Ontology-based Text Processing for Personal Information Management

25.04.2008
Guest lecture at Universitat Autònoma de Barcelona

Alexander Troussov, Ph.D.
IBM Dublin Center for Advanced Studies Chief Scientist

John Judge, Ph.D.
IBM LanguageWare

http://nepomuk.semanticdesktop.org
Demo in this talk use Galaxy Library, created in IBM Dublin:

http://www.alphaworks.ibm.com/tech/galaxy
GALAXY makes sense out of data provided by networks.
Galaxy - makes sense out of data provided by networks

- Galaxy as a product for semantic text analysis:
  - Reads the text
  - Memorises all the concepts
  - Uses networks of words to analyse which concepts sit well together

- Text which mentions *Mulhuddart, Lansdowne, Clontarf* is probably about *Dublin / Ireland / Europe*
Galaxy - makes sense out of data provided by networks

- Galaxy as a tool for building solutions for socio-semantic computing
  - Analyses networks of words, people, documents, Lotus Connections activities, tags, blogs etc
- In a corporate environment, we are all connected to other people through various links. Using these complicated networks, Galaxy help people, for instance, to answer questions like

"Who are the people with whom I share the most interests"?
Agenda

- **Context and Motivation**
  - Nepomuk project
  - What are the specifics of this type of ontology-based text processing

- **Technical organisation of lexico-semantic resources into lexically enriched ontology**

- **Linguistic Light Lexical Analyser**
  - How to link polysemous and ambiguous words with concepts and how to find concept mentions in free texts
  - Lexical layer for ontologies, LLA/LLS

- **Analytics on term mentions**
  - Building semantic models of texts and using this for tagging (keyword extraction) and for term disambiguation

- **Conclusion, Future work, Cooperation**
Context and Motivation of our Approach – The Needs of Nepomuk’s Social Semantic Desktop
What Nepomuk is about?

- **Problems with personal data**
  - How can I remember that this MS Word document is a draft reply to this email in IBM Lotus Notes email client?
  - How to remember who is working with whom, participates in which projects?
  - How to find a particular email?
  - How can I memorise the names of Claudia’s cats?

- **Solution is Social Semantic Desktop**

- **Based on semantic web technologies like PIMO**
What is semantic text processing?

- **EU 6th framework integrated project Nepomuk aims to build social semantic desktop (based on Semantic Web standards)**

- **If you'll start to use this desktop, you'll face two problems:**
  - First of all, to benefit from semantic web technologies you need to annotate your documents and emails with information as to what they are about.
  - Secondly, even if you do this tedious manual work, later you still might not be able to easily find what is needed.

- **For instance, you might annotate an email:**
  - *This email is about document of understanding*
  - *with a company operating in California.*
  - When later you'll look for all your legal documents related to *the United States*, this email will not be retrieved because parameters of your semantic search do not exactly match the semantic annotation of the email.

- **Galaxy library (IBM’s Contribution to Nepomuk) is a solution for these problems of semantic desktop consumability**
  - And lightweight scalable API to build solutions for socio-semantic processing
Applications to metadata generation in Nepomuk

• Currently the semantic web relies on semantic annotation mostly done manually by humans
  • Tedious, time consuming, and error prone work

• We in IBM Dublin developed a tool for automation of metadata creation, which can use PIMO of the user of semantic desktop:
  • Automatic ontology-based conceptual tagging
    (central concepts of the text with respect to the given lexico-semantic resource)
  • Text which mentions *Mulhuddart, Lansdowne, Clontarf* is probably about Dublin/Ireland/Europe/Earth; This fact can be inferred from geographical relations like *Mulhuddart* "is-part-of" *Dublin*
  • Disambiguation of terms
    • Based on the ontological knowledge from corresponding resource
Galaxy as a product for semantic text analysis:

- Reads the text
- Memorises all the concepts
- Uses networks of words to analyse which concepts sit well together
- Text which mentions *Mulhuddart, Lansdowne, Clontarf* is probably about *Dublin / Ireland / Europe*
Galaxy at work: automatic tagging based on concept mentions

Finding “focus” concept

Mapping of term mentions to concepts

NETWORK OF CONCEPTS

TEXT

Mention  Mention  Mention  Mention
Technical Organisation of Lexico-Semantic Resources into Lexically Enriched Ontology
What is an ontology?

• Our pragmatic approach for representation of lexico-semantic resources is in vein with the most cited (in Computer Science, and in the field of Artificial Intelligence) definition of ontology “An ontology is an explicit specification of a conceptualization.” (Gruber, 1993) with Guarino’s clarifications: “... engineering artefact, constituted by a specific vocabulary used to describe a certain reality... ".

• However, our goal is to design the means by which various types of knowledge can be organized in a computer system and applied to text processing for semantic annotation and IR applications
  • It is an established fact that knowledge plays a vital role in the comprehension and production of discourse. The process of interpreting words, sentences, and the whole discourse involves an enormous amount of knowledge which forms our background and contextual awareness. However, how various types of knowledge can be organized in a computer system and applied to the comprehension process is still a major challenge for semantic interpretation systems used in natural language processing.

  Weiqiang Ou, Adel Elsayed, and Roger Hartley
Aspects of our technical organisation of lexico-semantic resources

- **Provides conceptualisation**
  - In our approach concepts are organised as a semantic network (however, in contrast to the standard definition of semantic networks, links are not only typed, but might have weights)

- **Is Formal**
  - Knowledge and linguistic expressions representation has a formal software specification and is suitable for text processing + reasoning how concepts mentioned in a text sit together

- **Explicit**
  - Means that all types of primitives, concepts, and constraints used in the ontology specification are explicitly defined.
Aspects of our technical organisation of lexico-semantic resources (cont.)

- Shareable,
  - not necessarily shared
- **Not strictly a taxonomy**
  - with its only class-subclass hierarchy
- **Not strictly a thesaurus**
  - there are no (semantic) relations between lexical expressions
- **We focus on how to organise resource for text processing, in order to recognise mentions, or potential relevance, of concepts to text**
  - not on how to assign labels - which are strings, not linguistic units
  - not on concept descriptions - which is for human (not machine) consumption
Layered organisation of lexico-semantic knowledge for automatic text processing

Lexicon

Layered organisation

Free N x M mapping
Layered organisation

- Key principles of how we organise lexico-semantic knowledge into a lexically enriched ontology for automatic text processing:
  - Full separation of two layers: lexical entries and concepts
  - No distinctions between conceptual layer and instances layer
    - which are merged into semantic network
  - Labels != lexical entries
  - Lexical entries – any names of concepts, terms, MWU, ...
  - Free $N \times M$ mappings between lexical expressions and concepts
... and processing resources

- Lexicon
- Semantic Network
- Mapping
- Graph Mining
... and processing resources

- **Lexicon**
  - Used by lexical analyser to find mentions of concepts represented by nodes in the semantic network.

- **Semantic Network**
  - Used by some automated reasoners and miners which exploit the graph-theoretic features of the network during processing.

- **Mapping**
  - Mapping from text to concepts creates semantic model of a text (as a function on nodes of the network which shows how concepts are related to text).

- **Graph Mining**
  - Provides analytics on term mentions.
Why lexical enrichment?

• To improve recall in recognition of concept relevancy to texts
  • To detect that the text is about *Linguistics*, one can look for words *linguist*, *language*, *grammar*, *syntax*
  • Finance: *bank*, *currency*, *Euro*, *money*
  • Bank: *account*, *safe*, *money order*, *cheque*

• Graded levels of Salience
  • Some words are more indicative of concept then others
    • *Linguist* is more indicative for the concept *Linguistics*, than *language* and *syntax*

• Enrichment might be done exploiting machine learning techniques
Ontologies and Their Relatives

Catalogs
Thesauri
Glossaries & Terminologies
Semantic Networks
Formal isa
Formal Instance
General logical constraints
Axioms: Disjoint/Inverse...

© Paul Buitelaar: Lexical Semantics and Ontologies
Tutorial at ACL/HCSnet, July 2006, Melbourne, Australia
Knowledge encoded in semantic networks

- "Everything is deeply intertwingled. In an important sense there are no "subjects" at all; there is only all knowledge, since the cross-connections among the myriad topics of this world simply cannot be divided up neatly".
  Theodor Holm Nelson, the pioneer of information technology

- and nowadays many of this links are explicit in computer mediated communication networks (the web, semantic web, desktops data, intranets, ...), which can be studied using interdisciplinary methods

- We agree with J.R. Firth "You shall know the word by the company it keeps", then nowadays we can know the text, the person and the digital artefact by the company it keeps in the unified socio-semantic space
Knowledge encoded in semantic networks (cont.)

• If the knowledge encoded in semantic networks is “the truth”?
  • "the truth, the whole truth, and nothing but the truth"

• Probably not
  • After many years as an expert, I've become more and more uncomfortable about swearing to tell the truth, the whole truth and nothing but the truth, especially when I've looked back on cases in which two experts have said exactly the opposite and one's bound to ask which one was telling the truth? ...
    Professor Max Sussman, The Expert Witness Institute

• The truth is rarely pure and never simple.
  Oscar Wilde, Irish dramatist, novelist, & poet
Ontologies and their relatives

- Catalogs
- Thesauri
- Glossaries & Terminologies
- Semantic Networks
- Formal is-a
- General logical constraints
- Formal instance
- Axioms
- Disjoint/Inverse

Fuzzy Mathematics
Soft Computing
Network Analysis
Linguistic Light Annotation and the Linguistic Light Scanner
Motivation

• Semantic Web ontologies used in Information Extraction and Text Analytics applications
• Semantic Web community – encoding linguistic data in ontologies
  • Descriptive purposes
  • Localisation
• Lexical extensions drive OBIE, semantic annotation etc.
• Removes the need for mapping from language data to ontology
Lexical Entries

• Lexical entries for many concepts are not simple...they are often multi-word units
  • Molecular Biology
  • Lexical Functional Grammar
  • Linear Algebra
  • 11 beta-hydroxysteroid dehydrogenase
• And also have varying word order and capitalisation
  • Hydroxysteroid 11-beta Dehydrogenase
Multi-word Units

• Simple compound nouns are somewhat idiomatic
  - No great syntactic alteration
  - Simple plural inflection
• Car park + plural = Car parks
• Teórico-práctico + plural = Teórico-prácticos
Multi-word Units

• This is not the case for left headed compounds
  • *Part of speech + plural = Parts of speech*
  • *Attorney general + plural = Attorneys general*
  • *Sueldo base + plural = Sueldos base*
Listing Approach

• **Quick fix: Listing all variant forms**
  - Leads to lexical proliferation
  - Difficult to maintain

• **Not all MWUs are syntactically unalterable**
  - Marie Curie → Marie and Pierre Curie
  - Salvador Dalí → Salvador y Gala Dalí
It seems linguistics is needed

- **Linguistic theory allows us to capture these phenomena and generalise**
- **But...**
  - Which linguistic theory do we use?
  - Can it be implemented for automation?
  - Who is going to do it?
- **Knowledge engineers who work with ontology data are not linguists**
Lexical Layer Requirements

- Exploiting lexical layers in commercial TA requires
  - Quick to implement
  - Flexibility
  - Handling of multi-word and inflecting forms
  - Ease of use by non linguists
Our Solution – Linguistic Light Annotation

• Linguistic Light Annotation (LLA)

  • A set of citation forms and lightweight linguistic constraints which specify how lexical forms can vary
  • citation form: <String>
    fixed word order: <Boolean>
    intervening tokens: <Integer>
    exact capitalisation: <Boolean>
    exact string match: <Boolean>
Our Solution - Linguistic Light Scanner

- **Linguistic Light Scanner (LLS)**
  - Processes text
  - Detects “signature” of lexical entries
  - Determines if signature is a hit or not based on LLA constraints
LLS Algorithm (Simplified)

- Process text within a sliding window
- Normalise words
- Check against FSM dictionary for signature of MWU
- Signature found
  - Map back to original text
  - Check constraints
• Lexicon contains entries for

• “United Nations”
  – fixed word order: true
  – intervening tokens: 0
  – exact caps: true
  – exact string: true

• “debt consolidation,”
  – fixed word order: false
  – intervening tokens: 5
  – exact caps: false
  – exact string: false

• “… a nation united by its struggle….”
• “… consolidation of credit card debts…”
“... a nation united by its struggle....”

- Signature of “United Nations” is detected
- BUT
  - Variable word order, different capitalisation, and inflections are disallowed by constraints
- Rejected
“... consolidation of credit card debts...”

- Signature of “debt consolidation” is detected
  - Variable word order is allowed
  - Intervening tokens are allowed
  - Inflected forms are allowed
- Accepted
Advantages

• Full separation of lexical entries and concepts
• Free N x M mapping between semantic and lexical layers
• Requires little (no?) linguistic knowledge
• Handles inflected forms, varying word order and both
Experiments

• Chosen domain: medical documents
  • Many (complex) MWUs which refer to the same entities
  • Multiple citation forms
  • N x M mapping between concepts and surface forms
• MeSH (Medical Subject Headers) Ontology
• PubMed documents
  • 63,430 open-access documents for data mining
Experiments

- Compiled dictionary of MeSH concepts
- No constraints are provided in MeSH
- Assigned “default” constraints to each lexical entry
  - fixed word order: false
  - intervening tokens: 2
  - exact caps: false
  - exact string: false
- Constraints not strictly appropriate for all entries - represents a baseline
## Results

<table>
<thead>
<tr>
<th>Expression Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exprs with fixed word order and no intervening tokens</td>
<td>17,345,322</td>
</tr>
<tr>
<td>Exprs with variable word order and intervening tokens</td>
<td>17,666,135</td>
</tr>
</tbody>
</table>

### Breakdown

<table>
<thead>
<tr>
<th>Component</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable word order</td>
<td>473,557</td>
</tr>
<tr>
<td>Intervening tokens</td>
<td>672,697</td>
</tr>
<tr>
<td>Both</td>
<td>326,823</td>
</tr>
</tbody>
</table>
Results - Breakdown

• Results show that the LLA + LLS gives a 4.86% increase in detecting MWU lexical items
• Marked increase for such naïve assumptions regarding MWUs
• In a real life situation a KE provides accurate descriptions and constraints when adding data to the ontology
• We expect much improved results given correct data
Semantic Function Space Models for Representing Text Documents
Semantic function space models

• Vector Space Model
  • Traditional Vector Space Model of Information Retrieval

• Semantic Function Space Model
  • Model we introduced which covers Vector Space Models and is somewhat similar to it,
  • however, VSM is an algebraic model,
  • while Function Space Model can be studied by the methods of function analysis: find local maximums, make function more smooth, etc. involving graphmining
Graphmining technique used in Galaxy
Physical analogy of how Galaxy works

• To implement this vision we created software which makes sense out of data provided by networks. This software hinges on an algorithm inspired by the phenomena spread of activation.

• In physical analogy we replace the notion of activation by the notion degree of illumination and spread of activation by the notion of light propagation

• Each node of interest within the graph emits an amount of “light” which propagates around the graph along its links

• “Light” from multiple sources combines eventually leading to a point which is illuminated to a greater degree
  – Physical analogy allows to see that the most illuminated nodes are not necessarily those nodes which were originally chosen as light emitters, but rather the overlapping areas
Numerical simulation of spread of activation provided by Galaxy

Initial activation

Activation spread
Parameters affecting single action of activation spread

- The following parameters can be changed in configuration file of Galaxy Library:
  - base decay parameters for each link type
  - maximum distance of spread activation
  - Cut-of threshold (minimum value of weight to continue spread activation)

- The following parameters can potentially be customized by changing the source code:
  - Formula for calculating weights when spreading activation
    - base decay rates,
    - out-degree
    - different link types
Why we focus on using Spread of Activation technique

• This is from known to unknown, from seen to unseen

• "Je ne cherche pas, je trouve"
  Pablo Picasso
Summary of Galaxy Library: Components and applications

Library

- Lexico-semantic resources
- Text processing
- Graph mining
- Eclipse Workbench
- Natural Language Interface

Applications

- Term Disambiguation
- Producing meta-information
- Tag recommendation for Dogear
- Semantic Web Navigation

Solutions

- Context Smart tagging
- Semantic Search
Summary of Galaxy (cont.)

• Components of Galaxy software (beta version)
  • Resource (Dictionary) builder
    • Semantic Network + Lexical Dictionary
  • Text scanning (Mapping from text to concepts)
  • Graph mining on ontological network (our particular version of highly customised spread of activation)
  • Eclipse workbench
    • With full interface for text processing (semantic tagging, term disambiguation)
    • Prototype of natural language interface to work with social semantic networks
IBM LanguageWare Miner for Multidimensional Socio-Semantic Networks

The proliferation of Web 2.0 has brought about all kinds of digital artifacts: documents, people, concepts, vocabulary, tasks, activities, and more. IBM LanguageWare Miner for Multidimensional Socio-Semantic Networks provides a unified API that helps in creating solutions for these types of multidimensional networks (people, documents, tasks, etc.) and provides an integrated platform for combining social computing, semantic processing, and activity-centered computing for enhanced user experience.

http://www.alphaworks.ibm.com/tech/galaxy
Research cooperation
Research collaboration

• If the data in your task naturally lend themselves to be represented as a network, you might think about using Galaxy
  • Because Galaxy makes sense out of data provided by networks.

• To test Galaxy
  • Convert your network into the format which can be used by Galaxy (it is simple XML format)
  • ... and you’ll get first results in a matter of days

• To get good results
  • You might consider to work with us to leverage our experience with applying Galaxy to solve various problems
Research collaboration – contacts, links

• **Contact**
  - Alexander Troussov, CAS Chief Scientist,
    atrousso@ie.ibm.com

• **Links**
  - **NEPOMUK**
  - IBM CAS Dublin
  - Lotus Connections
  - Galaxy Library
Acknowledgements

• **Acknowledgements:**

  • IBM LanguageWare team, especially Mikhail Sogrin

  • Nepomuk project partners:
    • Brian Davis, Siegfried Handschue (NUIG)
    • Alexander Polonsky and Mikhail Kotelnikov (Cognium Systems)
    • Ying Du, Ernie Ong and Uwe Riss (SAP Research Team)
    • Leo Sauerman and Gunnar Grimnes (DFKI)
    • Heiko Haller and Max Voelkel (FZI)
Thank you!