

Semantic Metadata Extraction using GATE

Diana Maynard
Natural Language Processing Group
University of Sheffield, UK

OOA-HR Workshop, 11 October 2006
http://www.jarrar.info/OOA/OOA-HR_workshop.htm

h-TechSight project

- Integration of a variety of next generation knowledge management technologies, in the domain of chemical engineering.
- Knowledge Management Portal enables support for knowledge intensive industries in monitoring information resources on the Web:
 - observe information resources automatically on the internet
 - notify users about changes occurring in their domain of interest.
- Much effort in terms of knowledge management has been placed in the area of employment because it affects every organisation and business
- Monitoring job advertisements over time can alert users to changes such as requirements for skills, general trends in the field, comparison of salaries, etc.

An Architecture for Language Engineering

- GATE is used to enable the ontology-based semantic annotation of web-mined documents
- Instances in the text are linked with concepts in the ontology
- Performs analysis of unrestricted text to extract from the text instances of concepts in the ontologies
- Instances linked to the ontology are exported to a database, enabling monitoring of such instances and concepts over time, according to the user's interests

Gate IE system

- Architecture consists of a pipeline of processing resources which run in series
- Many of these processing resources are language and domain-independent
- Pre-processing stages include:
 - word tokenization
 - sentence splitting
 - part-of-speech tagging
- Main processing is carried out:
 - by a gazetteer linked to one or more ontologies
 - by a set of grammar rules

Demo Employment ontology

- Demo Employment ontology has 9 Concepts: Location, Organisation, Sectors, JobTitle, Salary, Expertise, Person and Skill
- Each concept in the ontology has a set of gazetteer lists associated with it, which help identify instances in the text
 - default lists - quite large and contain common entities such as first names of persons, locations, abbreviations etc.
 - domain-specific lists - need to be created from scratch.
 - keyword lists - collected for recognition purposes to assist contextually-based rules, also attached to the ontology, because they clearly show the class to which the identified entity belongs.
- Lists can be acquired automatically from the web or from training data

Populated ontology

- Lists are linked directly to an ontology, such that instances found in the text can then be related back to the ontology

The screenshot displays a software interface with three main panels:

- Ontology:** A tree view showing a hierarchy starting with 'Ontology_0003A' and 'DEFAULT_ROOT_CONCEPT'. Under 'Skill', there are sub-categories like 'Computing_Skill', 'Personal_Skill', 'Location', 'Recruitment', 'Sectors', and 'Expertise'. 'Computing_Skill' is further divided into 'Servers', 'User_Interface', 'Computational_Packages_Skill', 'Programming_Packages_Skill', and 'Simulation_Packages_Skill'.
- Linear Definition:** A list of ontology instances with their corresponding URIs. The selected instance is 'Computing_Skill/programming_packages.lst:skill', with its URI being 'http://gate.ac.uk/projects/htechsight/Employment.daml:Computing_Skill/programming_packages.lst:skill'.
- Gazetteer List:** A list of terms extracted from text, including 'Dymola', 'ASCEND', 'SIMPROCESS', 'BATCHCAD', 'MASSBAL', 'INDISS', 'Pascal', 'XML', 'Assembly', 'C', 'C++', 'SQL', 'Perl', 'Fortran', 'FORMTRAN', 'Prolog', 'Foxpro', 'BASIC', 'Tcl-Tk', 'Forth', 'MS-Access', 'Ada', 'lisp', 'COBOL', 'Cobol', 'VISUAL BASIC', 'Visual Basic', 'Visual C', and 'VISUAL C'.

The 'Mapping Definition' panel at the bottom shows the mapping between the ontology instances and their URIs, such as 'Application/CV.lst:http://gate.ac.uk/projects/htechsight/Employment.daml:CV.lst'.

Visualisation of Results

- Implemented as a web service.
- User selects a URL and the concepts in which he/she is interested
- System performs the analysis
- User can view analysis in different ways

Employment based Information Extraction

Enter URL :

Annotations (Employment Ontology) :

<input checked="" type="checkbox"/> Application	Concept Selection Area	<input checked="" type="checkbox"/> Salary
<input checked="" type="checkbox"/> Email		<input checked="" type="checkbox"/> Skills
<input checked="" type="checkbox"/> Organization		<input checked="" type="checkbox"/> Person
<input checked="" type="checkbox"/> Expertise		<input checked="" type="checkbox"/> JobTitle
<input checked="" type="checkbox"/> Location		

Run Gate

-Gate results analysis

Click here to view the Dynamics of instances.

Check the Analysis for the following months.

Click here to view the instances in the month of JAN

Click here to view the instances in the month of FEB

Click here to view the instances in the month of MAR

Click here to view the instances in the month of APR

OOA-HR Workshop, 11 October 2006

Visualisation of Results

A new web page is created with highlighted annotations

Information Extration for the URL : http://www.jobserve.com/it/jobserve/searchresults.asp?jobType=*&d=0&page=1&q=java&order=DateTime

Annotation Keys:

Application **Email** **Organization** **Expertise** **Location** **Salary** **Skills** **Person** **JobTitle**

GATE output :

Bonus Posted: 26/04/2004 14:13:47 Reference: JSPM1584



[Senior Java Developer, E-Commerce, Investment Banking, Lon](#)

Senior role for Senior Java Developer with lots of **Multithreading** experience and **current investment bank**. Essential: this is a senior roles so you will need no less than 4 years . Office environment, also good Sybase with **Multithreading programming** experience. Pi banking experience with **FX** and **money markets product** knowledge. Top tier investment requires a Java Developer with around **4 years Commercial** Experience to work in esales solutions [more ->](#).

Type: Permanent Location: City Of London Country: England Start: ASAP / One Month Salary/Rate: 50k-60

Bonus Posted: 26/04/2004 14:13:45 Reference: JSPM1585



OOA-HR Workshop, 11 October 2006

Database Output

The occurrences of the instances over time are stored dynamically in a database

Concept	Frequency
Contract	5
AnnotationString	Frequency
4-6 month contract assignment	1
Temporary	1
Contract	1
Project	1
6 MONTHS	1
*	0
Curriculum Vitae	18
AnnotationString	Frequency
resume	9
Resumes	8
resume/CV in word format	1
*	0
E-Mail	2
AnnotationString	Frequency
email	1
E-mail	1
*	0
Full-Time	21
Location	5
On-Line	2
Parmanent	1
Part-Time	1
Postgraduate	7
AnnotationString	Frequency
PhD in chemical engineering	2
related fields with an emphasis in chemical proces	1
Ph.D in Chemical Engineering	1
Polymer Science	1
Textile Technology	1
MS or PhD in Chemical engineering	1
*	0
Undergraduate	51
AnnotationString	Frequency
BS Chemical Engineer	32
BS degree in Chemical Engineering	8
Bachelor Degree Paid	4
Bachelor Degree Paid	4
Bachelor of Science Chemical Engineering	1
BS/MS Degree in Chemical Engineering	1
BS, Chemical Engineering	1
*	0
USA	10
AnnotationString	Frequency
Wayne IN	1
Frazer, PA	1
Pleasanton CA	1


Concepts	Annotations	Document ID	Time Stamp
Url	http://jobsearch.monster.com/getjob.asp?JobID=1687	16870161	Jan 15
Company	DuPont Human Resources US-TX-Orange	16870161	Jan 15
TempOrganizat	Orange	16870161	Jan 15
Title	Research Chemical Engineer - Level 04-05A	16870161	Jan 15
Location	USA	16870161	Jan 15
Sector	Chemical Engineer	16870161	Jan 15
TempOrganizat	Orange	16870161	Jan 15
TempOrganizat	Orange	16870161	Jan 15
Status	Full Time	16870161	Jan 15
Reference	ENG00030	16870161	Jan 15
Location	Orange	16870161	Jan 15
TempOrganizat	Orange	16870161	Jan 15
Qualification	PhD in chemical engineering	16870161	Jan 15
Qualification	related fields with an emphasis in chemical proces	16870161	Jan 15
JobTitle	Candidate	16870161	Jan 15
JobTitle	pilot	16870161	Jan 15
Expertise	adipic acid production	16870161	Jan 15
Expertise	crystallization	16870161	Jan 15
Expertise	solids handling	16870161	Jan 15
JobTitle	candidate	16870161	Jan 15
TempOrganizat	DTI	16870161	Jan 15
JobTitle	candidate	16870161	Jan 15
Expertise	process optimization	16870161	Jan 15
Expertise	process engineering design	16870161	Jan 15
Expertise	simulation for developmental projects	16870161	Jan 15
Expertise	communicate well	16870161	Jan 15

Dynamics of Concepts

Users may see tabular results of statistical data about how many annotations each concept had in the previous months, as well as seeing the progress of each instance in previous time intervals

Gate Analysis - Dynamics of Instances

Concepts	Count of instances for January	Count of instances for February
Application	4	46
Citizenship	0	3
Email	0	15
Expertise	64	1798
FirstPerson	8	313
JobTitle	20	513
Location	0	13
Money	0	42
MoneyNot	0	10
Organization	26	420
Period	20	553
Phone	0	5
Qualification	4	51
Salary	12	200
Skills	74	1044


**Click a Concept to
see Dynamics of its
Instances**

Dynamics of Instances

- DF is an elasticity metric that quantifies dynamics of an instance, taking account of volume of data and time period
- Instances for the concept "Organisation" can track the recruitment trends for different companies
- Monitoring instances for concepts such as Skills and Expertise can show which kinds of skills are becoming more or less in demand.

Instance	DF	Jan	Feb	Mar
ARC	145	-1	12	6
Archimedia SA	-1	0	1	0
Army	23	0	2	1
AT&T	-1	0	2	0
BA	23	0	3	1
BMI British Midland	-335	1	3	0

Evaluation and User feedback

- Overall, the system achieved 97% Precision and 92% Recall
- Tested by real users in industry, e.g. Bayer, JetOil, IChemE.
- Found to be “helpful in increasing efficiency in acquiring knowledge and supporting project work...helping to scan, filter, structure and store the wealth of information”
- Application areas spanned from R&D, engineering and production, to marketing and management
- Employment application was a “valuable means of graduates gaining a fresh insight into their jobs and related training which may be narrower than it ideally should due to company constraints (i.e. time and money)

OOA-HR Workshop, 11 October 2006