Annotation and Evaluation

Diana Maynard, Niraj Aswani

University of Sheffield

Topics covered

- Defining annotation guidelines
- Manual annotation using the GATE GUI
- Annotation schemas and how they change the annotation editor
- Coreference annotation GUI
- Methods for ontology-based evaluation: BDM
- Using the GATE evaluation tools

The easiest way to learn…

… is to get your hands dirty!

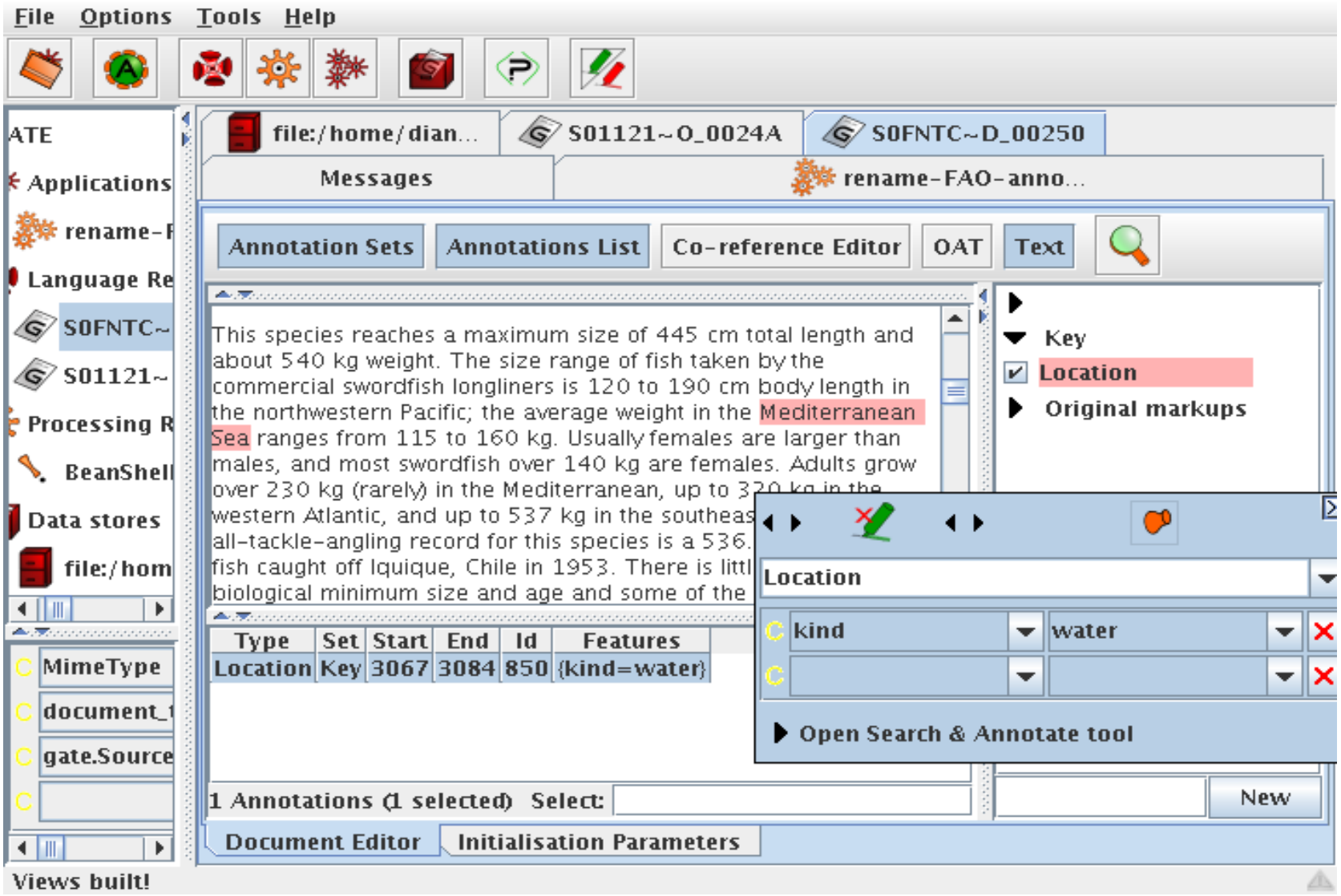
Before you start annotating...

- You need to think about annotation guidelines
- You need to consider what you want to annotate and then to define it appropriately
- With multiple annotators it's essential to have a clear set of guidelines for them to follow
- Consistency of annotation is really important for a proper evaluation

Annotation Guidelines

- People need clear definition of what to annotate in the documents, with examples
- Typically written as a guidelines document
- Piloted first with few annotators, improved, then “real” annotation starts, when all annotators are trained
- Annotation tools may require the definition of a formal DTD (e.g. XML schema)
  - What annotation types are allowed
  - What are their attributes/features and their values
  - Optional vs obligatory; default values

Manual Annotation in GATE



Annotation in GATE GUI (demo)

- Adding annotation sets
- Adding annotations
- Resizing them (changing boundaries)
- Deleting
- Changing highlighting colour
- Setting features and their values

Annotation Hands-On Exercise

- Load the Sheffield document

hands-on-resources/evaluation-materials/sheffield.xml

- Create Key annotation set
  - Type Key in the bottom of annotation set view and press the New button
- Select it in the annotation set view
- Annotate all instances of “Sheffield” with Location annotations in the Key set
- Save the resulting document as xml

Annotation Schemas

- Define types of annotations and restrict annotators to use specific feature-values
  - e.g. Person.gender = male | female
- Uses the XML Schema language supported by W3C for these definitions

```
<?xml version="1.0"?>
<schema xmlns="http://www.w3.org/2000/10/XMLSchema">
<element name="Person">
  <complexType>
    <attribute name="gender" use="optional">
     <simpleType>
       <restriction base="string">
         <enumeration value="male"/>
         <enumeration value="female"/>
       </restriction>
     </simpleType> ...
```

```
<Person gender=male/>
```

Annotation Schemas

- Just like other GATE Components
- Load them as language resources

  Language Resource → New → Annotation Schema
- Load them automatically from creole.xml

```
<resource>
  <name>Annotation schema</name>
  <class>gate.creole.AnnotationSchema</class>
  <autoinstance>
    <param name="xmlFileUrl" value="AddressSchema.xml" />
  </autoinstance>
</resource>
```

- New Schema Annotation Editor
- DEMO

Annotation Schemas Hands-on-Exercise

- Load evaluation-material/creole.xml
- Load the AddressSchema.xml schema
- Load the Schema Annotation Editor
- Load the Sheffield.xml document
- Explore the Schema Editor
- Change creole.xml to load AddressSchema.xml automatically?

Coreference annotation

- Different expressions refer to the same entity
  - e.g. UK, United Kingdom
  - e.g. Prof. Cunningham, Hamish Cunningham, H. Cunningham, Cunningham, H.
- Orthomatcher PR
  - co-reference resolution based on orthographical information of entities
  - Produces a list of annotation ids that form a co-reference chain
  - List of such lists stored as a document feature named “matches”

Coreference annotation DEMO

Messages | GATE corpus_0003D | ANNIE_0001E | example document

Text | Annotations | Annotation Sets | Coreference | Print

"It's an indication that Mike Armstrong is serious about local competition and serious about getting moving," said Anna-Marie Kovacs, an analyst for the brokerage firm Janney Montgomery Scott Inc. "AT&T and Teleport are going after the business market, which is where local companies make a lot of their money."

The deal also will give three major cable television companies, which are majority owners of Teleport, a collective 10 percent stake in AT&T.

By acquiring Teleport, AT&T can offer business customers local and long-distance
telephone service, and data and Internet access, under its own brand name. Using
Teleport's local facilities, the company also would be able to reduce the fees it pays to local phone companies for access to local telephone customers.

"It's going to permit us to be much more cost-effective as we go for that local business," Armstrong said at a news briefing. "This has competition and growth written all over it." AT&T is paying for Teleport with its stock. Teleport shareholders will receive 0.943 AT&T shares for each of their Teleport shares, putting the deal at $ 59 a share based on AT&T's closing price yesterday of $ 62.62 1/2 a share, up $ 2.62 1/2. Teleport closed down $ 3.62 1/2 at $ 54.12 1/2 a share. The companies expect the deal, which must be approved by regulators

Coreference data

- Default
  - [x] AT&T
  - [x] Teleport
  - [ ] Hughes
  - [ ] 1997
  - [ ] yesterday
  - [ ] 1996
  - [ ] WorldCom Inc
  - [x] Mike Armstron

Annotations Editor | Features Editor

Coreference annotation Hands-on-Exercise

- Load the Sheffield.xml document in a corpus and run ANNIE without Orthomatcher
- Open document and go to the Co-reference Editor
- See what chains are created?
- Highlight the chain with string “Liberal Democrats”
- Delete the members of this chain one by one from the bottom of the document to the top (note the change in the chain name)
- Recreate a chain for all the references to “Liberal Democrats”

Ontology-based Annotation

- This will be covered in the lecture on Ontologies (Wed afternoon)
- Uses a similar approach to the regular annotation GUI
- We can practise more annotation in the ad-hoc sessions for non-programmers – please ask if interested

Evaluation

“We didn’t underperform. You overexpected.”

Performance Evaluation

2 main requirements:

- **Evaluation metric**: mathematically defines how to measure the system's performance against human-annotated gold standard
- **Scoring program**: implements the metric and provides performance measures
  - For each document and over the entire corpus
  - For each type of annotation

Evaluation Metrics

- Most common are Precision and Recall
- **Precision** = correct answers/answers produced (what proportion of the answers produced are accurate?)
- **Recall** = correct answers/total possible correct answers (what proportion of all the correct answers did the system find?)
- Trade-off between Precision and Recall
- **F1** (balanced) Measure = 2PR / 2(R + P)
- Some tasks sometimes use other metrics, e.g. cost-based (good for application-specific adjustment)

AnnotationDiff

- Graphical comparison of 2 sets of annotations
- Visual diff representation, like tkdiff
- Compares one document at a time, one annotation type at a time
- Gives scores for precision, recall, F-measure etc.
- Traditionally, partial matches (mismatched spans) are given a half weight
- **Strict** considers them as incorrect; **lenient** considers them as correct

Annotations are like squirrels…

Annotation Diff helps with “spot the difference”

Annotation Diff

Annotation Diff Tool

	Document	Annotation Set				
Key:	ft-bank-of-england.xml_00016	Key	Annotation Type:	Date	F-Measure Weight	Do Diff
Response:	ft-bank-of-england.xml_00016	[Default set]	Features:	All / Some / None	1.00	

Start	End	Key	Features	=?	Start	End	Response	Features
1318	1332	second quarter	{kind=date}	-?				
1466	1474	Thursday	{}	-?				
212	222	early 1964	{kind=date}	~	218	222	1964	{kind=date, rule1=TempYear3, rule2=YearOnlyFinal}
23	31	Thursday	{kind=date, rule1=GazDate, rule2=DateOnlyFinal}	=	23	31	Thursday	{kind=date, rule1=GazDate, rule2=DateOnlyFinal}
1005	1015	last month	{kind=date}	=	1005	1015	last month	{kind=date, rule1=ModifierDate, rule2=DateOnlyFinal}
1582	1591	next week	{kind=date}	=	1582	1591	next week	{kind=date, rule1=ModifierDate, rule2=DateOnlyFinal}

			Recall	Precision	F-Measure
Correct:	3				
Partially Correct:	1	Strict:	0.50	0.75	0.60
Missing:	2	Lenient:	0.6667	1.00	0.80
False Positives:	0	Average:	0.5833	0.875	0.70

Export to HTML

AnnotationDiff Exercise

- Load the **Sheffield** document that you annotated and saved earlier.
- Load ANNIE and select Document Reset PR.
- Add “Key” to the parameter “setsToKeep” (this ensures Key set is not deleted)
- Run ANNIE on the Sheffield document.
- Open the **Annotation Diff** (Tools menu)
- Select **Sheffield** document
- Key contains your manual annotations. (select as **Key** annotation set)
- Default contains annotations from ANNIE (select as **Response** annotation set)
- Select the **Location** annotation
- Check precision and response
- See the errors

Corpus Benchmark Tool

- Compares annotations at the corpus level
- Compares all annotation types at the same time, i.e. gives an overall score, as well as a score for each annotation type
- Enables regression testing, i.e. comparison of 2 different versions against gold standard
- Visual display, can be exported to HTML
- Granularity of results: user can decide how much information to display
- Results in terms of Precision, Recall, F-measure

Corpus structure

- Corpus benchmark tool requires a particular directory structure
- Each corpus must have a **clean** and **marked** sub-directory
- **Clean** holds the unannotated version, while **marked** holds the marked (gold standard) ones
- There may also be a **processed** subdirectory – this is a datastore (unlike the other two)
- Corresponding files in each subdirectory must have the same name

How it works

- Clean, marked, and processed
- Corpus_tool.properties – must be in the directory where build.xml is
- Specifies configuration information about
  - What annotation types are to be evaluated
  - Threshold below which to print out debug info
  - Input set name and key set name
- Modes
  - Storing results for later use
  - Human marked against already stored, processed
  - Human marked against current processing results
  - Regression testing – default mode

Corpus Benchmark Tool

View Attachment: temp.html

ABC19980430.1830.0858.sgm

Annotation Type	Precision	Recall	Annotations
Annotation type: Organization	1.0 Precision increase on human-marked from 0.75 to 1.0	0.75 Recall increase on human-marked from 0.375 to 0.75	MISSING ANNOTATIONS in the automatic texts: **ABC**: *[2849,2852]* SPURIOUS ANNOTATIONS in the automatic texts: PARTIALLY CORRECT ANNOTATIONS in the automatic texts:
Annotation type: Person	0.9444444444444444 Precision increase on human-marked from 0.8947368421052632 to 0.9444444444444444	0.9444444444444444	
Annotation type: GPE	1.0	1.0 Recall increase on human-marked from 0.8571428571428571 to 1.0	

Corpus benchmark tool demo

- Setting the properties file
- Running the tool
- Visualising and saving the results

Ontology-based evaluation: BDM

- Traditional methods for IE (Precision and Recall) are not sufficient for ontology-based IE
- The distinction between right and wrong is less obvious
- Recognising a Person as a Location is clearly wrong, but recognising a Research Assistant as a Lecturer is not so wrong
- Integration of similarity metrics enable closely related items some credit

Which things are most similar?

Balanced Distance Metric

- Considers the relative specificity of the taxonomic positions of the key and response
- Unlike some algorithms, does not distinguish between the directionality of this relative specificity,
- Distances are normalised wrt average length of chain
- Makes the penalty in terms of node traversal relative to the semantic density of the concepts in question
- More information in the LREC 2008 paper “Evaluating Evaluation Metrics for Ontology-Based Applications” available from the GATE website

Examples of misclassification

Entity	Response	Key	BDM
Sochi	Location	City	0.724
FBI	Org	GovOrg	0.959
Al-Jazeera	Org	TVCompany	0.783
Islamic Jihad	Company	ReligiousOrg	0.816
Brazil	Object	Country	0.587
Senate	Company	Political Entity	0.826

BDM Plugin

- Load the **BDMComputation** Plugin, load a **BDMComputation** PR and add it to a corpus pipeline
- Set the parameters
  - **ontologyURL** (location of the ontology)
  - **outputBDMFile** (plain text file to store the BDM values)
- Result will be written to this file with BDM scores for each match
- This file can be used as input for the IAA plugin

IAA Plugin

- This computes inter-annotator agreement.
- Uses the same computation as the corpus benchmarking tool but can compare more than 2 sets simultaneously
- Also enables calculation using BDM
- Can be used for classification tasks also to compute Kappa and other measures
- Load the IAA Plugin and then an IAA Computation PR, and add it to a pipeline.
- If using the BDM, select the BDM results file

More on using the evaluation plugins

- More detail and hands-on practice with the evaluation plugins during the ad-hoc sessions for non-programmers
- Please ask if interested