Information Extraction and GATE

Valentin Tablan
University of Sheffield
Department of Computer Science
NLP Group

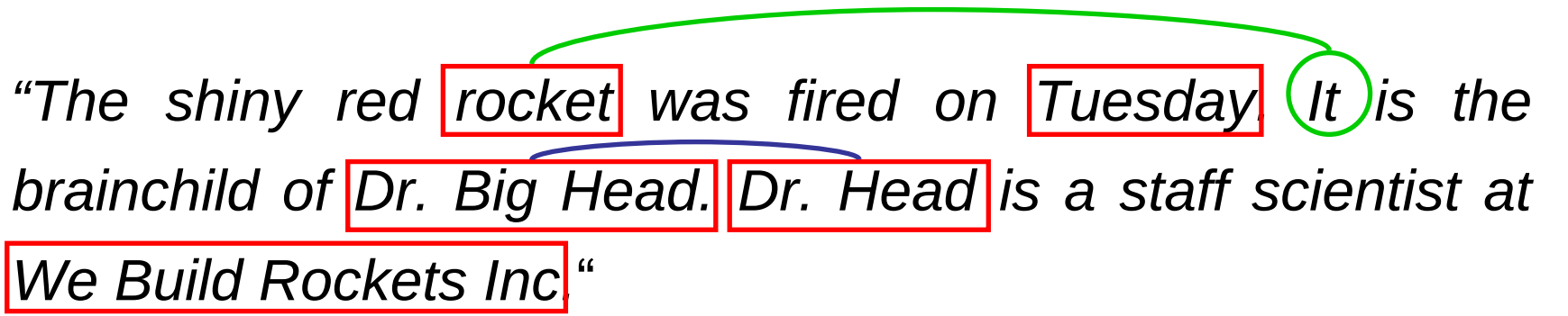
Information Extraction

- Information Extraction (IE) pulls facts and structured information from the content of large text collections.
- IR - IE - NLU
- MUC: Message Understanding Conferences
- ACE: Automatic Content Extraction

MUC-7 Tasks

- NE: Named Entity recognition and typing
- CO: co-reference resolution
- TE: Template Elements
- TR: Template Relations
- ST: Scenario Templates

An Example



- NE
- CO
- TE: the rocket is "shiny red" and Head's "brainchild".
- TR: Dr. Head works for We Build Rockets Inc.
- ST: a rocket launching event occurred with the various participants

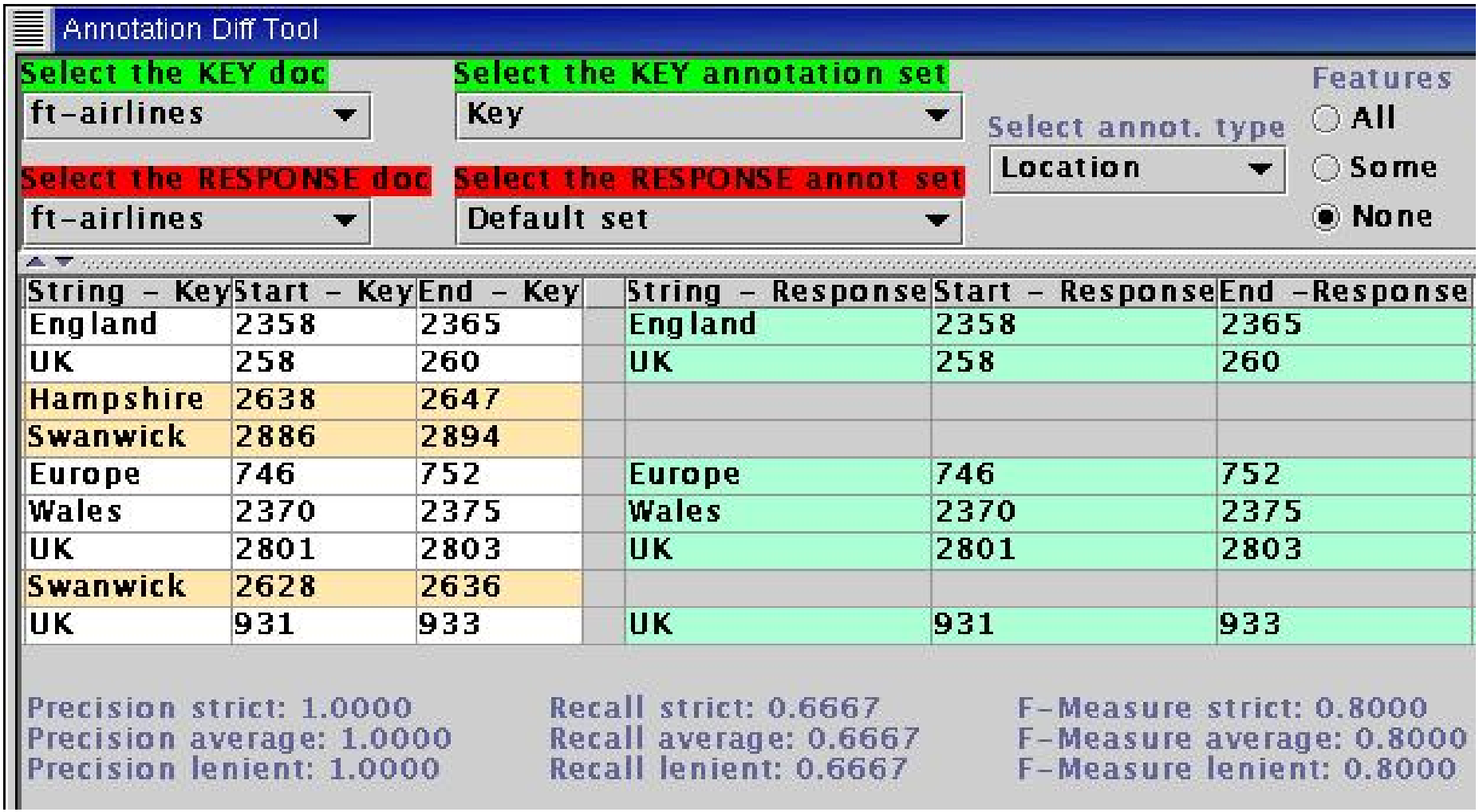
Performance Levels

- Vary according to text type, domain, scenario, language
- NE: around 97% (tested in English, Spanish, Japanese, Chinese)
- CO: 60-70% resolution
- TE: 80%
- TR: 75-80%
- ST: 60% (but: human level may be only 80%)

Evaluation

- Precision = correct answers/answers produced
- Recall = correct answers/total possible correct answers
- F-Measure = $\frac{2PR}{P + R}$ $\left(\frac{(\beta^2 + 1)PR}{\beta^2 P + R} \right)$

Evaluation


Annotation Diff Tool
Select the KEY doc
ft-airlines
Select the KEY annotation set
Key
Features
All
Some
None
Select annot. type
Location
Select the RESPONSE doc
ft-airlines
Select the RESPONSE annot set
Default set
String - Key
Start - Key
End - Key
String - Response
Start - Response
End -Response
England 2358 2365 England 2358 2365
UK 258 260 UK 258 260
Hampshire 2638 2647
Swanwick 2886 2894
Europe 746 752 Europe 746 752
Wales 2370 2375 Wales 2370 2375
UK 2801 2803 UK 2801 2803
Swanwick 2628 2636
UK 931 933 UK 931 933
Precision strict: 1.0000
Precision average: 1.0000
Precision lenient: 1.0000
Recall strict: 0.6667
Recall average: 0.6667
Recall lenient: 0.6667
F-Measure strict: 0.8000
F-Measure average: 0.8000
F-Measure lenient: 0.8000

A Typical IE System

1. Pre-processing
   - format detection
   - tokenisation
   - word segmentation
   - sense disambiguation
   - sentence splitting
   - POS tagging
2. Named entity detection
   - entity detection
   - coreference
3. Event detection
   - syntactic analysis
   - template filling
   - template merging
   - template relations
   - events detection

Two Approaches

Knowledge Engineering

- rule based
- developed by experienced language engineers
- make use of human intuition
- obtain marginally better performance
- development could 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
- requires large amounts of annotated training data
- some changes may require re-annotation of the entire training corpus

Named Entity Detection – more detail

- NE involves identification of proper names in texts, and classification into a set of predefined categories of interest.
- Three universally accepted categories: person, location and organisation
- Other common tasks: recognition of date/time expressions, measures (percent, money, weight etc), email addresses etc.
- Other domain-specific entities: names of drugs, medical conditions, names of ships, bibliographic references etc.

Basic Problems in NE

- Variation of NEs – e.g. John Smith, Mr Smith, John.
- Ambiguity of NE types:
- John Smith (company vs. person)
- May (person vs. month)
- Washington (person vs. location)
- 1945 (date vs. time)
- Ambiguity with common words, e.g. sentence initial “May"

More Complex Problems in NE

- Issues of style, structure, domain, genre etc.
- Punctuation, spelling, spacing, formatting, ... all have an impact:

```
Dept. of Computing and Maths
Manchester Metropolitan University
Manchester

United Kingdom
```

```
> Tell me more about Leonardo

> Da Vinci
```

List Lookup Approach

- System that recognises only entities stored in its lists (gazetteers).
- Advantages - Simple, fast, language independent, easy to retarget
- Disadvantages - collection and maintenance of lists, cannot deal with name variants, cannot resolve ambiguity

Shallow Parsing Approach (internal structure)

Internal evidence – names often have internal structure. These components can be either stored or guessed, e.g. location:

```
Cap. Word + {City, Forest, Center}
e.g. Sherwood Forest
Cap. Word + {Street, Boulevard,
Avenue, Crescent, Road}
e.g. Portobello Street
```

Shallow Parsing Approach (context)

External evidence - names are often used in very predictive local contexts, e.g. location:

- `"to the" COMPASS "of" CapWord`

e.g. to the south of London

- `"based in" CapWord`

e.g. based in London

- `CapWord "is a" (ADJ)? GeoWord`

e.g. London is a friendly city

Problems with Shallow Parsing

- Ambiguously capitalised words (first word in sentence)

  `[All American Bank] vs. All [State Police]`

- Semantic ambiguity

  `"John F. Kennedy"` = airport (location)

  `"Philip Morris"` = organisation

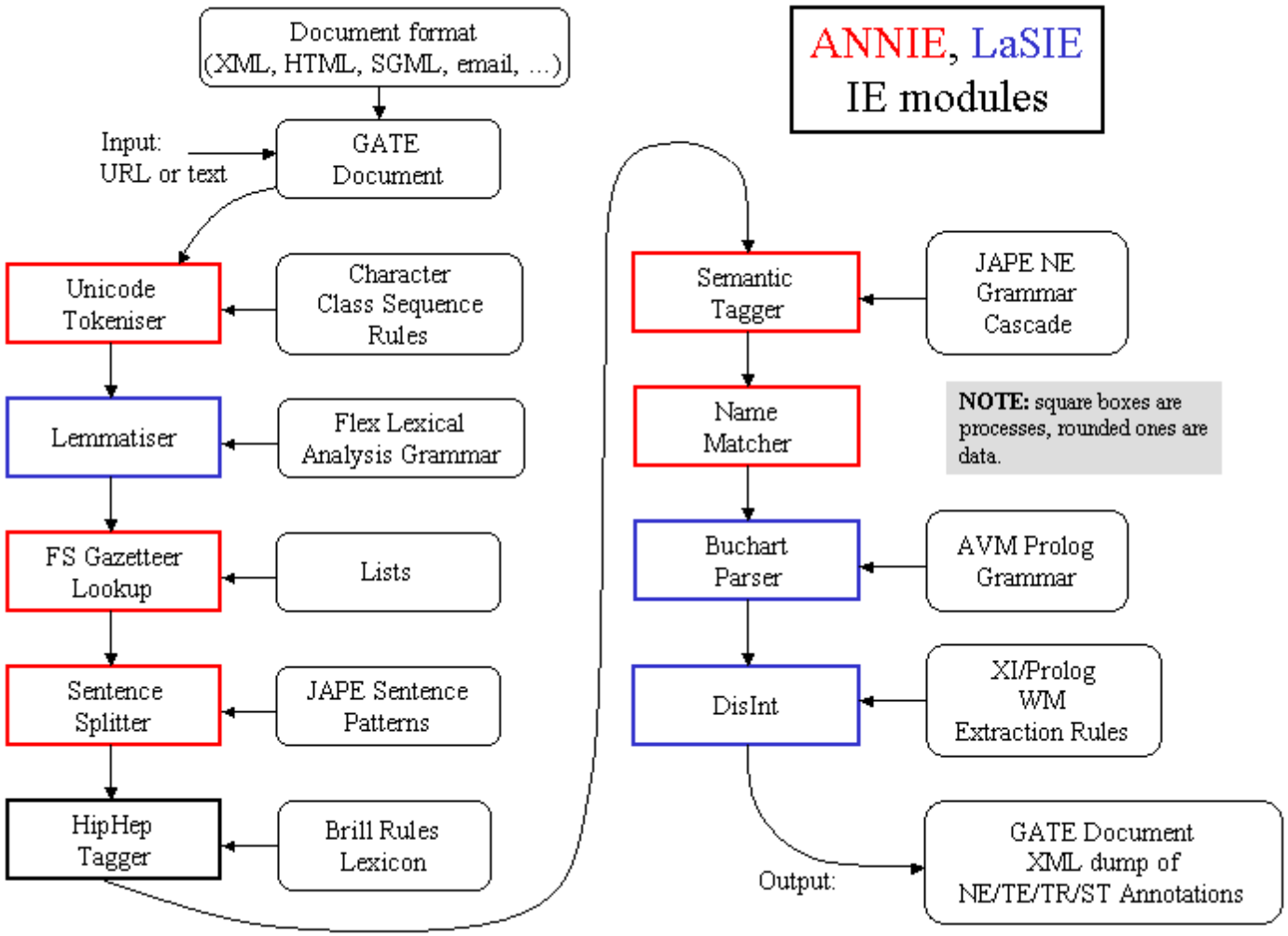
- Structural ambiguity

```
[Cable and Wireless] vs. [Microsoft] and [Dell]
[Center for Computational Linguistics] vs.
message from [City Hospital] for [John Smith].
```

IE at Sheffield

- LaSIE: a Large-Scale IE system
- VIE: a Vanilla IE system
- ANNIE: A Nearly-New IE system

ANNIE


Document format
(XML, HTML, SGML, email, ...)
Input:
URL or text
GATE
Document
Unicode
Tokeniser
Character
Class Sequence
Rules
Lemmatiser
Flex Lexical
Analysis Grammar
FS Gazetteer
Lookup
Lists
Sentence
Splitter
JAPE Sentence
Patterns
HipHep
Tagger
Brill Rules
Lexicon
ANNIE, LaSIE
IE modules
Semantic
Tagger
JAPE NE
Grammar
Cascade
Name
Matcher
NOTE: square boxes are processes, rounded ones are data.
Buchart
Parser
AVM Prolog
Grammar
DisInt
XI/Prolog
WM
Extraction Rules
Output:
GATE Document
XML dump of
NE/TE/TR/ST Annotations

Unicode Tokeniser

- Bases tokenisation on Unicode character classes
- Language-independent tokenisation
- Declarative token specification language, e.g.:

```
"UPPERCASE_LETTER" LOWERCASE_LETTER"* >

Token; orthography=upperInitial; kind=word
```

Gazetteer

- Set of lists compiled into Finite State Machines
- Each list has attributes MajorType and MinorType (and optionally, Language):

```
city.lst: location: city
currency_prefix.lst: currency_unit: pre_amount
currency_unit.lst: currency_unit: post_amount
```

- 60k entries in 80 types, inc.:

```
organization; artifact; location;
amount_unit; manufacturer;
transport_means; company_designator;
currency_unit; date;
government_designator; ...
```

The Named Entity Grammar

- Phases run sequentially and constitute a cascade of FSTs over annotations
- hand-coded rules applied to annotations to identify NEs
- annotations from format analysis, tokeniser and gazetteer modules
- use of contextual information
- rule priority based on pattern length, rule status and rule ordering
- Finds person names, locations, organisations, dates, addresses.

JAPE – a lightweight text processor

- Doug Appelt's CPSL: regular expressions over annotations
- IE is not NLU: light, regular-expression-based processing
- Cascaded finite state transduction.

JAPE Pattern Grammars

- A grammar is a set of phases, which are sets of rules
- A rule has a LHS and a RHS: pattern / action
- Pattern elements:

```
{Annotation.feature == value}
* + ? | & ( .... ):label
```

- Actions:
  create new annotations based on LHS match labels
  arbitrary Java code
- rule priority based on pattern length, rule status and rule ordering

Example of JAPE Pattern Rule

```
Rule:Company
Priority: 25
(
  ({Token.orthography == upperInitial} )+
  {Lookup.kind == companyDesignator}
):companyMatch
-->
:companyMatch.NamedEntity = { kind = "company" }
```

Coreference – the problem

- Entities referred to in many different ways
  - International Business Machines / IBM
  - General Motors Corporation / General Motors(!) / GM
  - William H Gates / Bill Gates / Mr. Gates / he

Coreference – the algorithm

1. mark each candidate (named entity/pronoun) with
   - type (location/person/etc.)
   - number (singular/plural)
   - gender
   - grammatical features (name/pronoun, definite/indefinite)
2. for each candidate find accessible antecedents
3. filter list for consistency
4. sort list using syntactic preferences

Coreference - accessibility domain

- Names - the entire preceding text.

  Match based on othographical similarities.

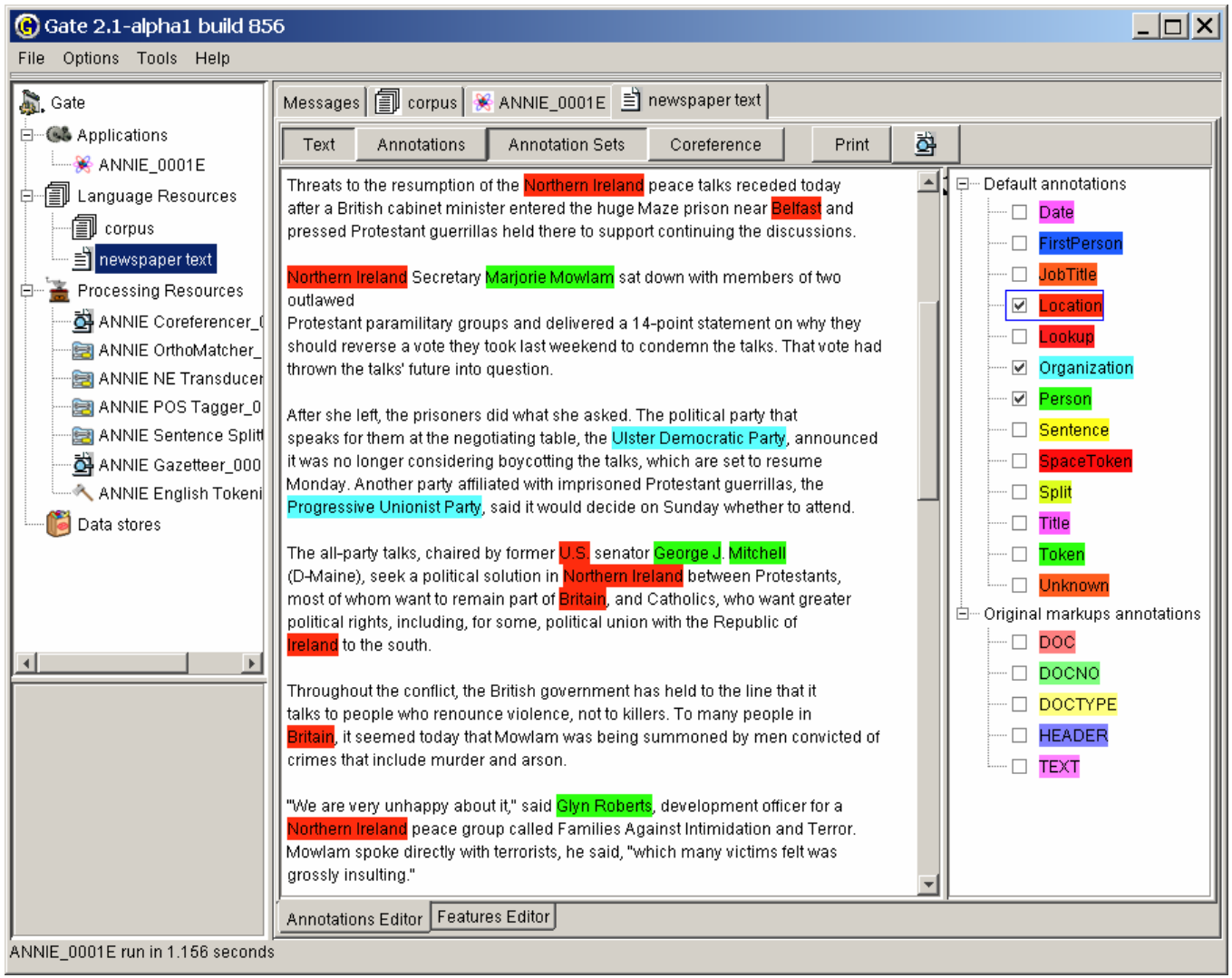
- Definite noun phrases - part of the preceding text.

  Typically determined experimentally.

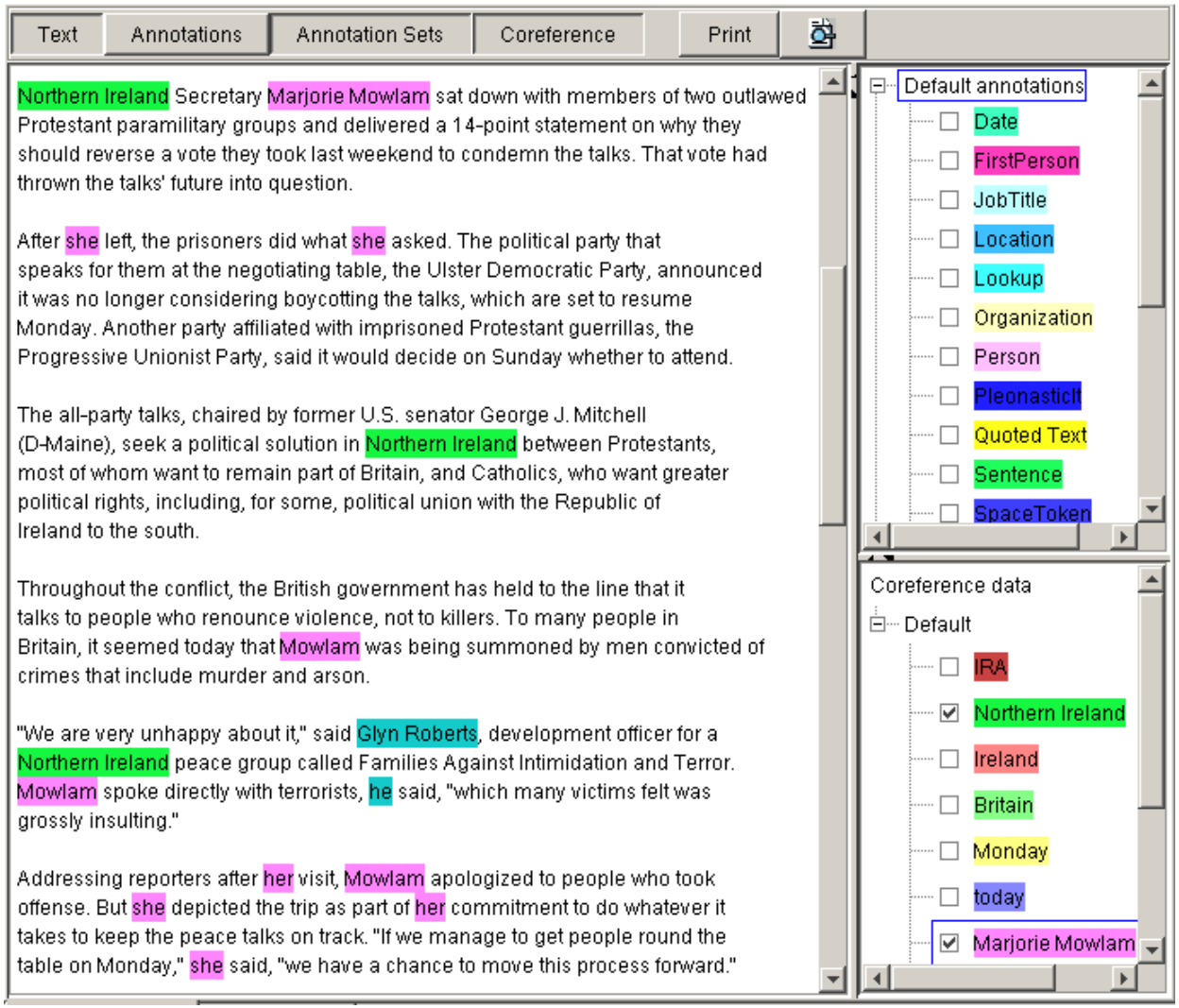
- Pronouns - a smaller part of the preceding text.

  Same paragraph perhaps.

NE Results


Gate 2.1-alpha1 build 856
File Options Tools Help
Gate
Applications
ANNIE_0001E
Language Resources
corpus
newspaper text
Processing Resources
ANNIE Coreferencer_(
ANNIE OrthoMatcher_
ANNIE NE Transducer
ANNIE POS Tagger_0
ANNIE Sentence Splitt
ANNIE Gazetteer_000
ANNIE English Tokeni
Data stores
Messages
corpus
ANNIE_0001E
newspaper text
Text
Annotations
Annotation Sets
Coreference
Print
Threats to the resumption of the Northern Ireland peace talks receded today after a British cabinet minister entered the huge Maze prison near Belfast and pressed Protestant guerrillas held there to support continuing the discussions.
Northern Ireland Secretary Marjorie Mowlam sat down with members of two outlawed
Protestant paramilitary groups and delivered a 14-point statement on why they should reverse a vote they took last weekend to condemn the talks. That vote had thrown the talks' future into question.
After she left, the prisoners did what she asked. The political party that speaks for them at the negotiating table, the Ulster Democratic Party, announced it was no longer considering boycotting the talks, which are set to resume Monday. Another party affiliated with imprisoned Protestant guerrillas, the Progressive Unionist Party, said it would decide on Sunday whether to attend.
The all-party talks, chaired by former U.S. senator George J. Mitchell (D-Maine), seek a political solution in Northern Ireland between Protestants, most of whom want to remain part of Britain, and Catholics, who want greater political rights, including, for some, political union with the Republic of Ireland to the south.
Throughout the conflict, the British government has held to the line that it talks to people who renounce violence, not to killers. To many people in Britain, it seemed today that Mowlam was being summoned by men convicted of crimes that include murder and arson.
"We are very unhappy about it," said Glyn Roberts, development officer for a Northern Ireland peace group called Families Against Intimidation and Terror. Mowlam spoke directly with terrorists, he said, "which many victims felt was grossly insulting."
Default annotations
Date
FirstPerson
JobTitle
Location
Lookup
Organization
Person
Sentence
SpaceToken
Split
Title
Token
Unknown
Original markups annotations
DOC
DOCNO
DOCTYPE
HEADER
TEXT
Annotations Editor
Features Editor
ANNIE_0001E run in 1.156 seconds

Coreference Results



More Information

http://gate.ac.uk

http://nlp.shef.ac.uk

Valentin Tablan: http://gate.ac.uk/valentin