

NLP for the Masses: Integrating GATE with Desktop Clients

René Witte

Semantic Software Lab
Concordia University
Montréal, Canada

FIG3, Montréal, 2010

The Problem...

Amazon.com: The Catcher in the Rye (9780316769174): J. D. Salinger: Books - Web Bro

File Edit View History Bookmarks Tools Help

http://www.amazon.com/Ca

amazon.com Hello. [Sign in](#) to get personalized recommendations. New customer? [Start here.](#)

Your Amazon.com Today's Deals Gifts & Wish Lists Gift Cards

Shop All Department Search Books

Books Advanced Search Browse Subjects New Releases Bestsellers The New York Times Bestsellers

Click to **LOOK INSIDE!**

The Catcher in the Rye [Paperback]
J. D. Salinger (Author)

★★★★★ (2,987 customer reviews)

List Price: ~~\$13.99~~
Price: **\$11.19** & eligible for **FREE Super Saver Shipping** on orders over \$25. [Details](#)

You Save: **\$2.80 (20%)**

In Stock.
Ships from and sold by **Amazon.com**. Gift-wrap available.

Want it delivered **Wednesday, June 23?** Order it in the next **2 hours and 45 minutes**, and choose **One-Day Shipping** at

Transferring data from ecx.ima... TMN: Off FoxyProxy: Disabled

Outline

- 1 Introduction
- 2 Semantic Assistants: NLP Web Services
- 3 Desktop Plug-Ins for NLP
- 4 Conclusions

Introduction

Information Overload

- Web 2.0 applications lead to more user-generated content (e.g., product reviews)
- News Business: professional and layperson as content creators (e.g., Twitter, blogs, social networks)
- Digitization of printed media

Information Processing

- Finding information is fast, analysing consumes a lot of time
- Applies to E-mail, Web documents, Intranets, ...
- GATE NLP pipelines can help (e.g., to summarize product reviews) – but how do we get them to the users?

Where are we today?

Status Quo in NLP

- A solid decade of NLP framework development
- Variety of robust NLP plugins and pipelines
 - Summarization, Question-Answering, Information Extraction, Opinion Mining, ...

Status Quo in Desktops

- But what is available to end users on their desktop today?
- No NLP in word processors, email clients, Web browsers, IDEs, ...

Why?

- 1 users don't want/need NLP?
- 2 lack of software engineering work covering NLP?

Supporting the “Knowledge Worker”

Typical Workflow, as of today

- Receives task/request via email, text message, etc.
- Searches for information via Google, Desktop Search, etc.
 - Note: typically involves lots of natural language documents
 - Information Retrieval (IR) alone not sufficient
- Read, understand, evaluate results. Solve task. Repeat.

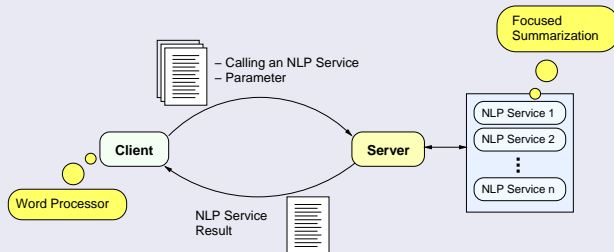
With Semantic Assistants

Users stay within (desktop) tool environment needed for their task

- tool (e.g., word processor) recognizes user’s need for information
- tool initiates data search in the background
- analysis tools (text/data mining) evaluate search results
- assistant offers results to user within his interface

Semantic Assistants

Workflow Overview



Features

- Focus on the user's needs, not the NLP tool
- Avoid context-switches through application integration

- 1 Introduction
- 2 Semantic Assistants: NLP Web Services
 - Requirements
 - System Architecture
 - Ontology-Based Context Model
 - Application Integration
- 3 Desktop Plug-Ins for NLP
- 4 Conclusions

Semantic Assistants

Support three distinct user groups

End Users: no knowledge in software engineering, NLP, or language engineering. Needs easy access to semantic (NLP) services within his desktop tools.

Language Engineer: develops NLP resources, tools, and pipelines. Presume no particular knowledge of server deployment or client (GUI) integration.

System Integrator: provides the integration of NLP services into desktop clients. Don't assume knowledge of NLP foundations or (GATE) framework details.

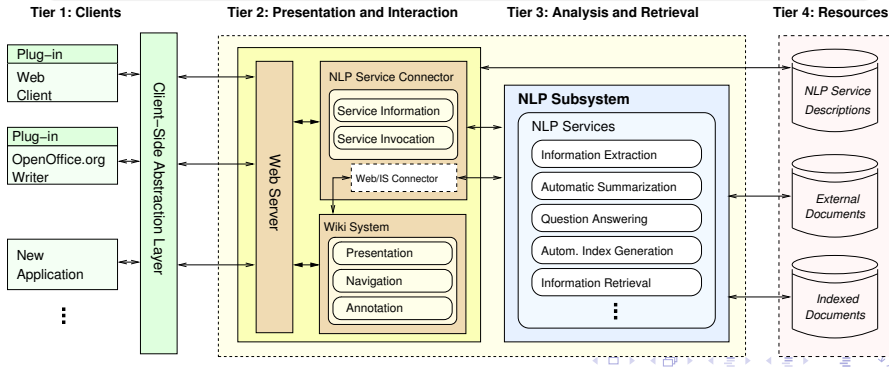
Semantic Assistants Approach

Provide a separation of concerns for these user groups.

Integration Architecture

Web Service-Based Architecture

- Standard (W3C) Web Services, using WSDL & SOAP
- Implemented using Java Web Services (Java 6)
- Client-Side Abstraction Layer (CSAL) for easy integration



Modeling the User's Context

Two-tiered Ontology Model

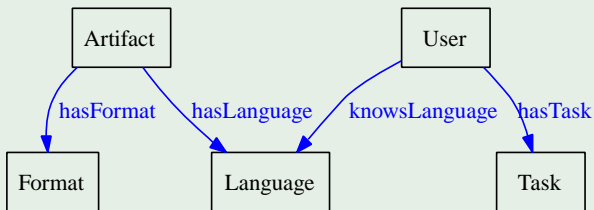
Modeling task, tools, available resources in an *upper ontology*

- reusable across domains (e.g., software process model)

Concrete ontology models domain-specific information

- user's languages, available (text) analysis pipelines, etc.

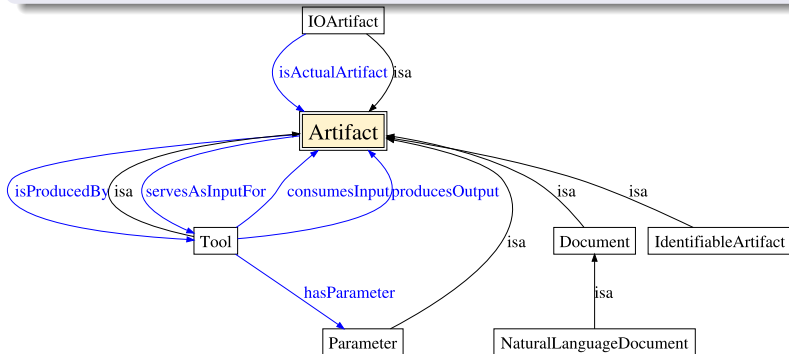
The five main concepts



Abstract Ontology

Contains the central **Artifact** concept

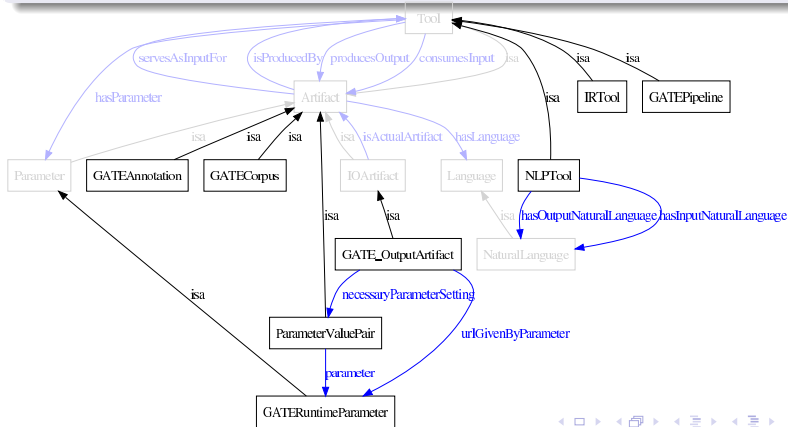
- Parent concept for programs, documents, parameters, etc.
- Important relations, e.g., *consumesInput*, *producesOutput*, *hasParameter*, ...



Concrete Ontology

Extending the upper ontology

- Domain-specific concepts, e.g., *NLPTool*
- Some implementation-specific concepts



Steps for Integrating GATE into a (Desktop) Application

NLP Work

- Develop the GATE Pipeline as usual (.gapp file)

Publish GATE Pipeline as Web Service

- Write meta-information for the pipeline (input, output, language, etc.) in a corresponding OWL file
- Server will publish the new pipeline to (all) clients

Develop Client Plug-In

- Very easy for Java-based plug-ins (using provided CSAL abstraction layer)
- Most other languages have support for using the published Web Service (WSDL) description

Code Examples

Client connecting to the server and finding available services

```
// Create a factory object
SemanticServiceBrokerService service
    = new SemanticServiceBrokerService();

// Get a proxy object, which locally represents the service endpoint
SemanticServiceBroker broker
    = service.getSemanticServiceBrokerPort();

// Proxy object is ready to use. Get a list of available NLP services.
ServiceInfoForClientArray sia = broker.getAvailableServices();
```

Making use of the OWL ontology model

Finding appropriate NLP services

Client plug-in can query server for NLP services that are useful for the current user, based on the context:

- Language capabilities of user/pipeline
- Current user client
- ...

Queries are implemented using SPARQL

Ask for services that produce English or German as output:

```
SELECT ?x ?name
WHERE { ?x sa:hasGATENAME ?name .
        {?x cu:hasFormat sa:GATECorpusPipeline_Format} . {
          {?x sa:hasOutputNaturalLanguage cu:en} UNION
          {?x sa:hasOutputNaturalLanguage cu:de}}
}
```


- 1 Introduction
- 2 Semantic Assistants: NLP Web Services
- 3 Desktop Plug-Ins for NLP
 - Word Processing
 - Software Engineering
- 4 Conclusions

An everyday task

Given

Access to the Web

Question

What is the role of DMSP [Dimethylsulfoniopropionate] in the Atlantic marine biology within the global climate change?

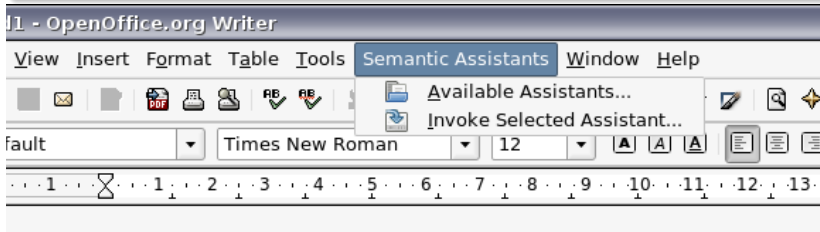
Your Task

Write a 450-word essay answering the question!

Semantic Assistants Client: OpenOffice.org *Writer*

Client-side plugin for word processor integration

- Users get a new menu item: “Semantic Assistants”
- Available services are discovered based on context through ontology queries (SPARQL)
- Services executed asynchronously in the background
- Results displayed based on type (new text window or browser window)

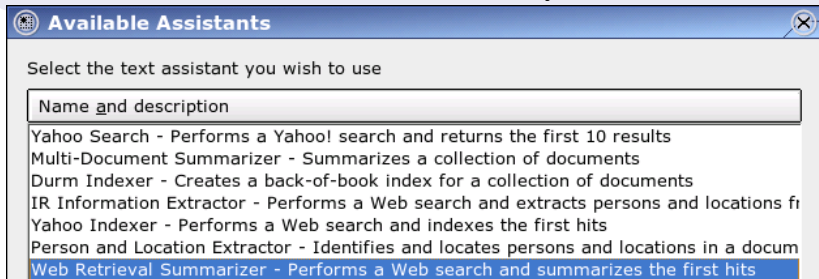


Invoking an Assistant

Example Scenario

User works on a report about “*DMSP in the Atlantic marine biology*” (DMSP = Dimethylsulfoniopropionate)

- The “Web Retrieval Summarizer” will initiate a web search, retrieve the results, automatically summarize the content of the top-*n* hits
- result can be further edited or refined by the user

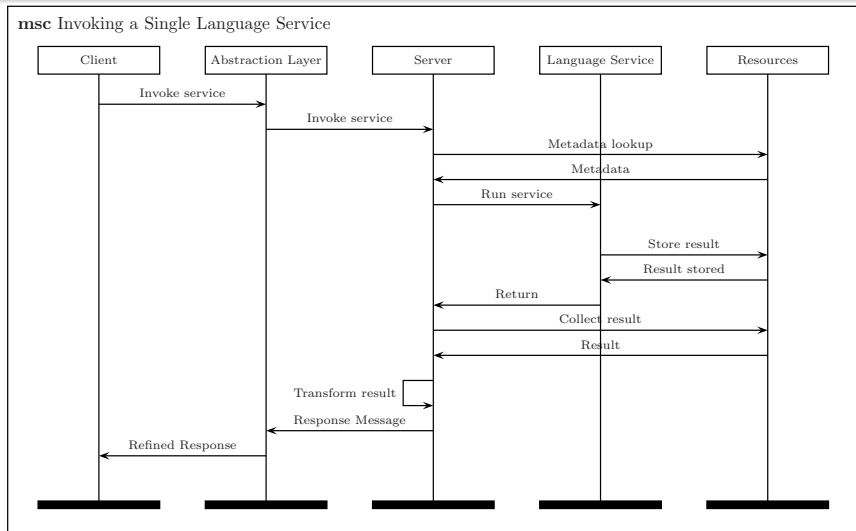


Available Assistants

Select the text assistant you wish to use

Name and description
Yahoo Search - Performs a Yahoo! search and returns the first 10 results
Multi-Document Summarizer - Summarizes a collection of documents
Durm Indexer - Creates a back-of-book index for a collection of documents
IR Information Extractor - Performs a Web search and extracts persons and locations fr
Yahoo Indexer - Performs a Web search and indexes the first hits
Person and Location Extractor - Identifies and locates persons and locations in a docum
Web Retrieval Summarizer - Performs a Web search and summarizes the first hits

Assistant Invocation – Communication Flow





Default Times New Roman 12 B I U

1 2 3 4 5 6 7 8 9 10 11 12 13

We are examining the activity and regulation of DMSP lyase in the oxidative str the phytoplankton. The researchers discovered that the bacterioplankton in the SAR11 groups are the primary plankton involved with directing DMSP away from making sulfur unavailable to atmospheric processes. Because of potential mism some Roseobacter group members and the MALF-1 probe, clones known to belong to the Roseobacter group (based on partial 16S rRNA gene sequences) but having varying complementarity to the probe (zero, one, or four mismatches) were used in quantitative hybridizations. Including both strongly and weakly hybridizing clones, the percentage positive for the MALF-1 probe were similar for the surface samples outside (sample 11; 8%) the eddy and slightly lower in the deep-water sample (sa

- 1 Introduction
- 2 Semantic Assistants: NLP Web Services
- 3 Desktop Plug-Ins for NLP
 - Word Processing
 - Software Engineering
- 4 Conclusions

NLP Support for Software Engineers

Software Artefacts Written in Natural Language

- Requirements Documentation
- Issue Tracker Messages
- SVN Commit Messages
- Source Code Comments

Quality Assessment

- No efficient means of analysing the quality of content
- Guidelines usually enforced manually
- Automatic quality assessment tools needed

Automatic Quality Assessment of Source Code Comments

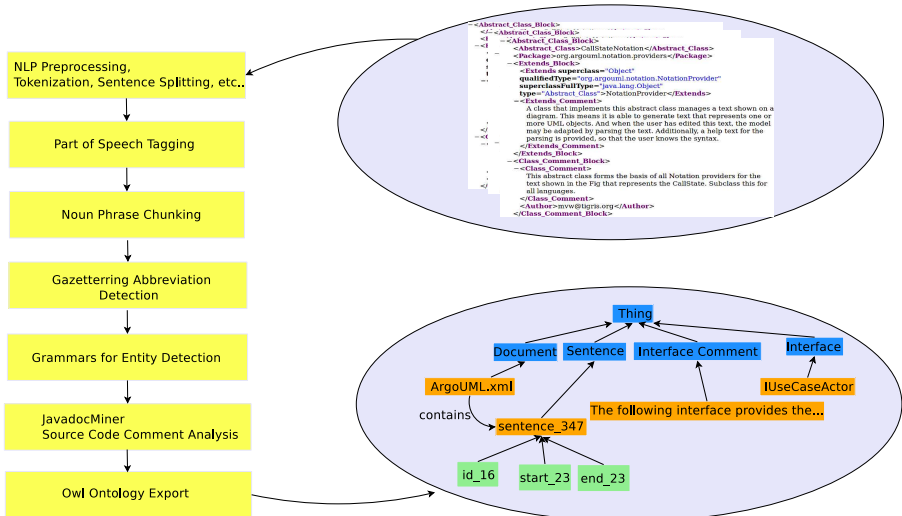
Analysis Heuristics

- Internal (NL Quality) Comment Analysis: A set of heuristics targeting the natural language quality of the in-line documentation itself
- Code/Comment Consistency Analysis: The following heuristics analyse in-line documentation in relation to the source code being documented

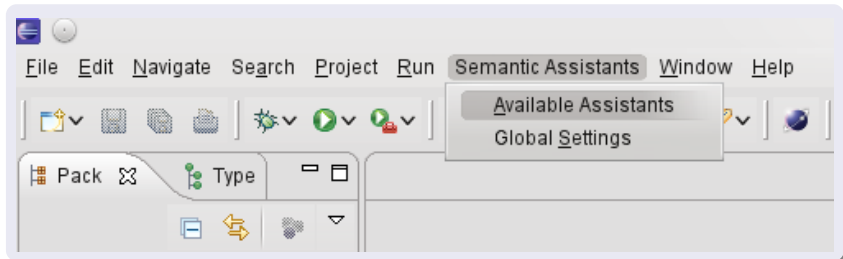
Semantic Assistants for Software Engineers

- NLP analysis of source code comments implemented as a GATE pipeline, 'JavadocMiner'
- Published as Web service through the Semantic Assistants framework
- Integrated into Eclipse through an SA plugin

JavadocMiner Implementation



Eclipse Plug-In



Features

- Execute GATE pipeline with file loaded into Eclipse (e.g., Java Source Code)
- Display results, create marks when line numbers are available

ANNIE results in Eclipse (Table View)

Project	Class Name	Type	Content	Start	End	Features
FirstProject	ClassOne.java	Location	England	231	238	locType=
FirstProject	ClassOne.java	Location	Hollywood	558	567	locType=city
FirstProject	ClassOne.java	Location	Manhattan	1017	1026	locType=
FirstProject	ClassOne.java	Location	New York	1028	1036	locType=city
FirstProject	ClassOne.java	Location	21 60 Clinton Avenue	1381	1400	locType=
FirstProject	ClassOne.java	Person	Stanley Kubrick	104	119	gender=male
FirstProject	ClassOne.java	Person	Kubrick	291	298	gender=male
FirstProject	ClassOne.java	Person	Stanley Kubrick	947	962	gender=male
FirstProject	ClassOne.java	Person	Jacques Leonard Kubrick	1072	1095	gender=male
FirstProject	ClassOne.java	Person	Gertrude	1119	1127	gender=female
FirstProject	ClassOne.java	Person	Barbara	1165	1172	gender=female

Java - Test/src/org/argouml/uml/diagram/collaboration/ui/UMLCollaborationDiagram.java - Eclipse P

File Edit Source Refactor Navigate Search Project Run Semantic Assistants Window Help

UMLCollaborationDiagram.java

```

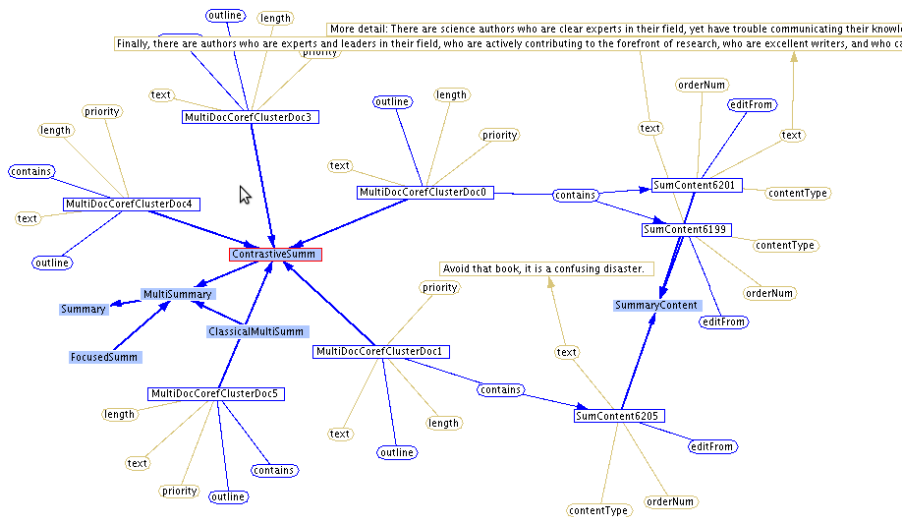
182     }
183     }
184
185     /**
186     * Get the actions from which to create a toolbar or equivalent
187     * graphic triggers.
188     * {@inheritDoc}
189     */
190     MethodCommentStyle=Use 3rd person (descriptive)not 2nd person (prescriptive).(Gets the label) |
191     NumberOfNouns=4 | NumberOfVerbs=3 | NumberOfTokens=18 | ABBCOUNT=0 | line=190 |
192     KincaidMetric=10.941429 |
193
194     getAssociationActions(),
195     getActionGeneralize(),
196     getActionDepend(),
    
```

Problems Javadoc Declaration Semantic Assistants Semantic Assistants Status

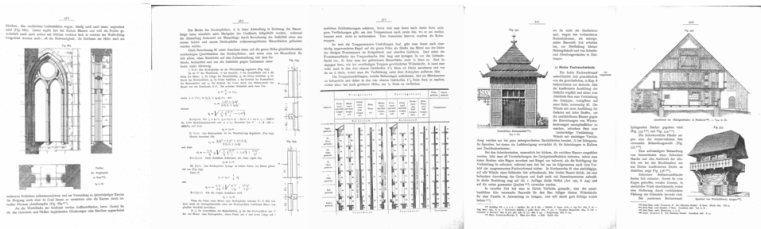
Project	Class Name	Type	Content	Start	End	Features
Test	UMLCollaborat	Method_Comr	Method to perform a	1029	1504	MethodCommentStyle=Use
Test	UMLCollaborat	Method_Comr	Get the actions from	1609	1703	MethodCommentStyle=Use
Test	UMLCollaborat	Method_Comr	After loading the dia	1755	1936	MethodCommentStyle= Nur
Test	UMLCollaborat	Method_Comr	A sequence diagram	2789	2954	MethodCommentStyle=Use

- 1 Introduction
- 2 Semantic Assistants: NLP Web Services
- 3 Desktop Plug-Ins for NLP
- 4 Conclusions
 - Product Reviews
 - Heritage Data Analysis
 - Moving on. . .

Analysing Product Reviews



Heritage Data Analysis



Large unstructured corpora

- Outdated terms, style of writing, huge amount, no categorization or assessment
- Comparing and evaluating with current content

Wiki/NLP Integration

Link to Discussion
Page for this Section

Link to Chapter
of Section

Full Text Search

New Page in
Original Document

Link to Figures in
Original Document

Subsection
Heading

The screenshot shows a Wikipedia article titled "Durm: Seitliche Begrenzung". The page is annotated with red circles and lines connecting to labels on the left. The "Diskussion" link is circled. The main heading "Durm: Seitliche Begrenzung" is circled. The "in Kapitel 29" link is circled. The "Inhaltsverzeichnis" table is circled. The "Seite 442" and "Scan" links are circled. The "Grundrißgestaltung. (394.)" subsection heading is circled. The "Werkstoff. (395.)" subsection heading is circled. A large red circle highlights a figure showing architectural drawings of a window opening, with a "Grafik" label circled. The page includes a navigation sidebar, a search box, and a list of tools.

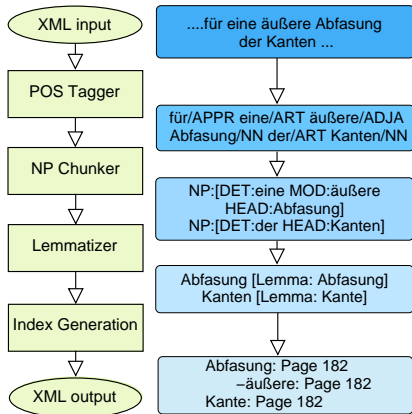
Section Heading
Main Table
of Content

Section Table
of Content

Link to Scan
of Original Page

Preview of Scan
of Original Page
and Link to
High-Resolution
Version

Back-of-the-Book Index Generation



[artikel](#) | [diskussion](#) | [bearbeiten](#) | [versionen](#)

Index

Abbinden	137		
Abbrechen	143		
Abdecken	125		
Abdeckung	131	108	
Abfallstoffen			
oft	129		
werdenden	129		
lästig	129		
Abfassung	19	181	139
ausgedehnte	186		
angebrachte	180		
äußere	182		
äußerer	182		

Conclusions and Future Work

Semantic Assistants Architecture

Semantic “glue” allows to model users, their context, and NLP services, and rapidly integrate them to provide support to users

- OWL ontologies and SPARQL queries
- Integration using W3C Web services
- Abstraction layer for easy client integration

Further Improvements

- Enhancing the context model
- Domain-specific ontology refinements
- Improvement of many technical details
- Development of further client plug-ins

More information

Semantic Assistants Distribution

Papers and additional documentation is available on

- <http://www.semanticsoftware.info>
- <http://rene-witte.net>

Source code on SourceForge:

- <http://sourceforge.net/projects/semantic-assist/>

Acknowledgements

Semantic Assistants

- Bahar Sateli, Nikolaos Papadakis, Tom Gitzinger

JavadocMiner

- Ninus Khamis
- Juergen Rilling

Heritage Data Analysis

- Thomas Kappler
- Ralf Krestel
- et al.

Summarization, Product Review Analysis

- Sabine Bergler, Ralf Krestel, et al.