(Christea et al, 2000)

Hun læste en fransk roman højt for den gamle fru Fiorenzo …
den gamle tante der lod som_om hun lyttede til oplæsningen men ikke forstod et kuk; (Pirandello: La Buon Anima)

Læse op sub-type-of **Læse**

(Navarretta, in press)

The missing referent can be reconstructed by means of relations in DanNet.