Statistical term ontology exploration in the R statistical analysis software

an adventure with ambiguous text:
statistical categorization for words, terms and concepts

Seppo Nyrkkö

2nd MOLTO project meeting
March 2011

Contents

- Background
- Tools and Data of Experiment
- Demo
  - Handling RDF in R
  - A work-flow
- Milestones Ahead
- Impact on MOLTO

1) Background

Motivation: some ambiguous terms in text (from wikipedia:)

"Five **species** of **Plasmodium** can infect humans"
"The **disease** results from the multiplication of **malaria parasites"**

This is analogous to a common problem with hand-written term ontologies: Parallel Term Matching (Merging)

- by hand, it is easy but time-consuming
- automatically, it produces gibberish

The Approach: Domain-specific text analyzed with multiple term ontologies: Findings on co-occurrences within text will support aligning the terms together

2) Tools and Data of Experiment

Tools:

- a language for statistical computing (R)
- a syntactic parser (Stanford)
- an ontology reading interface (Jena)

Development Data

- Ontologies:
  - PULS - medical / disease term knowledge ontology
  - TAP - knowledge base for generic content annotation
  - SUMO - an upper model ontology
- Corpus: Wikipedia: Malaria, Otitis, Cat scratch fever

The aim:

- o **a distance model: {similarity, coverage} measure**
- o **a work-flow model for term work**

Demo #1: Handling RDF in R

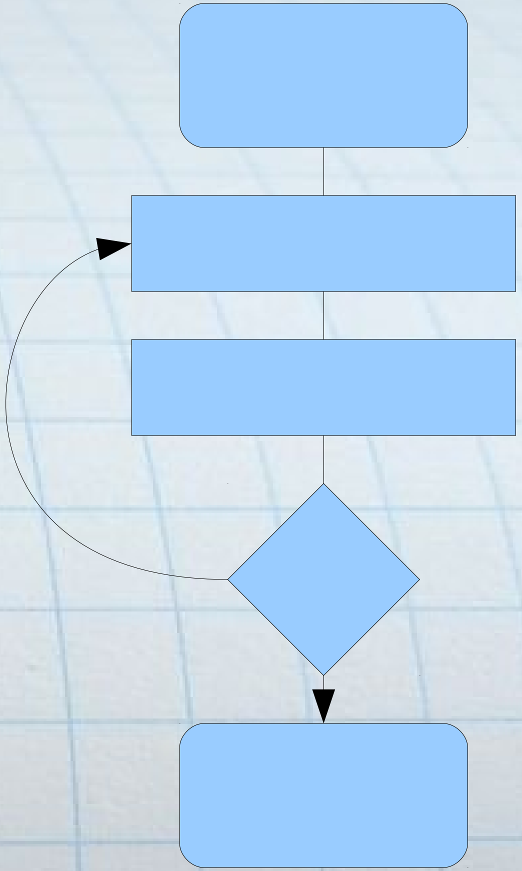
Short synopsis

```
o = ontoread(file, language="RDF/XML")
propsOf(o, uri)
propsTo(o, uri)
```

```
# statements mapped as property vectors

proptbl(onto,propURI) -> property table px
image(px,URI) -> get O by S
domain(px,URI) -> get S by O
```

Demo #2: a workflow



- import ontologies & corpora

- quick term analysis (w/ default weigths)

- corpus scan using key terms

- fine tune the weights

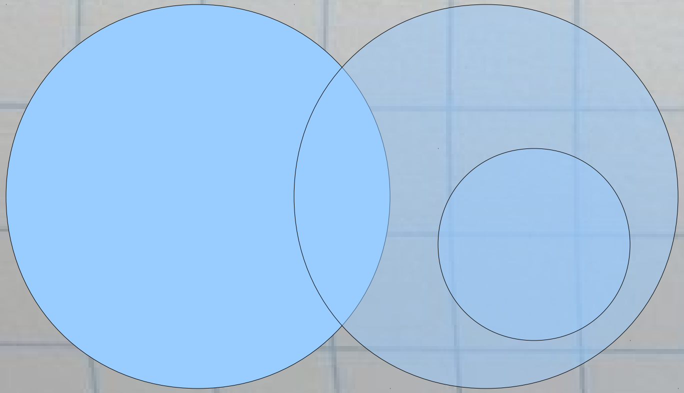
- see the results, create corrected alignments

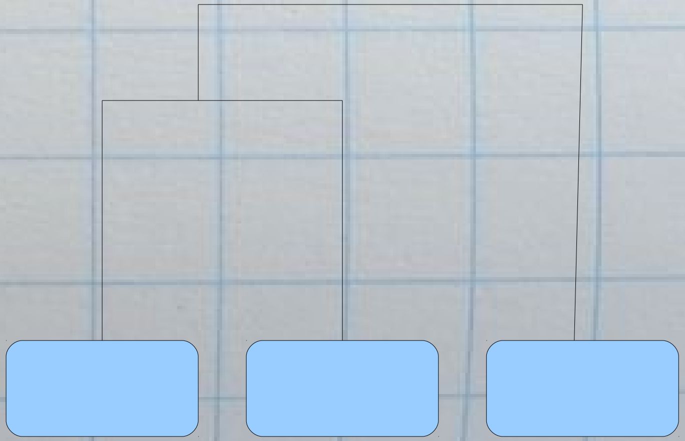
- continue building merged domain lexicon, creating syntax, etc ...

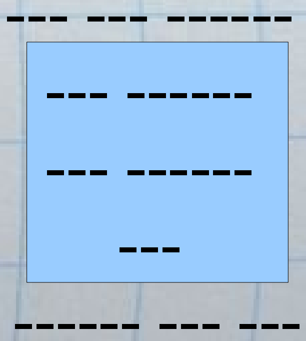
Demo #2: a workflow...

Expected outcomes of analysis:

- Super / subclass relations
- Similarity and coverage estimates
- Seeing usage of terms







3) Milestones Ahead

Some Refining needed...

- Fuzzy matching for typos and accidents in URI term names
- Syntactic analysis on descriptions in rdfs:labels
- Treating word compounds as syntactic branches

goals

1. ==> similarity + coverage metrics
2. ==> workflow for ontology validation / filling / merging

4) Impact on MOLTO

Expected vocabulary improvement

- Using multiple ontologies as MT term resources
- vocabulary to ^2

MOLTO-based systems development

- example: quickly modeling and extending multilingual dialogue systems with imported term ontologies

MOLTO-driven term ontology development

- syntactic pattern-based term ontology harvesting from text corpora
- Ontology validation by natural language generation

Thank you

Seppo . Nyrkko (at) Helsinki . fi