How Co-Occurrence can Complement Semantics?

Atanas Kiryakov & Borislav Popov
ISWC 2006, Athens, GA

9 Nov, 2006
How Co-occurrence can Complement Semantics
Semantic Annotation: How and Why?

• Information extraction (text-mining) for annotation
• Massive world knowledge is complementary
• Ontology population – extraction of structured data
• One needs a scalable semantic repository; OWLIM came to existence
• Semantic indexing and retrieval; match a query like:

  Find documents about a telecom company in Europe, John Smith, and a date in the first half of 2002.

with a document, containing:

  “At its meeting on the 10th of May, the board of Vodafone appointed John G. Smith as CTO”
How Co-occurrence can Complement Semantics
Semantic Features

- **Entities**: NE + "key phrases" (extracted through TF/IDF)
- Entities form a **reduced dimension feature space**
  - Documents are characterized by the occurring entities
  - It can still be extended to the full-dimension FTS
  - But it is interesting what these semantic features are good for
- **Documents are considered contexts**
  - Document sets (corpora) represent compound contexts
- **Occurrence indicates association** between entity and context
  - It can also be considered as "popularity" in this context
- **Co-occurrence indicates associative relationship** between entities
  - The exact relation type might not be known, but there is a link
Ranking and Timelines

• Entities can be **ranked by popularity** in a context

• The dates of the documents are used to provide **temporal dimension** to the context space
  
  – Suppose a corpus is partitioned into equal time-intervals
  
  – Popularity/association is measured in each of the partitions

• This allows for **timeline analysis**:
  
  – Popularity trends (of specific entity in specific context)
  
  – Proximity trends
Where is the Semantics?

• As a start – it is all based on semantic annotations

• One can combine co-occurrence with Related Concepts
  – Sort of “semantic closure” of the set of co-occurring entities
  – One can make “statistical closure” of the related concepts, also
KI M Platform

• KIM is a popular semantic annotation platform
  – **Automatic annotation** based on Information Extraction
  – Indexing and **retrieval**
  – **Hyper-linking**
  – **Semantic queries**
CORE Module in KIM

• CORE stands for “Co-occurrence and Ranking of Entities”
• CORE provides:
  - Tracking occurrences of entities in documents
  - Efficient querying for co-occurrences under various restrictions
  - Ranking of entities based on their “popularity”, i.e. frequency of occurrence
• There are also two Web UIs, based on it:
  - CORE Search performs interactive faceted search
  - The Timelines tool computes and draws popularity timelines …
How Co-occurrence can Complement Semantics
Demo with 1 Million Documents

• CORE Demonstration:
  » 1 million documents
  » International News Articles (2002-2006)
  » Approx. 1000 articles per business day

• Statistics
  » More than 1 million entities (50K pre-populated)
  » Described in about 10 million RDFS/OWL triples
  » On average, 30 entities occurring per document
  » Number of occurrences: 27 M
Different sorts of \textit{normalizations of the ranks} are considered

\textbf{Entity profiles}
- The 10 entities most strongly associated with a specific one
- Derive associative relationships

\textbf{Document/ Context profiles} – we have it
- The entities which occur in it
- A feature vector in the entity feature-space

\textbf{Identity resolution} via matching document and entity profiles
- It may also work for database integration (record-linkage)
IST World

- IST World is a portal for IT research
  - Experts, Organizations, Projects, Publications
- Social-networking and research trends analysis
  - Spectacular “research intelligence” tools
  - It can help you find FP7 partners
- Most of it based on comprehensive statistical text-mining from JSI
- How we use CORE here?
Applications

- **News aggregation** and analysis
- **Media Research**: identification and tracking of campaigns
- **Digital libraries** of technical documents and clinical studies
- Management of **digital audio-visual archives**
- **Opinion mining** from Web forums and blogs
- **Social Networks** extraction
- **Job Intelligence** – automatic extraction of job vacancies
  - JOCI - [http://www.innovantage.co.uk](http://www.innovantage.co.uk)
  - Extracting about 100k jobs from 30k UK organizations websites
JOCI: Recruitment Intelligence for UK

UK Web Space

Focused Crawler
- Crawler
- Classifier

Information Extraction
- Single-Document IE
- Object Consolidation

KI M Server
- Semantic Repository
- Document Store

Web User Interface

How Co-occurrence can Complement Semantics
9 Nov, 2006
How Co-occurrence can Complement Semantics
Thanks!

Give KIM a try

http://www.ontotext.com/kim/