

Introduction to Text Mining

Module 4: Applications (Part 2)

Rich News Multimedia Application

Multimedia annotation: Prestospace project

- Broadcasters produce many of hours of material daily (BBC has 8 TV and 11 radio national channels)
- Some of this material can be reused in new productions
- Access to archive material is provided by some form of semantic annotation and indexing
- Manual annotation is time consuming (up to 10x real time) and expensive
- Currently some 90% of BBC's output is only annotated at a very basic level

RichNews Tool

- A prototype addressing the automation of semantic annotation for multimedia material
- Not aiming at reaching performance comparable to that of human documentarists
- Fully automatic
- Aimed at news material, further extensions possible
- TV and radio news broadcasts from the BBC were used during development and testing

Overview

- Input: multimedia file
- Output: OWL/RDF descriptions of content
 - Headline (short summary)
 - List of entities (Person/Location/Organization/...)
 - Related web pages
 - Segmentation
- Multi-source Information Extraction system
 - Automatic speech transcript
 - Subtitles/closed captions
 - Related web pages
 - Legacy metadata

Key Problems

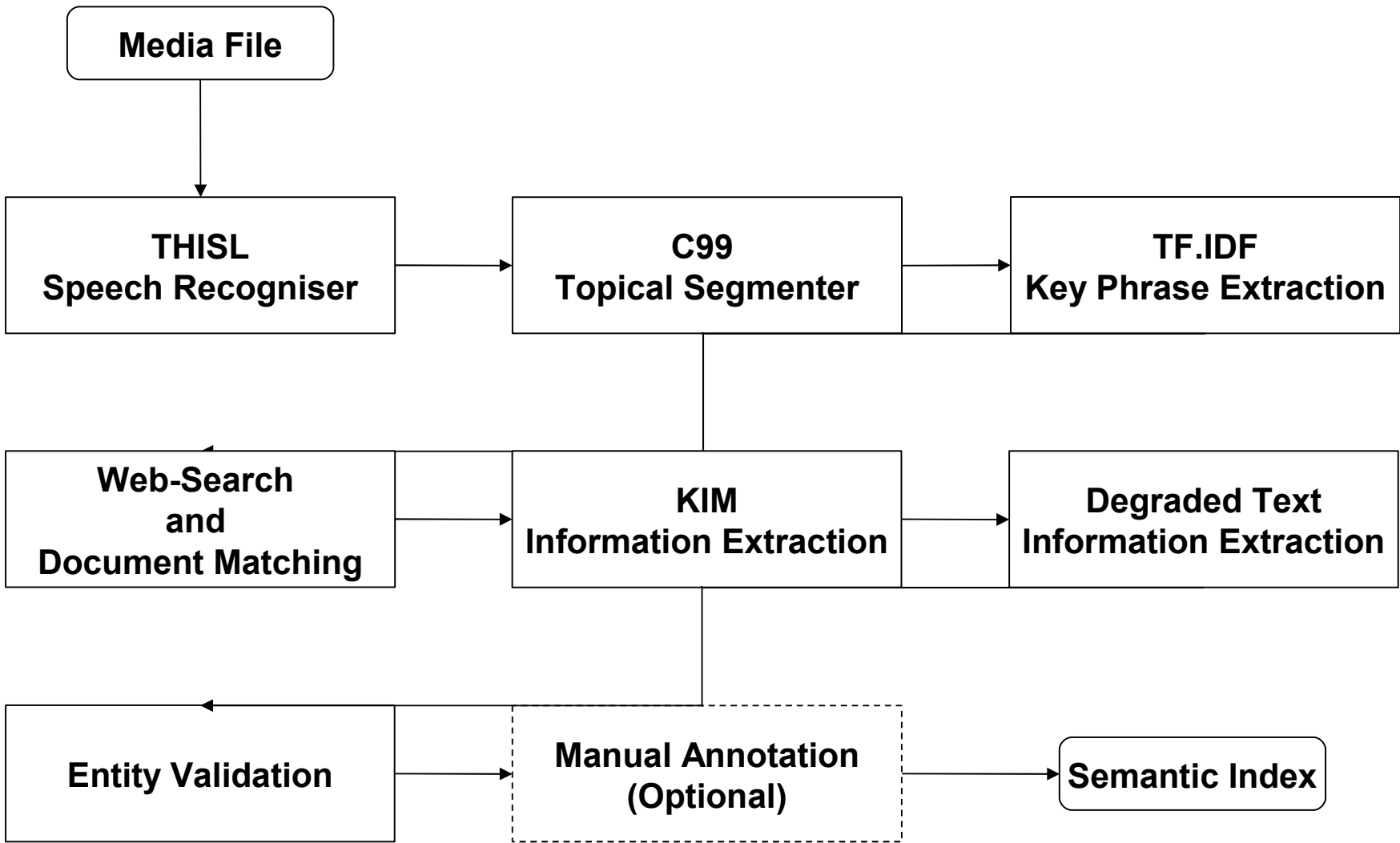
Obtaining a transcript:

- Speech recognition produces poor quality transcripts with many mistakes (error rate ranging from 10 to 90%)
- More reliable sources (subtitles/closed captions) not always available

Broadcast segmentation:

- A news broadcast contains several stories. How do we work out where one starts and another one stops?

Architecture



Using ASR Transcripts

ASR is performed by the THISL system.

- Based on ABBOT connectionist speech recogniser.
- Optimised specifically for use on BBC news broadcasts.
- Average word error rate of 29%.
- Error rate of up to 90% for out of studio recordings.

ASR

he was suspended after his
arrest [SIL] but the process
were set never to have lost
confidence in him



he was suspended after his
arrest [SIL] but the Princess
was said never to have lost
confidence in him

and other measures weapons
inspectors have the first time
entered one of saddam
hussein's presidential palaces



United Nations weapons
inspectors have for the first
time entered one of saddam
hussein's presidential
palaces

Topic Segmentation

Uses C99 segmenter:

- Removes common words from the ASR transcripts.
- Stems the other words to get their roots.
- Then looks to see in which parts of the transcripts the same words tend to occur.
- These parts will probably report the same story.

Key Phrase Extraction

Uses term frequency inverse document frequency (tf.idf):

- Chooses sequences of words that tend to occur more frequently in the story than they do in the language as a whole.
- Any sequence of up to three words can be a phrase.
- Up to four phrases extracted per story.

Web Search and Document Matching

- The Key-phrases are used to search on the BBC, and the Times, Guardian and Telegraph newspaper websites for web pages reporting each story in the broadcast.
- Searches are restricted to the day of broadcast, or the day after.
- Searches are repeated using different combinations of the extracted key-phrases.
- The text of the returned web pages is compared with the text of the transcript to find matching stories.

Using the Web Pages

The web pages contain:

- A headline, summary and section for each story.
- Good quality text that is readable, and contains correctly spelt proper names.
- They give more in depth coverage of the stories.

Semantic Annotation

- The KIM knowledge management system can semantically annotate the text derived from the web pages:
- KIM will identify people, organizations, locations etc.
- KIM performs well on the web page text, but very poorly when run on the transcripts directly.
- This allows for semantic ontology-aided searches for stories about particular people or locations etcetera.
- So we could search for people called *Sydney*, which would be difficult with a text-based search.

Entity Matching

RichNews - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Refresh Print Mail New Tab

Address C:\tmp\RichNews\WebDemo\main_800x600.htm Go Links



THE UNIVERSITY OF SHEFFIELD
Department of
computer science




RichNews

Keys: Person Location Organization

buckingham palace in these lanes at all about flower shop [SIL] in **north wales** thought to be country first [s] as many statement from one family about the collapse of the strike [SIL] made the briefest of reactions from buckingham palace were [SIL] **the queen's** office said the outcome of the trial was entirely a matter for the judicial authorities [s] and there's a similar distance in comments from st **james's** palace [SIL] which is you know has established an inquiry [SIL] by prince charles' private secretary sir **michael** peat said edmund lawson q. c. [SIL] that into the conduct of [SIL] the prince of **wales's** household in events surrounding the ball trial collapsed [SIL] and the spokesman there said it was most unlikely that that inquiry would be wide and [SIL] pointing out that these **haroldbrown** case [SIL] revolves around the princess of **wales's** household not prince charles is it the michael's inquiry is limited [SIL] and internal [SIL] be delaying who support our farm shop in **north wales** the collapse of this case put his case was linked to the collapse [SIL] of the trial of **pau** bowen [SIL] has been there this morning as in any site of the money reaction from the bowels [SIL] plan that would be nice to know what the first but the things about a second but that wouldn't be outside bordeaux shoppers to you martha chest which was damaged by fire a couple of days ago **police** investigation still underway to [SIL] see whether that [SIL] relates it all [SIL] turn to **pau** bowles trial in the way he's been telling his story over the past two weeks [SIL] is what's inside them reading some of the cleaning up but to no sign of him were told [SIL] by friends of the family that he had of the country may well be back in



Ontotext
Knowledge and Language
Engineering Lab of Sofia



KIM
Knowledge and Information
Management Platform

Paul Burrell, a **Man**

Property	Value
hasMainAlias	Paul Bu
hasAlias	Burrell
hasAlias	Paul Bu

Copyright © 2004 Ontotext Lab, Sima AI, Bulgaria

My Computer

Story Retrieval

The screenshot shows a Microsoft Internet Explorer window titled 'KIM WEB UI - Microsoft Internet Explorer'. The address bar is empty. The menu bar includes File, Edit, View, Favorites, Tools, and Help. The toolbar contains Back, Forward, Stop, Home, Search, Favorites, Media, and Print. The main content area is divided into two sections: 'Document Detail' and 'Document Content'. The 'Document Detail' section contains a table with the following data:

Feature Name	Feature Value
TITLE	'Equipment missing' at Iraq arms site
SOURCE	One O'Clock News
UNIQUE_URL	file:///Z:/sins/richnews/mpegs/021203.mpg
TIMESTAMP	3/12/2002

The 'Document Content' section contains the following text:

United Nations weapons inspectors say equipment has gone missing from a missile factory in Bagh

"In 1998, the site contained a number of pieces of equipment tagged by the United Nations Special C cameras," the statement issued by the inspectors said.

It was claimed that some [equipment] had been destroyed by the bombing of the site

Hiro Ueki
Unscm inspector

"None of these are currently present at the facility."

Evaluation

- Success in finding matching web pages was investigated.
- Evaluation based on 66 news stories from 9 half-hour news broadcasts.
- Web pages were found for 40% of stories.
- 7% of pages reported a closely related story, instead of that in the broadcast.
- Results are based on earlier version of the system, only using BBC web pages.

Ongoing Improvements

- Use teletext subtitles (closed captions) when they are available
- Better story segmentation through visual cues and latent semantic analysis
- Use for content augmentation for interactive media consumption

RichNews demonstration

<http://gate.ac.uk/demos/prestospace-london/prestospace-london.html>

Business Intelligence: the MUSING project

The problem

- Business intelligence requires the collecting and merging of information from many different sources
- This is needed to analyse financial risks, operational risk factors, follow trends, perform credit risk management etc.
- Traditional data mining tools make use of numerical data and cannot easily be applied to knowledge extracted from free text
- Traditional IE is not adapted for the financial domain, or does not address the issue of information integration.
- Musing aims at the analysis of financial information and news about mergers and acquisitions

The solution

- Apply NLP techniques to transform unstructured sources into the structured knowledge more suitable for analysis
- content mining using domain-specific ontologies
- Enables extraction of relevant information to be fed into models for financial risk analysis and business intelligence
- Use of XBRL standard for business reporting, for information exchange

Merging information across different sources

- Framework makes use of a domain ontology
- Ontology acts as a bridge between text and a KB, which in turn feeds reasoning systems or provides info to end users.
- 2 main issues concerning identity resolution:
 - variation across sources
 - ambiguity across sources

Variation and Ambiguity

- [Johann Sebastian Bach](#) (1685–1750), composer and organist, the most well-known of the Bachs
- [Wilhelm Friedemann Bach](#) (1710–1784), composer and organist
- [Carl Philipp Emanuel Bach](#) (1714–1788), composer, harpsichordist and pianist
- [Johann Aegidus Bach](#) (1645–1716), organist and conductor
- [Edward Bach](#) (1886-1936), medical doctor known for his work in alternative medicine
- [Sebastian Bach](#) (born 1968), former lead singer of Skid Row

Information Extraction in MUSING

- Document format and structure analysis
- Linguistic pre-processing (tokenisation, splitting..)
- Information extraction:
 - gazetteer lookup
 - pattern matching rules for semantic analysis
- Export of annotations to database / ontology
- Different applications needed for recognising information from different sources

Company Profiles

- Require structured information from company profiles to
 - feed into statistical models of financial risk assessment or investment
 - provide services to companies looking for commercial partners in same sector in a different country
- e.g. system extracts the fact that Russia's investment Fitch rating is BBB+, increased from BBB
- Risk assessment model can then revise risk downwards

International Enterprise Intelligence application

- Provides customers with up-to-date information about companies, mined from different sources (web, financial news, structured data sources, etc.)
- Extract set of relevant concepts from company profiles downloaded from Yahoo!
- Each concept is associated with relevant information, e.g. “number of employees = 200”
- Also need to extract country and region information (population, currency etc) from CIA World Factbook

Extracting information from financial statements

- Information only available as pdf
- Other binary formats difficult to process automatically
- When a bank needs financial information, it has to be manually copied from the balance sheet and re-entered into the system
- Impossible to obtain key information that is not explicit
- *“What were the net assets of the company on 31 December 2001?”*

Processing balance sheets

- PDF is loaded into GATE and pre-processed
- Spatial and graphical information is partially lost, so analysis has to be performed on figures, e.g. identifying totals, based on positional information
- For each concept, features and their values are extracted, e.g.
 - <string = Total Current Liabilities>
 - <value = 73,000>
 - <year = 2005>

Web-based annotation tool

Annotator GUI [Connected to: ADVO Inc ___ 1169480433950 ___ 1452@gate.ac.uk/docservice/services/docservice]

Type | Set | Start | End | Features

ADVO Inc.

Class : **Company_Name** Del Del All

Apply Apply To All Cancel

DETAILS
 Index Membership: S&P 600 SmallCap
 S&P 1500 Super Comp
 Sector: Services
 Industry: Marketing Services
 Full Time Employees: 3,700

BUSINESS SUMMARY
 ADVO, Inc., a direct mail media company, engages in soliciting and processing printed advertising from retailers, manufacturers, and service companies in the United States and Canada. It offers direct mail marketing products and services, such as shared mail, which provides the addresses of the households receiving the mail packages; and sorts, processes, and transports the advertising material for ultimate delivery primarily through the United States Postal Service (USPS). The company also offers solo mail services, which includes list procurement, addressing, processing, and distribution of brochures and circulars for an individual customer through USPS. In addition, ADVO provides ancillary services to complement its direct mail marketing programs; offers private carrier delivery in various markets; and provides direct mail advertising solutions, as well as operates a consumer Web site, ShopWise.com that allows retailers to electronically target promotions and values to subscribers. ADVO serves

Ontology Tree Options

Jena Ontology_00016

- Company_ID_Number
- Former_Company_Names
- Company_Address
- Company_Website
- Company_Fax
- Company_Phone
- Money
- Financial_Assets
- Company_Information_Other
- Disclosure_of_businesses_of_t
- Main_Product
- Main_Services
- Stock_Exchange_Listings
- Industry_Sector
- Shares
- Net_Income_or_Net_Loss_for_t
- Company_Stakeholders_Details
- Person
- Market_Outlets
- Disclosure_of_shares_of_mem
- Parent_Company
- Disclosure_of_all_members_of

Document: ADVO Inc ___ 1169480433950 ___ 1452

KIM CORE Search

KIM CORE

- Co-Occurrence and Ranking of Entities Search
- Hybrid technology combining Semantic Web technology, information extraction and relational databases.
- Idea is to record information about the co-appearance of entities in the same context, which speaks of "soft" or "associative" relations between them
- This means you can narrow a search to something more specific
- Also can be used to calculate statistics about the popularity of entities in a given context, information sub-space and period.
- Technique is known as **timelines generation**

CORE Timelines demo

- Allows the tracking of trends and tendencies, and the association of the each point in the timeline with a set of documents forming it
- Allows the navigation from the timeline to the documents, where the events forming the peaks or drops are evident.
- <http://people.aifb.kit.edu/dvr/videos/kimsearch.html>

GATE Mimir

<http://vimeo.com/11334635>

What to do with annotations

- GATE applications tend to produce LOTS of annotations
- There are lots of things you can do with them
 - Export GATE documents to XML
 - Custom PRs may export data
 - or you can use them to search the documents

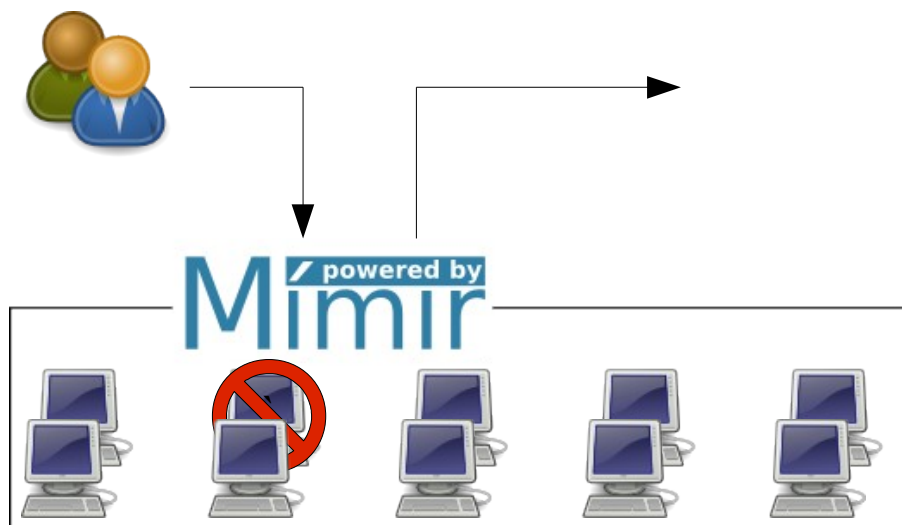
Mímir: The Big Idea

- Multi-paradigm Information Management Index and Repository
- Mímir is an IR engine that can search over:
 - text
 - semantic annotations
 - ontologies and KBs
- Built on top of
 - the MG4J text indexing engine
 - GATE's annotation index
- Scales to millions of documents

Mimir: Indexing

- For large scale annotation and indexing tasks, we have the GATE Cloud Paralleliser (GCP)
- GCP can run multiple instances of an application on a single machine
- GCP can be run on multiple machines to spread the load or to reduce processing time
- GCP is configured via XML files and can process documents directly from ARC files and send them direct to an open Mimir index

Mimir: Indexing



- Mimir supports federated indexes – an index that consists only of sub-indexes
- A sub index can be removed or replaced
- New indexes can be added at any time
- This allows for the gradual update of the index when new annotations are added or when improvements are made

Mimir: Querying

- Traditional search engines (e.g. Google) treat queries as a bag-of-words
- Documents that contain any or all of the words in any order are considered as matching the query
- Mimir always treats the query as a sequence
- Each result represents one instance where the sequence exactly matches the document

Mimir: Querying via GUS

The screenshot shows a web browser window with the URL `demos.gate.ac.uk/mimir/gpd/search/gus#page=1`. The page features a search bar with the query `{Person} root:say` and a 'Search' button. Below the search bar, a blue banner indicates 'Searching index: News Demo'. The results section, titled 'Results 1 - 10 of 3,681', lists several news items:

- Germany finds bailing out is hard to do (cached)**
so. SPD leader Frank **Walter Steinmeier said** his party would be willing
- How to sell music the Proper way (cached)**
managing director Stephen Kersley. **Graham Jones says** record shops will never die
- How to sell music the Proper way (cached)**
" Back to basics Mr **Mills says** he has no need to
- The demise of the Elgar £20 note (cached)**
no longer legal tender, **Mrs Cleland says**. Manufacturers of note-taking machines
- BT expands super-fast broadband network (cached)**
," he added. **Ms Garfield said** that she was unaware of
- Gove spells out education priorities for 'a new era' (cached)**
Education correspondent, BBC News **Michael Gove says** this is a new era
- IPCC's Pachauri says climate body must 'listen and learn' (cached)**
2007 Fourth Assessment report, **Dr Pachauri said**: "We have not

Using SPARQL to Restrict A Query

- As well as text and annotations Mimir queries can include SPARQL to restrict against an ontology
- SPARQL is embedded in a query using the synthetic “sparql” feature of the annotation you wish to restrict
- This is most helpful when the annotations are already linked to an ontology, probably via the Large Knowledge Base (LKB) Gazetteer.

Sparql Query for “people born In Sheffield”

```
{Person sparql = "SELECT ?inst WHERE { ?  
inst :birthPlace  
<http://dbpedia.org/resource/Sheffield>}"  
}
```

Sparql query for the location of steel industries

```
{Organization sparql = "SELECT ?inst  
WHERE { ?inst :industry  
<http://dbpedia.org/resource/Steel>} "  
[0..4] in {Location}}
```

Creating SPARQL Constraints

- You can develop SPARQL queries independently from the Mimir queries.
- Try issuing a couple of SPARQL queries (see previous slides) directly against
 - SKB: <http://skb.ontotext.com/sparql>
 - Dbpedi: <http://dbpedia.org/sparql>

Query Interfaces

- Useful Mimir queries are complex!
- The query syntax allows for unrestricted search
- Custom built interfaces could take the pain out of generating complex queries
 - Calendar controls for date constraints
 - A globe image for location restriction
 - ...

Using Mimir can be RESTful!

- As well as GUS, the Mimir web app supports an XML-based RESTful interface
 - this interface supports the same query syntax
 - allows access to all result information
 - is easy to use
 - can be used to build custom interfaces

Customised Querying

The screenshot shows a web browser window with the following elements:

- Browser Tab:** Nightly | People in the News
- Address Bar:** demos.gate.ac.uk/pin/?name=&bornIn=Sheffield&famousAs=Politician|OfficeHolder&aft...
- Search Bar:** Google
- Header:** PEOPLE IN THE NEWS (with a world map background)
- Search Filters:**
 - Looking For...**
 - Name: [Empty text box]
 - Fuzzy Name Matching:
 - Born In: Sheffield
 - Famous As: Politician
 - In Articles...**
 - Published Between: 01/04/2011 and 30/04/2011
 - Classified As: Scotland
 - Ignore Boilerplate Text:
- Search Button:** Search
- Results Bar:** Results 1 to 2 of 2 | Show Underlying Mimir Query
- Search Results:**
 - [Scottish election: Respect Coalition Against Cuts profile](http://www.bbc.co.uk/news/uk-scotland-13048761)
 - ... Bow - whose sitting MP Oona King had voted for the war ...
 - ... success came when Galloway overturned Oona King's 10,000- ...
- Page-Footer:** Powered by GATE Mimir | © The University of Sheffield, 2011

<http://demos.gate.ac.uk/pin/>

Demo: Adding Semantic Search to BBC News Articles

The Premise

- Use multiple GATE technologies to...
 - Build a GATE application to process BBC news articles
 - Populate a Mimir index to enable multi-paradigm search of the annotated articles

Start with ANNIE

- Use ANNIE for linguistic pre-processing and NE Recognition
 - Sentence Splitting
 - Tokenisation
 - Named Entity Recognition
 - Co-reference
 -
- ANNIE is almost always a good starting point when developing a new GATE application

Extend The Application

- To ANNIE we added
 - Date Normalisation
 - Measurements
 - LKB (Large Knowledge Base Gazetteer)
 - BoilerPipe Content Detection
- The LKB was initialised using DBpedia
 - Used to annotate Person, Organization and Locations wrt DBpedia
 - All relevant entities are thus associated with a URI

Extend The Application

- These extensions allow us to
 - search for a number of new types/features
 - link existing types to an ontology

We could have stopped at this point and still had a useful application, but...

BBC Classification

- Each BBC news article contains a classification (a label stating which section of the BBC website the article is published under)
- A simple JAPE grammar can extract the classifications for an article
- These annotations can be linked to a simple ontology (built from within GATE)
- Provide another axis on which the resulting annotations can be searched

BBC Classification

The screenshot shows the GATE (General Architecture for Text Engineering) software interface. On the left is a sidebar with a tree view of applications and resources. The main window is titled 'Messages' and 'bbc-classificat...'. It features two panes: 'Classes & Instances' and 'Properties'. The 'Classes & Instances' pane shows a hierarchical tree of ontology classes, with 'Classification' selected. The 'Properties' pane shows details for the selected class, including its URI, type, and property values.

Class	Instance
Also_in_the_News	Also_in_the_News
Asia_Pacific	Asia_Pacific
Business	Business
Economy	
Edinburgh	
England	
Entertainment_Arts	
Europe	
Fife_East	
Glasgow_West	
Have_Your_Say	
Health	
Highlands_Islands	
In_Pictures	
Latin_America	
Magazine	
Mid	
Middle_East	
NE	
N_Ireland	
North_East	
North_West	
Orkney_Shetland	
Politics	
Sci_Environment	
Scotland	
Scotland_Business	
South	
South_Asia	
South_East	
South_West	
Special_Reports	
Tayside_Central	
Technology	
UK	
US_Canada	
Wales	
World	
Your_Money	

The screenshot shows the BBC News website for the 'Highlands & Islands' region. The page features a navigation bar with various news categories and a search bar. The main headline is 'Northern Constabulary reports crime rate drop', dated 13 May 2011. Below the headline is a sub-headline: 'Crime in the Highlands and Islands fell by 4% last year, equating to 568 fewer victims, Northern Constabulary said.' To the right, there is a 'Top Stories' section with a photo of a man and a headline: 'IMF head charged over 'sex crime''. Other top stories include 'Nato 'must widen' Libya targets' and 'Sony begins PlayStation relaunch'.

The Final Application

The screenshot displays the GATE (General Architecture for Text Engineering) application interface. The main window shows a document titled "How to sell mus..." with various NLP annotations applied to the text. The annotations include entities like "Robert Plummer", "BBC News", "Malcolm Mills", "UK", "Big Four", "Joan Baez", "Nick Lowe", "Ella Fitzgerald", "John Coltrane", "Chris Anderson", and "Mr Mills". The text is color-coded according to these annotations.

On the left side, there is a "Processing Resources" panel listing various tools and modules:

- Ontology Coreference
- ANNIE OrthoMatcher
- Measurement Tagger
- ANNIE+ NE
- Document Info Extractor
- Date Normalizer
- Document Date Analyzer
- ANNIE NE Transducer
- Convert LKB Lookups
- DBpedia Gazetteer
- ANNIE Gazetteer
- Number Tagger
- Morphological Analyser
- POS Tagger
- Sentence Splitter
- Content Detection
- Tokeniser
- Document Reset

At the bottom left, a table shows the current document's metadata:

MatchesAnnots	{null=[[6365, 636
MimeType	text/html
gate.SourceURL	http://www.bbc.c
liveURL	http://www.bbc.c
normalized-date	20100429

On the right side, there is a "Text" panel with a list of categories and their corresponding colors. The categories are:

- Content (Red)
- Date (Green)
- Document (Yellow)
- DocumentClassification (Blue)
- DocumentDate (Pink)
- DocumentTitle (Cyan)
- FirstPerson (Red)
- Location (Green)
- Lookup (Yellow)
- Measurement (Blue)
- Money (Pink)
- Number (Cyan)
- Organization (Red)
- Person (Green)
- Ratio (Red)
- Sentence (Green)
- SpaceToken (Yellow)
- Split (Purple)
- Token (Pink)
- Unknown (Cyan)
- Original markups (Blue)

The bottom of the window shows "Document Editor" and "Initialisation Parameters" tabs, and a "New" button.

GCP and Mimir

- We downloaded 8,255 BBC news articles
- We used the GATE Cloud Paralleliser to...
 - annotate the articles using the application
 - push the resulting annotations into Mimir

Mímir

- This resulted in a Mímir index of
 - 8,255 documents
 - 13 annotation types
 - 2 ontologies

People Born In Sheffield

Nightly GUS - GATE Unified Search - {P...
http://services.gate.ac.uk/mimir2/gpd/search/gus#page=1 Google

Search powered by Mimir

Searching index: News Demo

```
{Person sparql = "SELECT ?inst WHERE { ?inst :birthPlace <http://dbpedia.org/resource/Sheffield>"}"}
```

Search

Results 1 - 10 of 95

- What is a rating agency? (cached)**
was started in 1909 by **John Moody**, who published an analysis
- Oona King's knife crime pledge in mayoral candidate bid (cached)**
BBC News - **Oona King's** knife crime pledge in
- Oona King's knife crime pledge in mayoral candidate bid (cached)**
reddit StumbleUpon Twitter Email Print **Oona King's** knife crime pledge in
- Oona King's knife crime pledge in mayoral candidate bid (cached)**
pledge in mayoral candidate bid **Ms King** lost her parliamentary seat to

Location of Steel Industry

Nightly ▾ GUS - GATE Unified Search - {O...

← → http://services.gate.ac.uk/mimir2/gpd/search/gus#page=1 ☆ ↻ Google 🔍

Search powered by Mimir

Searching index: News Demo

```
{Organization sparql = "SELECT ?inst WHERE { ?inst :industry <http://dbpedia.org/resource/Steel>}"} [0..4] in {Location}
```

Search

Results 1 - 10 of *unknown*

Spending cuts 'to hit north harder' (cached)
Your stories Teesside voices The **Corus steel works in Redcar** announced 1,600 job

Spending cuts 'to hit north harder' (cached)
's large industrial employers, **Corus steel works in Redcar**, announced the partial mothballing

Page 1

Mimir 3.2.0-snapshot , © GATE 2011.

A Labour Party member being quoted in a document written in 2011 and classified as Scotland by the BBC

The screenshot shows a web browser window with the following elements:

- Browser Tab:** Nightly | GUS - GATE Unified Search - {{P...
- Address Bar:** <http://services.gate.ac.uk/mimir2/gpd/search/gus#page=1>
- Page Header:** Search powered by Mimir
- Search Bar:** Searching index: News Demo
- Query Input:**

```
{Person sparql = "SELECT ?inst WHERE { ?inst :party <http://dbpedia.org/resource/Labour_Party_%28UK%29>}"} root:say IN ({Document date > 20110000} OVER {DocumentClassification sparql = "SELECT ?inst WHERE { ?inst a bbc:Classification . FILTER (?inst = bbc:Scotland)"}")
```
- Search Button:** Search
- Results Section:** Results 1 - 10 of *unknown*
- Result 1:** Trident nuclear fleet cuts ruled out by Liam Fox (cached)
bn. Former Prime Minister **Gordon Brown said** in 2009 that he would
- Result 2:** Trident nuclear fleet cuts ruled out by Liam Fox (cached)
bn. Former Prime Minister **Gordon Brown said** in 2009 that he would
- Result 3:** Councils set out budget savings (cached)
'priorities' Council leader **Gordon Matheson said**: "I've always
- Result 4:** Scottish election: Green policy and business on agenda (cached)
businesses. And former chancellor **Alistair Darling said** Labour would put families first

BBC News Demos

- MIMIR demo: <http://demos.gate.ac.uk/mimir2/gpd/search/gus>
- PIN interface demo <http://demos.gate.ac.uk/pin/>

Summary

- In this module, we have seen how the various techniques can be implemented and used in real life applications
- In particular, we see how text mining can be used to make common tasks easier by
 - providing better or faster ways of searching for specific information
 - merging information from different sources to give a more accurate picture
 - adding semantics to the information to relate it with known existing information or to provide disambiguation