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

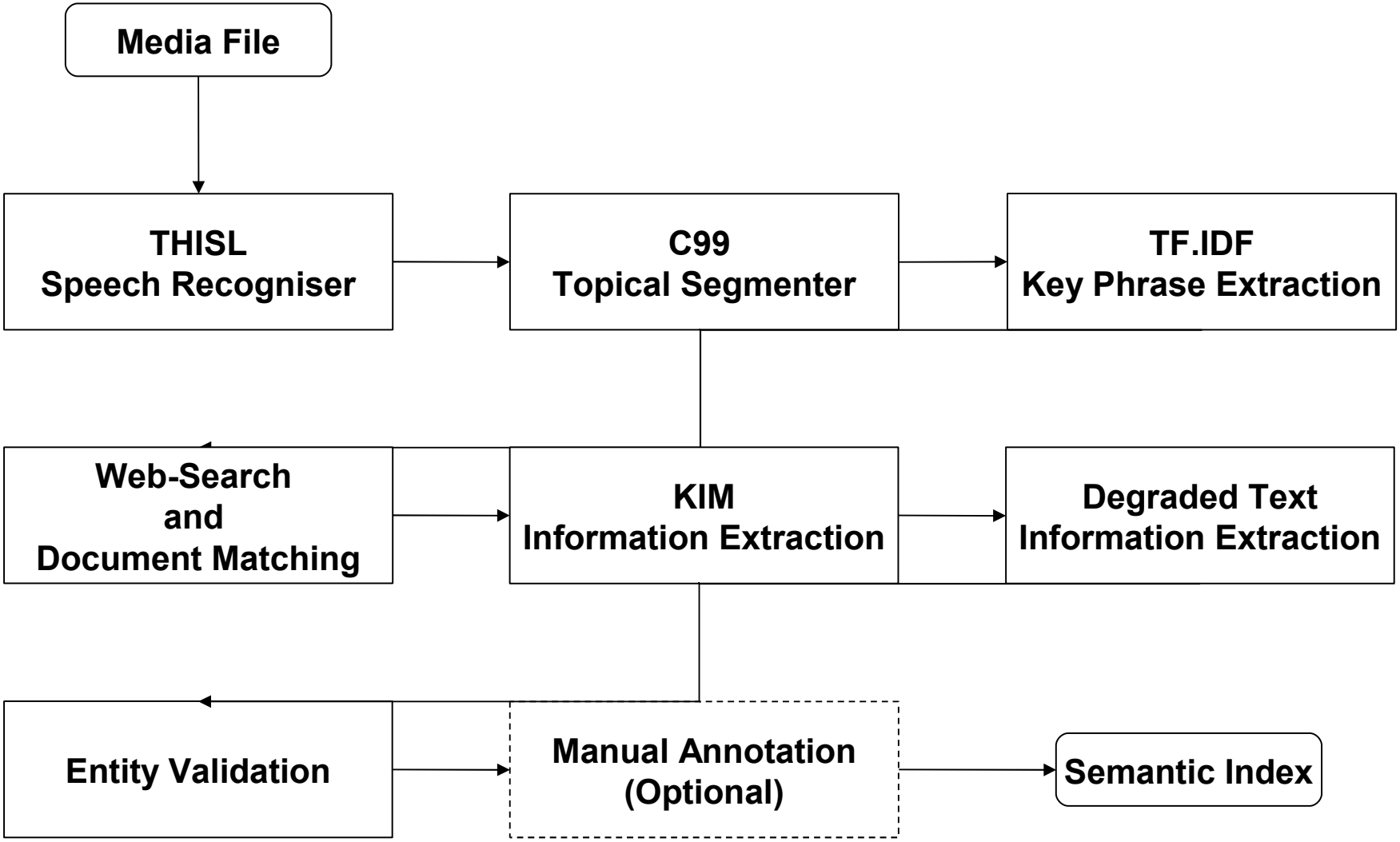
Obtaining a transcript:

- 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:

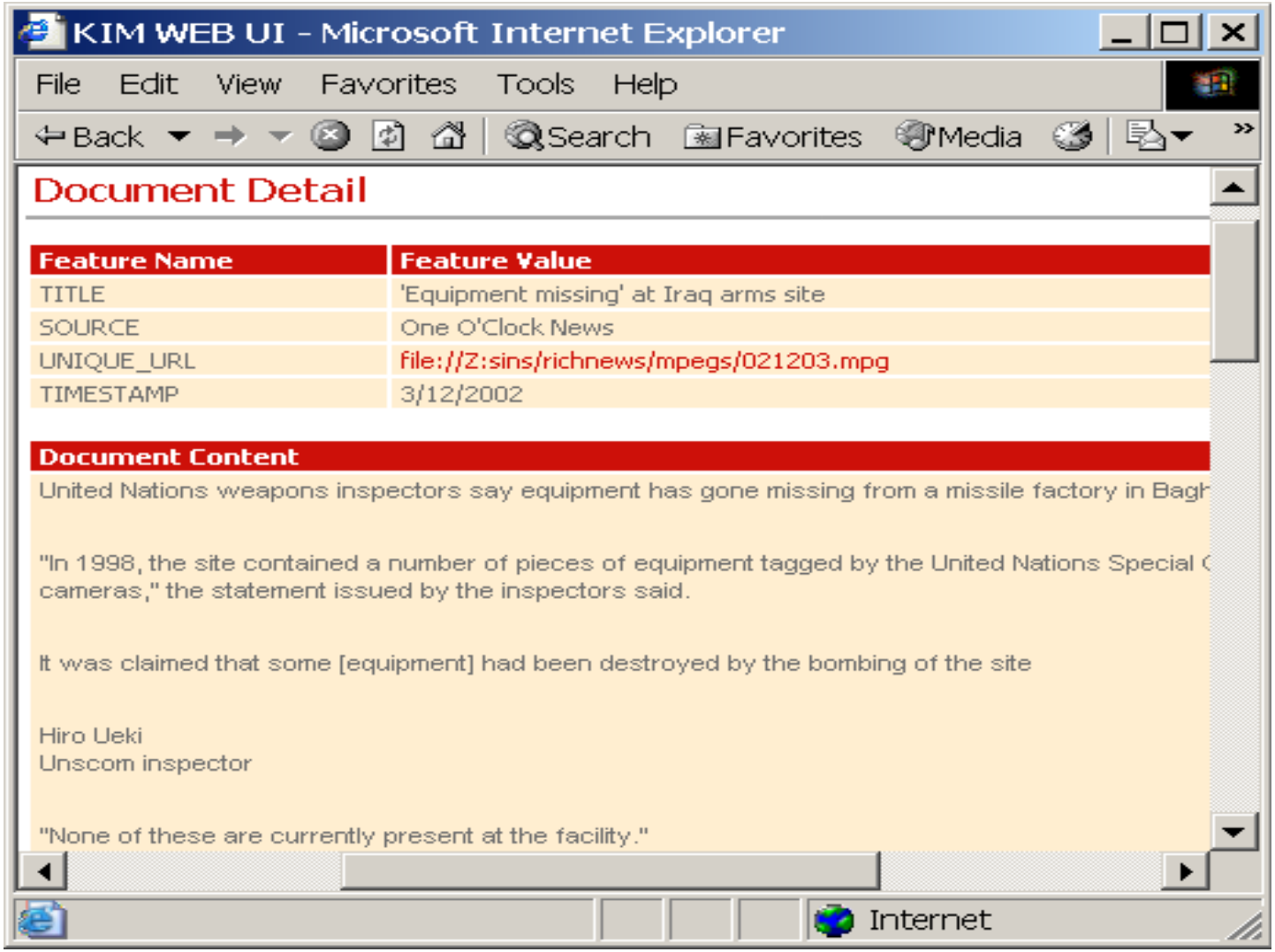
- A headline, summary and section for each story.
- Good quality text that is readable, and contains correctly spelt proper names.
- They give more in depth coverage of the stories.

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

Story Retrieval



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

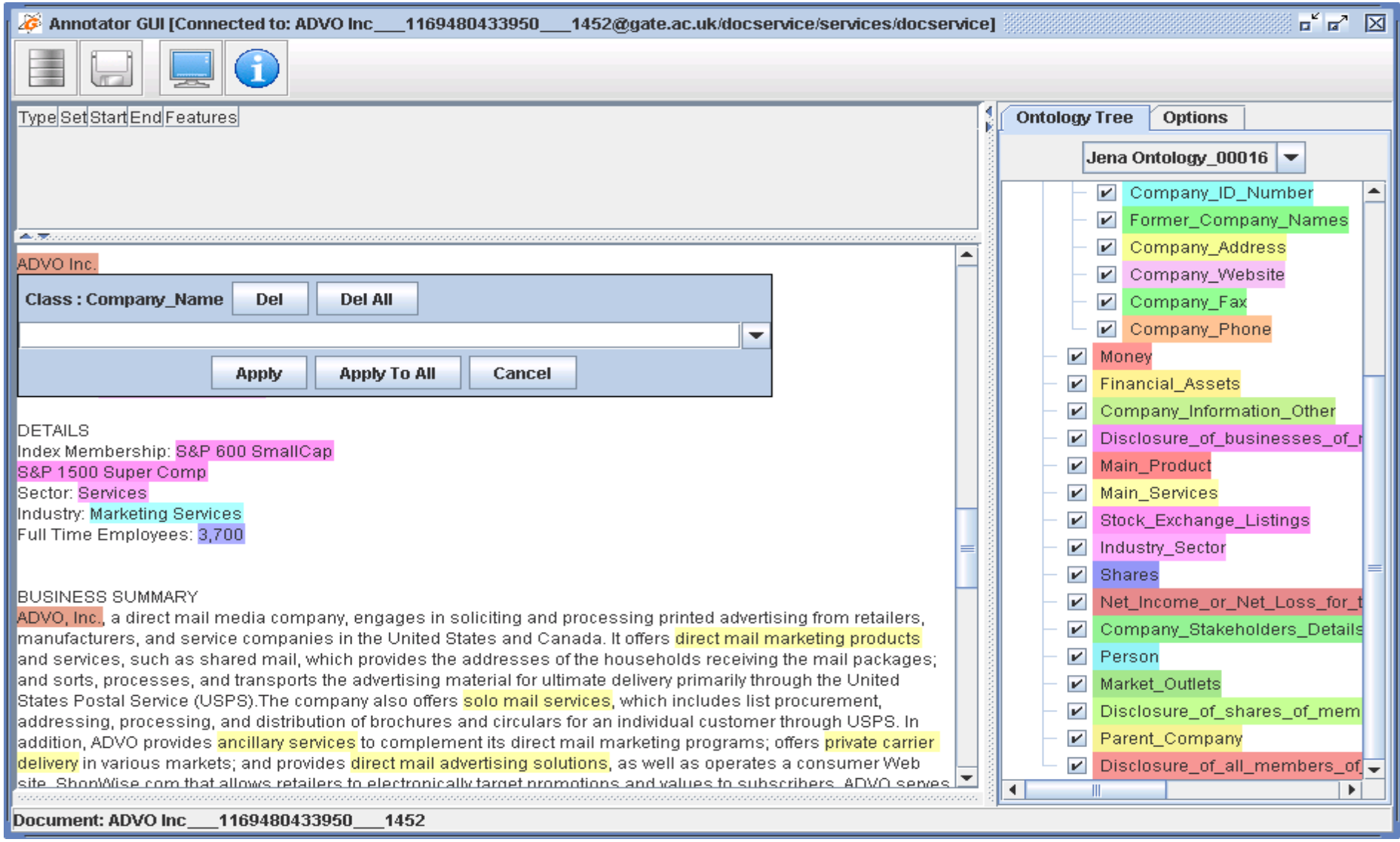
Extracting information from financial statements

- 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



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 Mímir

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

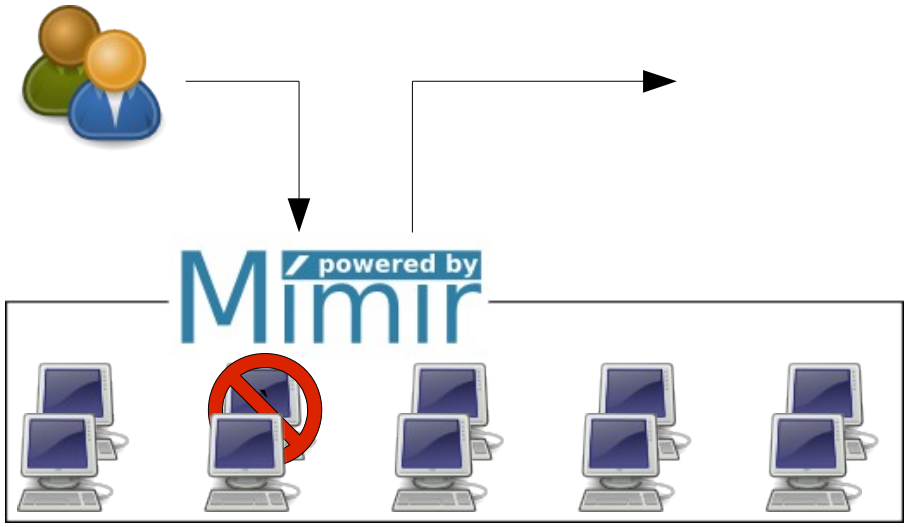
Mímir: The Big Idea

- 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

Mímir: 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 Mímir index

Mímir: Indexing

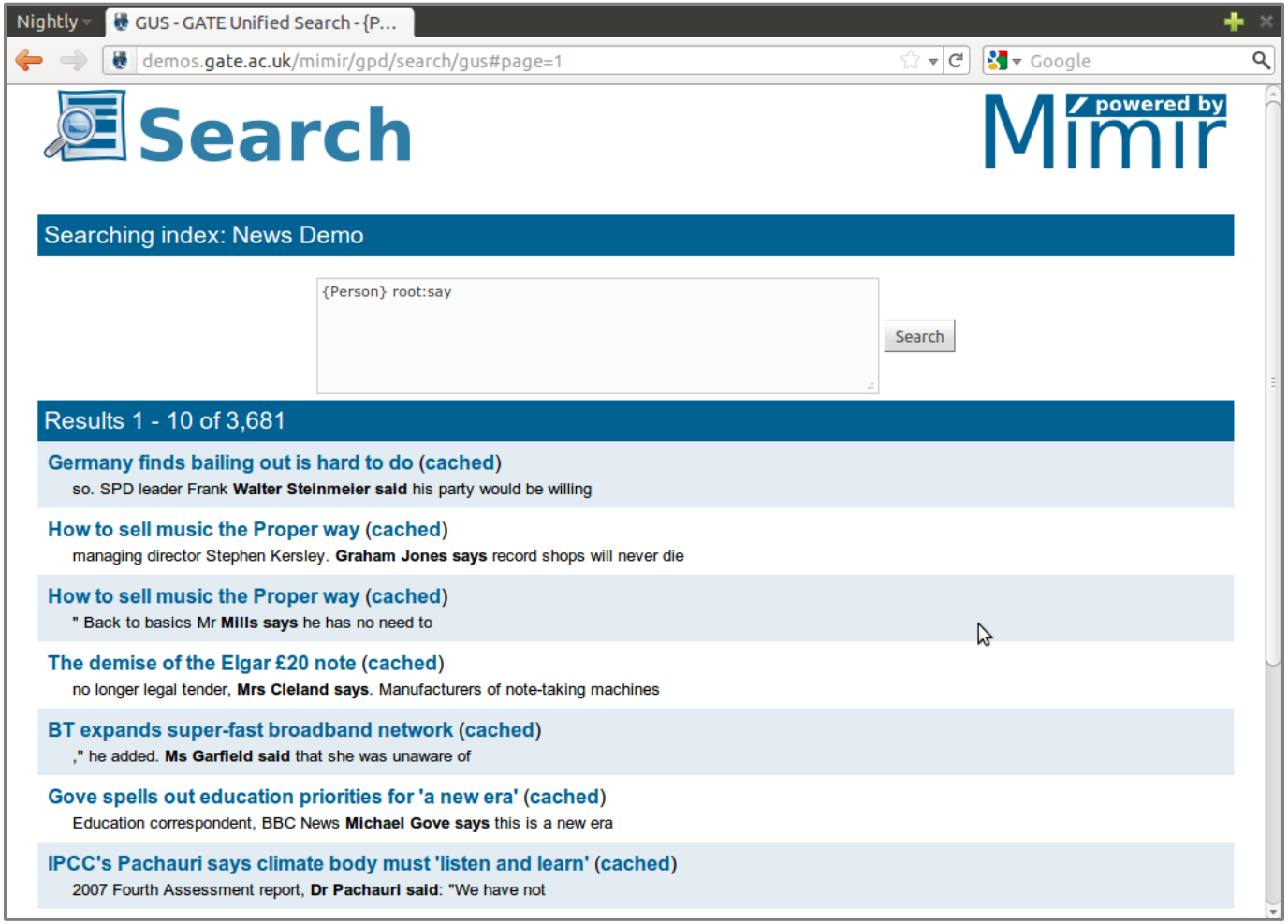


- Mímir supports federated indexes – an index that consists only of sub-indexes
- A sub index can be removed or replaced
- New indexes can be added at any time
- This allows for the gradual update of the index when new annotations are added or when improvements are made

Mímir: 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
- Mímir always treats the query as a sequence
- Each result represents one instance where the sequence exactly matches the document

Mímir: Querying via GUS



Using SPARQL to Restrict A Query

- As well as text and annotations Mímir 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 Mímir 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 Mímir can be RESTful!

- As well as GUS, the Mímir 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

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 Mímir index to enable multi-paradigm search of the annotated articles

Start with ANNIE

- Use ANNIE for linguistic pre-procesing 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

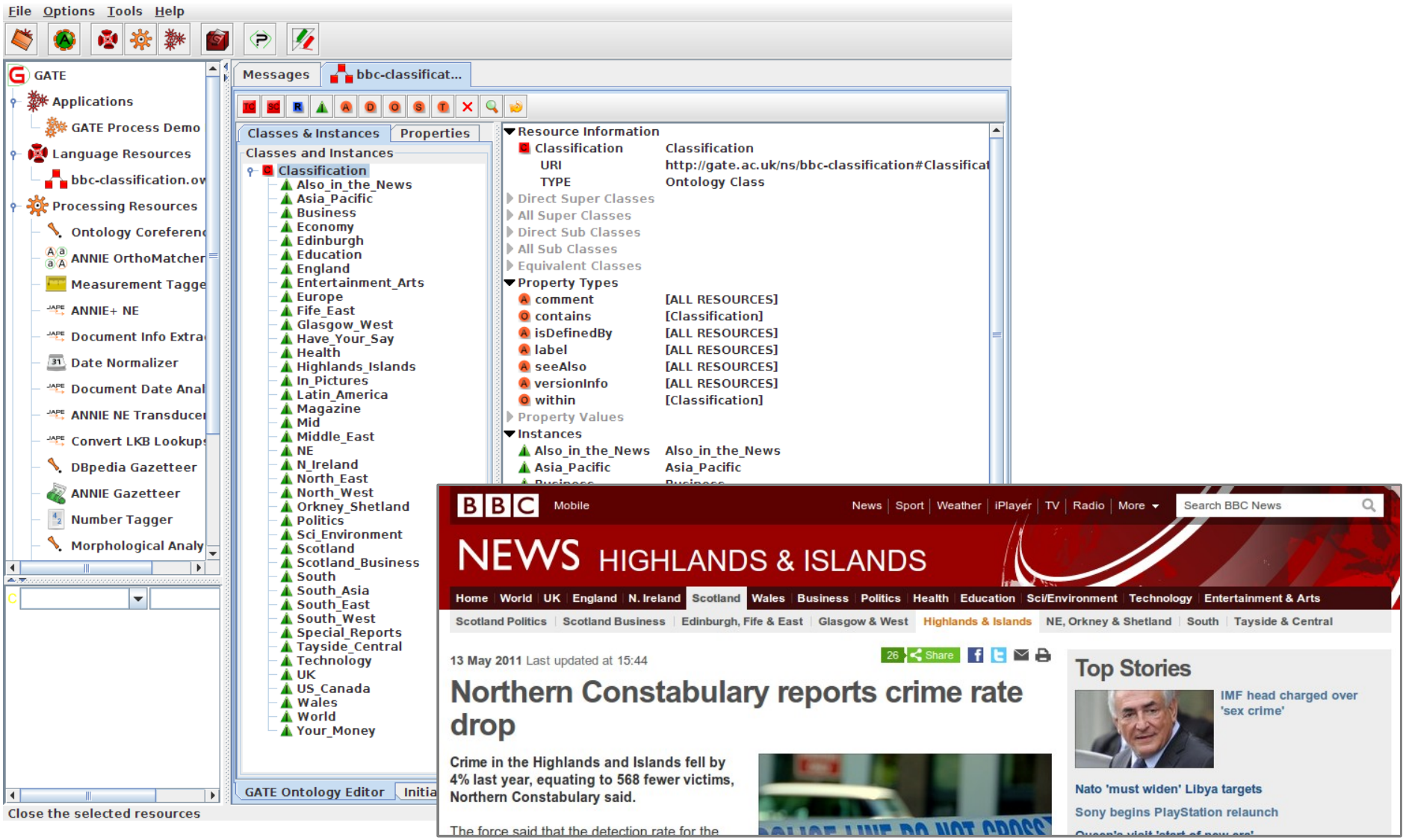
- These extensions allow us to
  - search for a number of new types/features
  - link existing types to an ontology

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

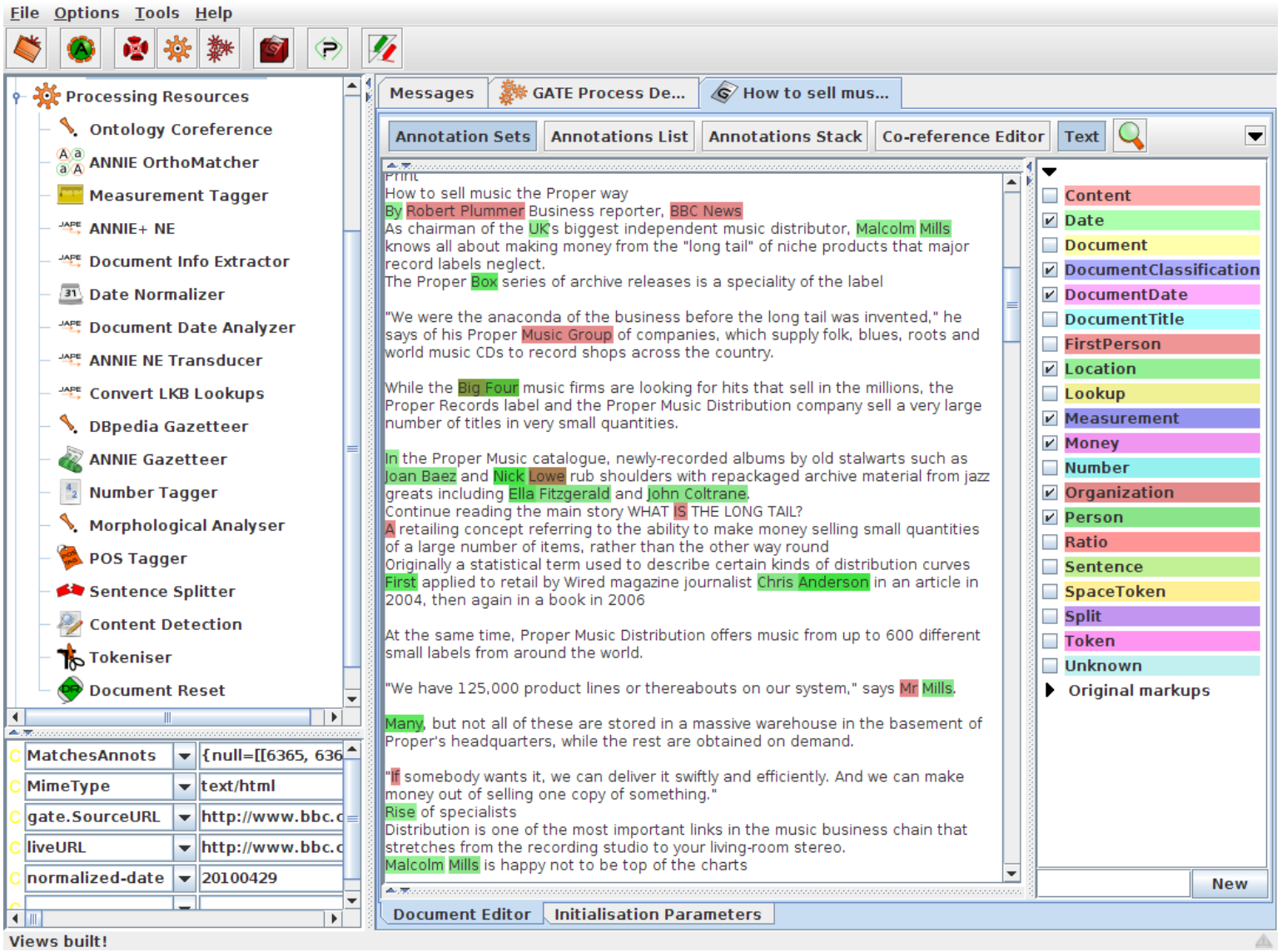
BBC Classification

- 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 Final Application



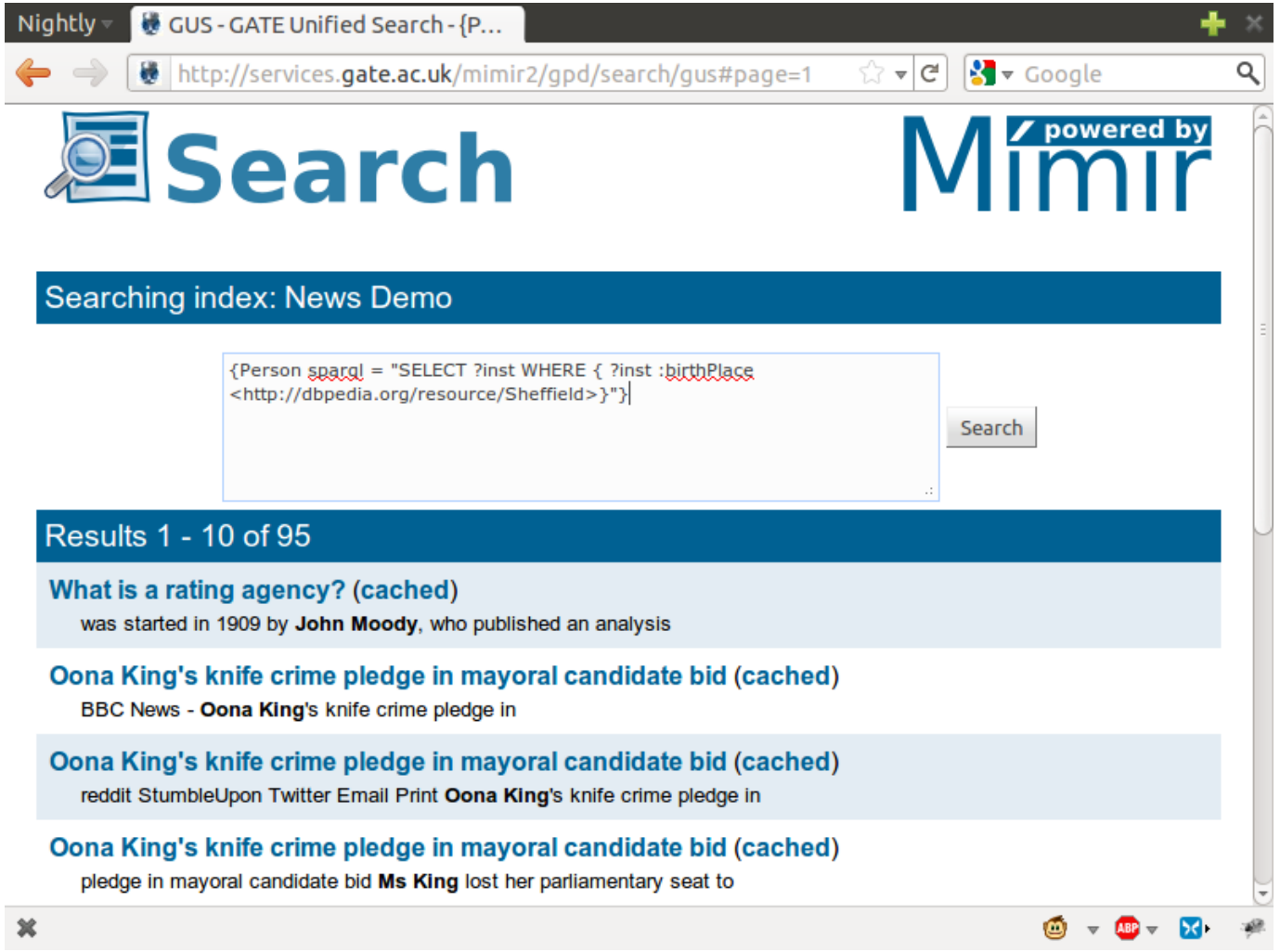
GCP and Mímir

- 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 Mímir

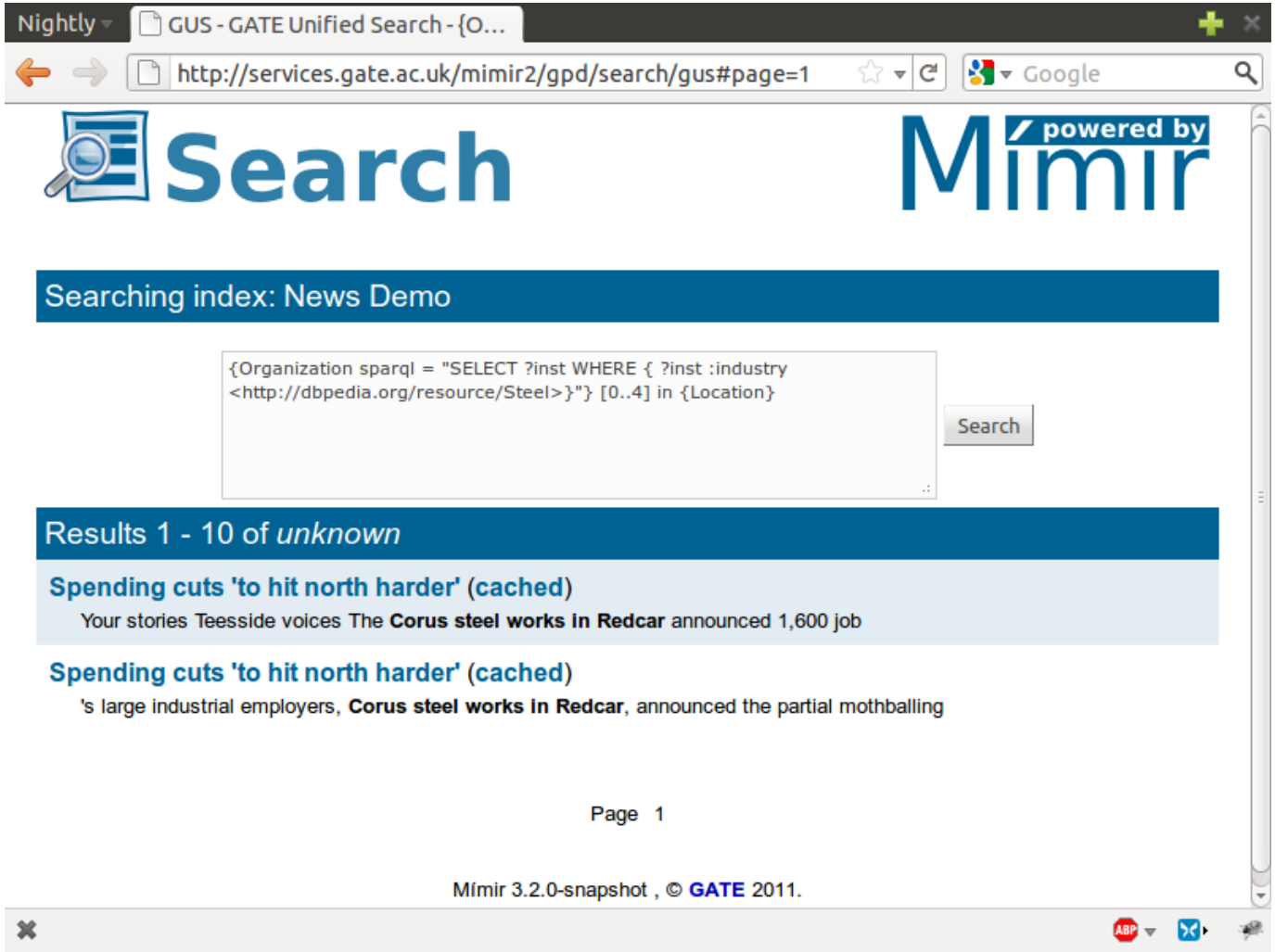
Mímir

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

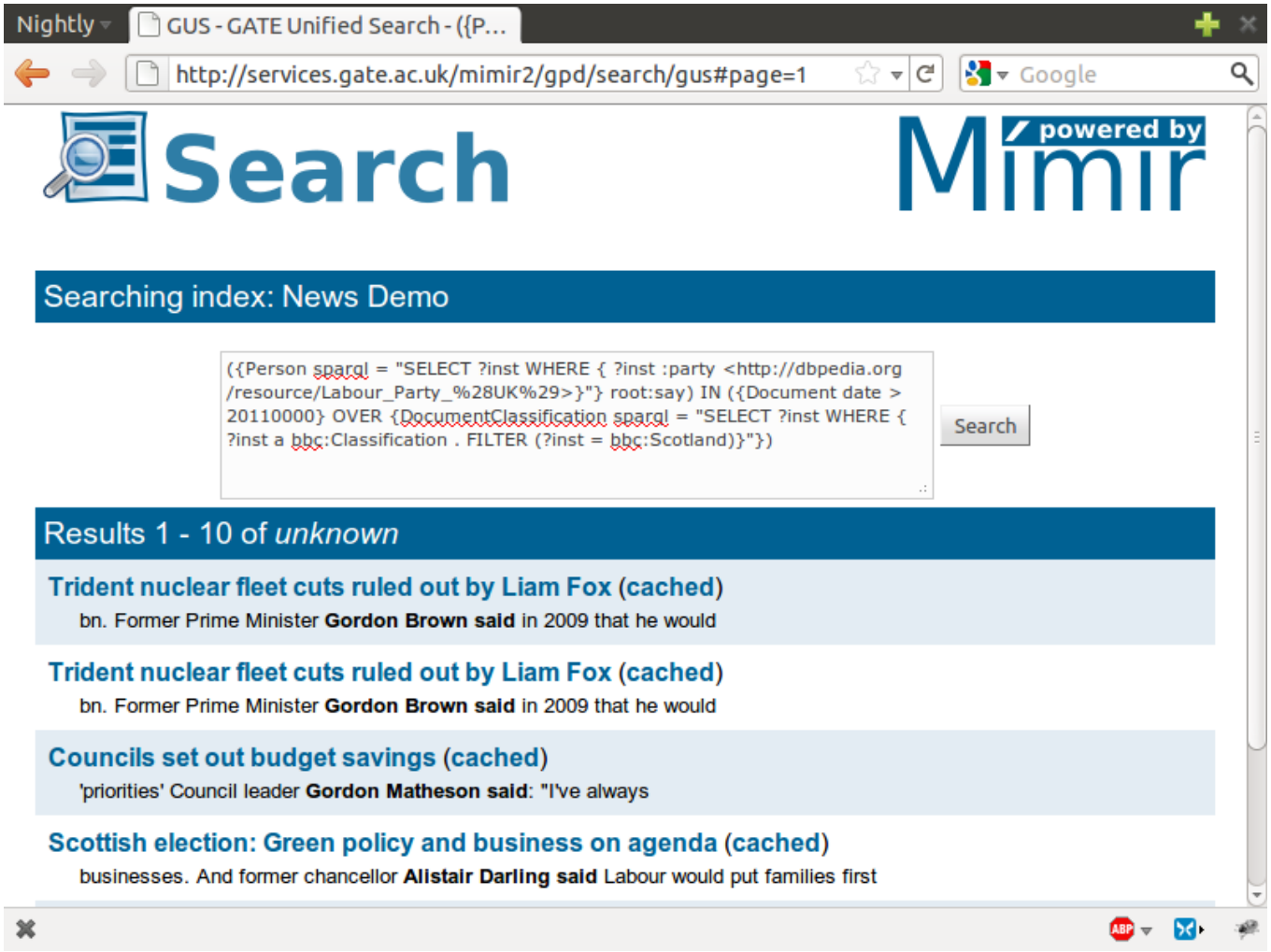
People Born In Sheffield



Location of Steel Industry



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



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