

4A Framework

Annotations Anywhere, Annotations Anytime

Jaroslav Dytrych

Natural Language Processing Group
Faculty of Information Technology
Brno University of Technology

April 28, 2011

Supervisor: Doc. RNDr. Pavel Smrž, Ph.D.

Outline

- 1 Motivation and Objectives
 - Introduction
 - Sources of inspiration
 - Goals of work
- 2 4A Framework Infrastructure
 - Annotation format
 - Exchange protocol
- 3 Proof-of-Concept Implementation
 - 4A Server
 - 4A Clients
- 4 Future Work

Outline

- 1 Motivation and Objectives
 - Introduction
 - Sources of inspiration
 - Goals of work
- 2 4A Framework Infrastructure
 - Annotation format
 - Exchange protocol
- 3 Proof-of-Concept Implementation
 - 4A Server
 - 4A Clients
- 4 Future Work

Motivation

- Success of knowledge-based systems (and the Semantic Web) depends on metadata
- There is a lot of unstructured data
- Automatic extraction techniques are unreliable
- Manual annotation is tedious
- Structure of knowledge is known in a limited number of (simple) cases only
- Annotations can be used for knowledge structuring

Desiderata

- Creating complex annotations should be as simple as tagging
- Users should be able to annotate in the applications they know and use for reading and content creation
- Automatic methods should facilitate the work by suggesting annotations
- Whenever possible, annotations should be anchored in text
- Knowledge structuring should build on popularity of social tagging

Annozilla

The screenshot displays a Mozilla browser window titled "Annozilla project - Mozilla" with the address bar showing "http://www.w3.org/2001/Annozea/". The browser's sidebar on the left shows a tree view of annotations for the current page. The main content area features the W3C Annozea logo and the heading "About Annozea". A context menu is open over the annotations list, showing options like "View Annotation body", "Edit annotation body", and "Delete annotation". An "Annotation body" dialog box is overlaid on the page, containing a rich text editor and form fields for "Author" (with the value "your name here"), "Server" (set to "http://localhost/annotations"), "Type" (set to "Comment"), and "Language" (set to "English"). There are also checkboxes for "Keep this window open after the operation finishes" and buttons for "Post" and "Cancel".



Sources of inspiration

Annotating

- Anozilla, SharedCopy, Stickis ...
- PREP Editor, Bundle Editor, FAST ...
- DiLAS

Annotation exchange

- Annotea (W3C)
- API for project InsightLink (IBM)

Real-time collaboration

- Google Wave
- Novell Pulse
- Google Docs
- Microsoft Office Live

Shortcomings of existing solutions

- It is hard to share and reuse annotations
- The content being annotated cannot be edited
- There is no structuring of annotations
- Fixed set of annotation types
- Limited integration of advanced information extraction tools

Annotations Anywhere, Annotations Anytime

- Simultaneous editing and annotating in heterogeneous environments
- Working on the same text in various formats
- Real-time collaboration
- Simple and structured annotations
- Annotation suggestions and automatic update of annotations
- Documents can change

Outline

- 1 Motivation and Objectives
 - Introduction
 - Sources of inspiration
 - Goals of work
- 2 4A Framework Infrastructure
 - Annotation format
 - Exchange protocol
- 3 Proof-of-Concept Implementation
 - 4A Server
 - 4A Clients
- 4 Future Work

Annotation format

- Based on RDF
- Annotations and tags together
- Simple and structured annotations
- Relations among annotations
- Dynamic structures
- Robust positioning

Annotation format

Encoded information:

- type of annotation
- date and time of creation
- author
- URI of a server copy of the annotated document
- textual fragments (XPath, offset, length and textual content),
- content of annotation
- attributes

Attributes:

- structuring of annotations, dynamically added
- simple data types, links to annotations or nested annotations

Exchange protocol

- Based on XML
- Transport in various protocols on lower level
- Messages can be combined to make communication efficient
- Two-way real-time asynchronous communication
- Synchronization and actualization of documents
- Subscription to annotations from defined sources
- Exchange of annotations and types of annotations
- Annotation suggestions

Exchange protocol

- session management
- users and groups
- subscription to annotations
- synchronization of annotated documents
- exchange of annotation types
- exchange of annotations
- suggested annotations
- settings
- errors and warnings
- explicit confirmation

Outline

- 1 Motivation and Objectives
 - Introduction
 - Sources of inspiration
 - Goals of work
- 2 4A Framework Infrastructure
 - Annotation format
 - Exchange protocol
- 3 Proof-of-Concept Implementation
 - 4A Server
 - 4A Clients
- 4 Future Work

Proof-of-Concept Implementation

Server

- core of the system functionality
- modular design
- universal

Clients

- addons and plugins for various applications (document viewers, editors, e-mail clients, etc.)
- varying complexity

4A Server

- implemented in Java
- communication with clients over Comet (Grizzly Framework)
- ORM EclipseLink (JPA 2.0)
- MySQL database
- annotation suggestions from the KiWi project
- other modules under development

4A Clients

- specification of required and optional functionality forms a part of the framework
- implemented by several authors
- the first functional client – a plugin for TinyMCE
- addon for Mozilla Firefox being developed
- plugin for Microsoft Internet Explorer being developed

Structured annotation

The next **Technical plenary meeting** will be held 1-5 March 2004, in Cannes-Mandelieu, France. The group discussed meeting other groups face to face.

Will participate:

(PC) **Patrick Curran** (Sun Microsystems)

(KD) **Karl Dubost** (W3C, WG co-chair)

Annotation

Selection:

Type:

Content:

Attributes

Type:

Selection:

Content:

Attributes List:

- Solver
- Term
- Attendees**

Adding of attribute

The next Techni
Cannes-Mandelie
face to face.
Will participat
(PC) Patrick
(KD) Karl Dub

Annotation

Selection: meeting

Type: Task

Content: Prepare m

Attributes

- Solver
- Term
- Status

New attribute

Name: Task number

Type:

- Simple
 - Date and time
 - Numeric
 - Integer number
 - Decimal number
 - String**
 - URI
- Structured
 - Person
 - Employee
 - External collaborator
 - Customer
 - Business object

Add a new as a subtype of selected

Add

Add to the basic attributes of given annotation type
(to show others when entering a new annotation)

Required attribute

Cancel OK

Save

Visualization

Annotation by idytrych (2011-01-05)
Type: Place
Content: Place of meeting
City: Cannes 🔍

Task from idytrych (2011-05-01) 📌
Prepare materials for Patrick

Solver: Alan Wright
Term: 2011-03-01 07:30
Place: Cannes-Mandelieu 🔍

City

Annotation by idytrych (2011-01-04) 📌
Text: Cannes
Type: Thing -> Settlement -> City
Content: City in France
Location: France 🔍

The next Technical plenary meeting will be held 1-5 March 2004, in Cannes-Mandelieu, France. The group discussed meeting other groups face to face.
Will participate:
(PC) Patrick Curran (Sun Microsystems)
(KD) Karl Dubost (W3C, WG co-chair)

Annotation

Selection:

Type:

Content:

Advanced visualization

Description by idytrych (2010-12-14) 🗨️
Demonstrative dummy text

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum passages, and more recently with desktop publishing software like **Aldus PageMaker** including versions of Lorem Ipsum.

Contrary to popular belief, Lorem Ipsum is not simply random text. It has roots in a piece of classical Latin literature from **45 BC**, making it over 2000 years old. **Richard McClintock**, a Latin professor at **Hampden-Sydney College** in **Virginia**, looked up one of the more obscure Latin words, **consectetur**, from a Lorem Ipsum passage, and going through the cites of the word in classical literature, discovered the undoubtable source. Lorem Ipsum comes from sections 1.10.32 and 1.10.33 of "de Finibus Bonorum et Malorum" (The Extremes of Good and Evil) by **Cicero**, written in 45 BC. This book is a treatise on the theory of ethics, very popular during the Renaissance. The first line of Lorem Ipsum, "Lorem ipsum dolor sit amet..", comes from a line in section 1.10.32.

Adapted from <http://www.lipsum.com/>

Person - annot. by idytrych (2010-12-21) 🗨️
First name: Marcus Tullius
Surname: Cicero
e-mail: N/A



Annotation

Selection:

Type:

Browse

Content:

Save

Working solution

AnnoEditor v0.1 | Status Bar | Doc annotations | Suggest annotations | Annotate document | Connect | Create Annotation

Boosted Decision Trees for Deep Learning

About 4 years ago, I speculated that decision trees qualify as a deep learning algorithm because they can make decisions which are substantially nonlinear in the input representation. **Ping Li** has proved this correct, empirically at **UAI** by showing that **boosted decision trees** can beat **deep belief networks** on versions of **Mnist** which are artificially hardened so as to make them solvable only by deep learning algorithms.

This is an important point, because a deep learning algorithm is a deep learning algorithm because it is a boosted decision tree, they are not.

Geoff Hinton once told me that the substantial difficulty in getting

Submit | Reset

Create Annotation

Annotation

Selection: Ping Li has proved this correct, empirically at UAI by showing that boosted decision trees can beat deep belief networks on versions of Mnist which are

Type: Comparison -> ML Method Comparison | Browse

Content: Comparison of machine learning methods

Attributes

- ML Methods
 - ML Method 1
 - ML Method 2
- Best
- Comparison Author
- Blog author
- Source of comparison
- Test set
- Modified

Type: <i -> ML -> ML Method | Browse

Selection: boosted decision trees | Select

Content: Boosted Decision Trees

Save | Cancel

StatusBar

Connected. Connection was

Suggestions

The screenshot shows the AnnoEditor v0.1 interface. At the top, there is a toolbar with buttons for 'Status Bar', 'Doc annotations', 'Suggest annotations', 'Annotate document', 'Connect', and 'Create Annotation'. Below the toolbar, there are two overlapping 'Suggested Annotation' dialog boxes. The left dialog box shows a suggestion with the following details: **Type:** Science -> Mathem Algorithm -> Momentum; **Content:** Momentum met. The right dialog box shows a suggestion with the following details: **Type:** Science -> Mathematics -> Algorithm -> Conjugate Gradient; **Content:** Conjugate Gradient method. Both dialog boxes have 'Accept' and 'Decline' buttons. In the background, a document snippet is visible with the following text: 'momentum is the same as CG! Really?!?! There's tons of stuff that I want to look more deeply into, such as robust mirror descent, some work by Candès about SVD when we don't care about near-zero SVs, regularized stochastic gradient (Xiao) and sparse eigenvector work. Lots of awesome stuff. My favorite part of HIPS, ...'. The text 'momentum is the same as CG!' is highlighted in yellow.

Semiautomatic annotation

AnnoEditor v0.1 | Status Bar | Doc annotations | Suggest annotations | Annotate document | Connect | Create Annotation

Styles | Paragraph | Font family | 3 (12pt)

Scaling up, a 100 topic model run on 35 million tweets took 3 hours and 15 minutes to complete on my laptop. Ramage et al. report training a circa 800 topic Labelled LDA model on 8 million tweets in 96 machine-days (24 machines for 4 days). It's not quite apples-to-apples, but I figure the online LDA implementation in vowpal is somewhere between 2 and 3 orders of magnitude faster.

Create Annotation

- Annotation
 - Model size
 - Number of tweets
 - Duration

Type: Number of topics

Fragment: 800 topic

Content:

Multimedia ontology

The screenshot shows the AnnoEditor v0.1 interface. The top bar includes buttons for 'Status Bar', 'Doc annotations', 'Suggest annotations', 'Annotate document', 'Connect', and 'Create Annotation'. The main editing area displays a document titled 'Vojtěch Svátek - CV, topics of interest, projects and publications'. The document content includes a list of 'Important past projects' with entries for 'K-Space (IST Network of Excellence, 2006-2008)' and 'MedIEQ'. A 'Create Annotation' dialog box is open in the foreground, featuring a tree view on the left with categories like 'Annotation', 'Acronym', 'Title', 'Theme', 'Type', 'Start date', 'End date', 'Coordinator', 'Scientific Manager', 'Other Partners', 'Partner', 'Roles in the Project', 'Team Leader', 'Team Members', 'Team Member', 'Topic', 'Who says it', and 'When'. The dialog has input fields for 'Type' (set to 'University -> Division'), 'Fragment' (containing 'Queen Mary College, University of London'), and 'Content'. It also includes 'Browse', 'Select', 'Save', and 'Cancel' buttons.

AnnoEditor v0.1 | Status Bar | Doc annotations | Suggest annotations | Annotate document | Connect | Create Annotation

Vojtěch Svátek - CV, topics of interest, projects and publications

...

Important past projects:

- **K-Space (IST Network of Excellence, 2006-2008)**, coordinated by **Queen Mary College, University of London**, focusing on the **semantics of multimedia**. I was local contact person for the project. Our group was involved in **multimedia ontology design, complementary textual resource analysis, semantic repository querying and similarity-based image retrieval**. I was also the **Editor-in-Chief** of the 6-monthly K-Space Newsletter.
- **MedIEQ** - Quality Labeling of Multimedia Information in English (Greece), focusing on support of information extraction. In particular, the **Ex** information extraction system was used in 2006-2008, coordinated by ICSR, involved in web information extraction (in

Create Annotation

Type: University -> Division [Browse]

Fragment: Queen Mary College, University of London [Select]

Content:

Save Cancel



Outline

- 1 Motivation and Objectives
 - Introduction
 - Sources of inspiration
 - Goals of work
- 2 4A Framework Infrastructure
 - Annotation format
 - Exchange protocol
- 3 Proof-of-Concept Implementation
 - 4A Server
 - 4A Clients
- 4 Future Work

Future development and validation

- Finish clients for web browsers
- A client for annotating PDFs
- Server enhancements
 - better annotation suggestions
 - many instances of the server, distributed environment
 - additional modules (e.g., workflow support)
- Validation by means of annotation experiments