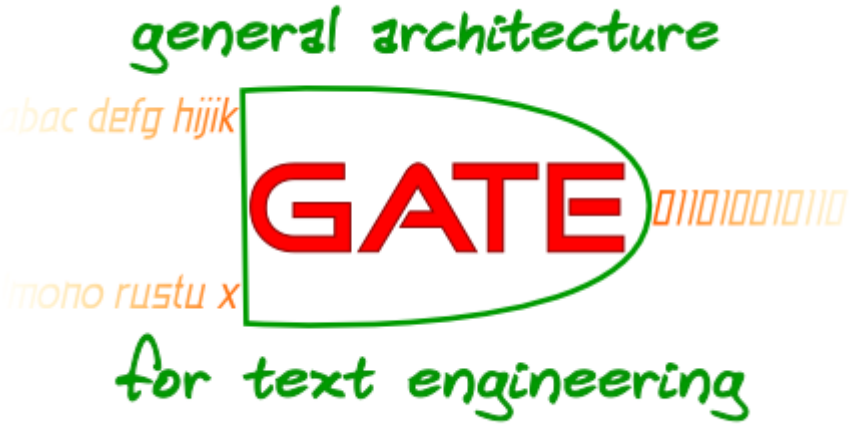


GATE and Semantic Annotation of Web Services

Adam Funk

University of Sheffield (UK)

01.09.2009

Aims

- Investigate some aspects of natural language processing (NLP) and information extraction (IE) for the Semantic Web
- Provide an introduction to the GATE toolkit
- Explain how we develop novel extensions for GATE and show some extensions developed for Service-Finder

Challenge of the SW

- requires machine-readable, flexible (multi-purpose) data ... easier to classify, search, and monitor
- two types:
  - **Explicit** metadata describe the document (HTML titles, datestamps)
  - **Implicit** metadata are **deduced from the text (**named entities, relations, categories)

Some terminology

- **Semantic annotation** – annotate in the texts all mentions of instances relating to concepts in the ontology
- **Ontology learning** – automatically derive an ontology from texts
- **Ontology population** – given an ontology, populate the concepts with instances derived automatically from a text

Knowledge engineering & machine learning

Knowledge Engineering

- rule based
- developed by experienced language engineers
- make use of human intuition
- require only small amount of training data
- development can be very time consuming
- some changes may be hard to accommodate

Learning Systems

- use statistics or other machine learning
- developers do not need LE expertise
- require large amounts of annotated training data
- some changes may require re-annotation of the entire training corpus

Entity Recognition: the cornerstone of IE

- Traditionally, identifying proper names (...) in texts and classifying them into a set of predefined categories of interest
  - Persons
  - Organisations
  - Locations
  - Dates

Typical NE pipeline

- Pre-processing (tokenization, sentence splitting, morphological analysis, POS tagging)
- Entity recognition (gazetteer lookup, NE grammars)
- Coreference (orthographic coreference, anaphora resolution)
- Export to database / XML
  - and now RDF

the VW Beetle of NLP

- **Over ten** years old, with 000s of users at 00s of sites
- **An architecture** for organizing & integrating language engineering tools
- **A framework** for programmers: an object-oriented class library that implements the architecture
- **A development environment**: a GUI for language engineers, computational linguists, etc.
- **Some free components** ... and wrappers for other people's components
- **Tools** for evaluation, visualizing, editing, persistence, IR, IE,ontologies, etc.
- **Free** software (LGPL)

the VW Beetle of NLP

- http://gate.ac.uk/
  - download official releases, nightly builds
  - a fairly comprehensive user guide
  - examples, movies
  - API documentation
  - links to other parties' plug-ins
  - gate-users mailing list
- http://sourceforge.net/projects/gate/develop
  - svn

the VW Beetle of NLP

- **Over ten** years old, with 000s of users at 00s of sites
- **An architecture** for organizing & integrating language engineering tools
- **A framework** for programmers: an object-oriented class library that implements the architecture
- **A development environment**: a GUI for language engineers, computational linguists, etc.
- **Some free components** ... and wrappers for other people's components
- **Tools** for evaluation, visualizing, editing, persistence, IR, IE,ontologies, etc.
- **Free** software (LGPL). Download at http://gate.ac.uk/download/

GATE Users

- American National Corpus project
- Perseus Digital Library project, Tufts University, US
- Longman Pearson publishing, UK
- Hewlett Packard ESC
- British Telecom labs
- Merck KgAa, Germany
- Canon Europe, UK
- Knight Ridder, US
- BBN (leading HLT research lab), US
- SMEs: Melandra, SG-MediaStyle, Sirma AI, ...
- a large number of UK, US and EU Universities and research centres: DERI, Stanford, Imqperial College London, University of Manchester, University of Karlsruhe, Vassar College, the University of Southern California, ...
- many UK and EU projects

Projects using GATE

- **MUSE**: multi-genre multilingual IE
- **HSL**: IE in domain of health and safety
- **Old Bailey**: IE on 17th century court reports
- **Multiflora**: plant taxonomy text analysis for biodiversity research in e-science
- **EMILLE**: creation of S. Asian language corpus
- **ACE / TIDES**: IE competitions and collaborations in English, Chinese, Arabic, Hindi
- **h-TechSight:** ontology-based IE and text mining
- **ETCSL**: Language tools for Sumerian digital library
- **SEKT**: Semantic Knowledge Technologies
- **PrestoSpace**: Preservation of audiovisual data
- **KnowledgeWeb:** Semantic Web network of excellence
- **MEDIACAMPAIGN:** Discovering, inter-relating and navigating cross-media campaign knowledge
- **TAO** : Transitioning Applications to Ontologies
- **MUSING** : SW-based business intelligence tools
- **NEON** : Networked Ontologies
- **Service-Finder!**

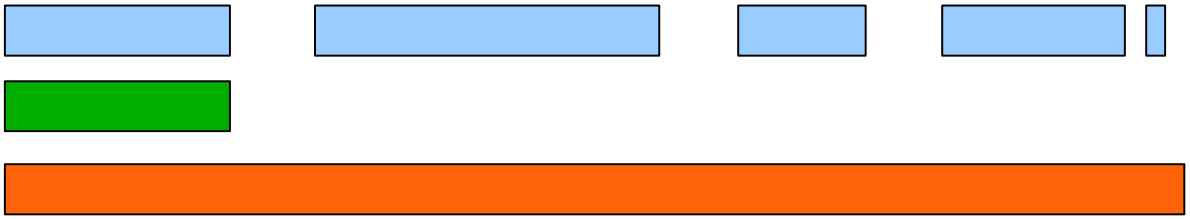
GATE

- Language Resources: data
  - documents (features, content, annotation sets, annotations, features)
  - corpora, ontologies, etc.
- Processing Resources: executable (JAPE, arbitrary code)
- Applications: pipelines of PRs
- Datastores: persistent storage
- Plug-ins

Annotations Example

Text: **Cyndi savored the soup.**

Nodes: **|0...|5...|10..|15..|20**



Annotation spans:

Annotation descriptions:

Id	Type	Start	End	Features
1	token	0	5	pos=NP
2	token	6	13	pos=VBD
3	token	14	17	pos=DT
4	token	18	22	pos=NN
5	token	22	23	
6	name	0	5	type=person
7	sentence	0	23	

What is ANNIE?

- ANNIE is a vanilla information extraction system with a set of core PRs
  - Tokenizer
  - Sentence Splitter
  - POS tagger for English
  - Gazetteers
  - Named entity tagger (JAPE transducers)
  - Orthomatcher (orthographic coreference)

ANNIE's Gazetteer Lists

- Set of lists compiled into Finite State Machines
- 60,000 entries in 80 types, such as organization; artifact; location; amount_unit; manufacturer; transport_means; company_designator; currency_unit; date; government_designator; ...
- Each list has attributes MajorType and MinorType and Language):

  city.lst: location: city: english

  currency_prefix.lst: currency_unit: pre_amount

  currency_unit.lst: currency_unit: post_amount

- List entries may be entities or parts of entities, or they may contain contextual information (e.g. job titles often indicate people)

NE transducer

- Gazetteers find terms that suggest entities, and their context
- These terms may be ambiguous:
  - Mrs. May Jones / 1st May 2006
  - Mr. Parkinson / Parkinson's disease
- Hand-crafted grammars are used to define patterns over the Lookups and other annotations
  - Disambiguate
  - Combine annotations: numbers, dates, money, names
- JAPE: regular expressions over annotation graphs

JAPE grammars

- JAPE is a pattern-matching language
- The LHS of each rule contains patterns to be matched
- The RHS contains details of annotations (and optionally features) to be created
- More complex rules can also be created, using Java code in the RHS

Machine learning plug-in

- Can be trained to make annotations based on presence and features of document annotations
  - NER (marking spans)
  - text classification (opinion mining, service categorization)
- SVM is the main type we use

GATE in Service-Finder

- Preprocessor
  - Heritrix arc.gz + index → 30 datastores
  - 8,000 corpora/providers, 23,000 services
  - 100,000 documents (300,000 duplicates)
- AA pipeline
  - ANNIE + custom PRs
  - Voting PRs
  - RDF-XML generator
- XML collator

GATE in Service-Finder

- Service categorization

	traditional %			BDM %		
	P	R	F1	AP	AR	AF1
Crawl 1	36	16	22			
Crawl 2	39	12	19			
Crawl 3	20	32	24	38	40	39

- so the keyword+rules system is not effective
- → try machine learning

Evaluation Metrics

- **Precision** = correct answers/answers produced

  true_pos / (true_pos + false_pos)
- **Recall** = correct answers/possible correct answers

  true_pos / (true_pos + false_neg)
- **F1** (balanced) = 2 P R / (2 (R + P))
- For ontological classification, we use the Balanced Distance Metric (BDM), which gives partial credit for "near misses" in the class tree

GATE in Service-Finder

- Recent experiments with machine-learning

Max	traditional %			BDM %		
categories	P	R	F1	AP	AR	AF1
1	58	53	55	79	55	65
2	50	54	52	69	56	62