

Machine Learning

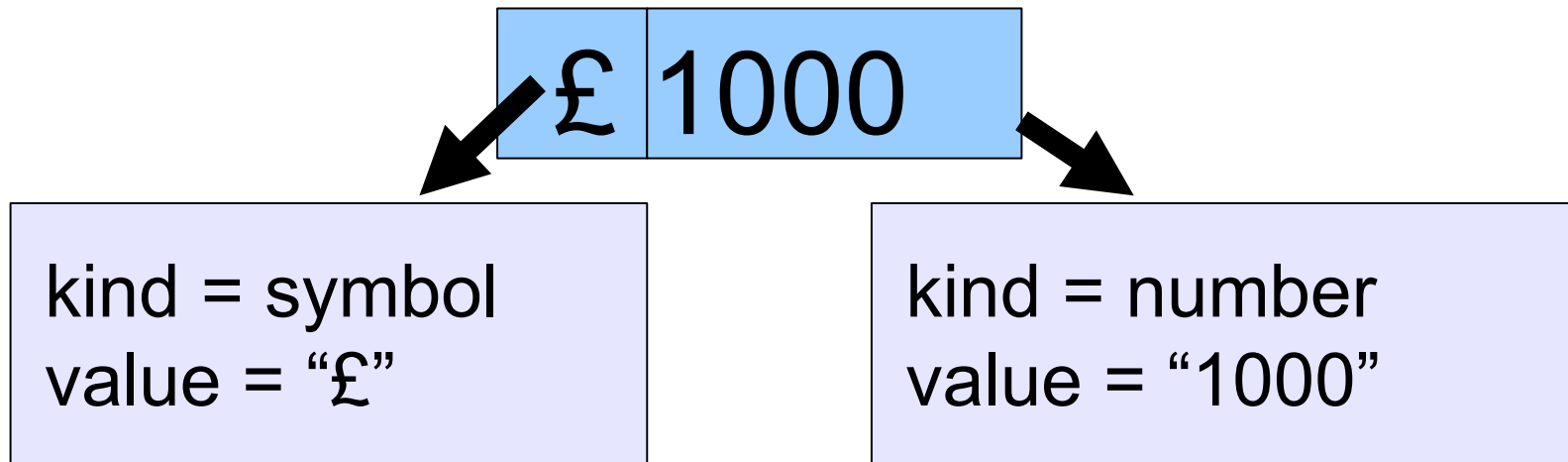
What is Machine Learning and why do we want to do it?

What is ML?

- Automating the process of inferring new data from existing data
- We will introduce ML by providing an overview of terminology only
- We cannot provide a tutorial on ML. Try:
 - Playing with Weka and reading the Weka book
<http://www.cs.waikato.ac.nz/ml/weka/index.html>
 - Andrew Ng's course:
<https://www.coursera.org/course/ml>

Learning a pattern

- In GATE, that means creating annotations by learning how they relate to other annotations
- For example, we have “Token” annotations with “kind” and “value” features



- ML could learn that a “£” followed by a number is an amount of currency

How is that better than making rules?

- It is different to the rule-based approach
- Humans are better at writing rules for some things, and ML algorithms are better at finding some things
- With ML you don't have to create all the rules
- However, you have to manually annotate a training corpus (or get someone else to do it!)
- Rule-based approaches (e.g. JAPE) and ML work well together; JAPE is often used extensively to prepare data for ML



Terminology: Instances, attributes, classes

California Governor Arnold Schwarzenegger proposes deep cuts.

Instances:

Any annotation
Tokens are often convenient



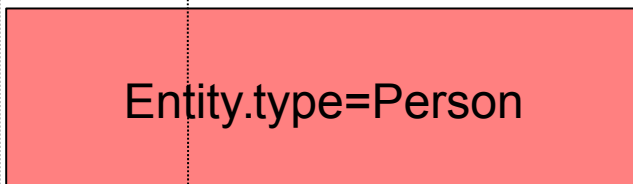
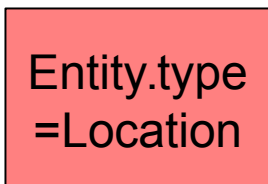
Attributes:

Any annotation feature relative to instances
Token.String
Token.category (POS)
Sentence.length



Class:

The thing we want to learn
A feature on an annotation



Instances

- Instances are cases that may be learned
- Every instance is a decision for the ML algorithm to make
- To which class does this instance belong?
 - “California” → Location



Terminology: Instances, attributes, classes

California Governor Arnold Schwarzenegger proposes deep cuts.

Instances:

Any annotation
Tokens are often convenient



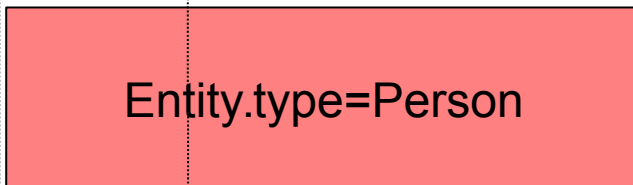
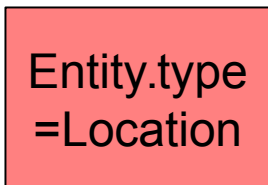
Attributes:

Any annotation feature relative to instances
Token.String
Token.category (POS)
Sentence.length



Class:

The thing we want to learn
A feature on an annotation



Attributes

- Attributes are pieces of information about instances
- They are sometimes called “features” in machine learning literature
- Examples
 - `Token.string == “Arnold”`
 - `Token.orth == upperInitial`
 - `Token(-1).string == “Governor”`



Terminology: Instances, attributes, classes

California Governor Arnold Schwarzenegger proposes deep cuts.

Instances:

Any annotation
Tokens are often convenient



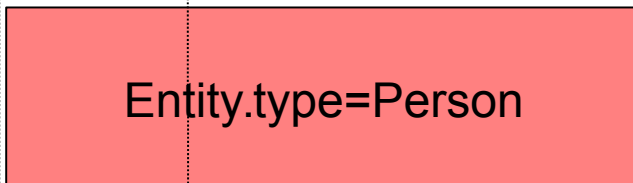
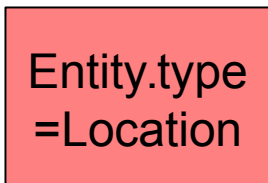
Attributes:

Any annotation feature relative to instances
Token.String
Token.category (POS)
Sentence.length



Class:

The thing we want to learn
A feature on an annotation



Classes

- The class is what we want to learn
- Suppose we want to find persons' names: for every instance, the question is “is this a person name?” and the classes are “yes” and “no”
- Sometimes there are many classes, for example we may want to learn entity types
 - For every instance, the question is “which type from the list does this instance belong to?”
 - One answer is “none of them”



Terminology: Instances, attributes, classes

California Governor Arnold Schwarzenegger proposes deep cuts.

Instances:

Any annotation
Tokens are often convenient



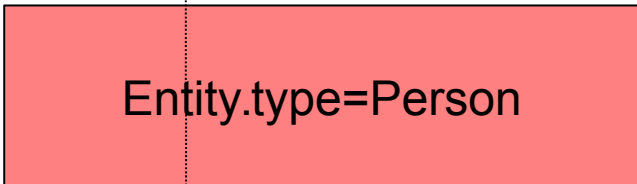
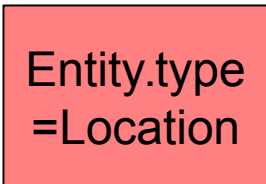
Attributes:

Any annotation feature relative to instances
Token.String
Token.category (POS)
Sentence.length



Class:

The thing we want to learn
A feature on an annotation



Machine Learning in GATE

- GATE supports machine learning in several ways
- Some of the **standard PRs** are ML-based e.g.
 - ANNIE POS tagger
 - Stanford parser
- **Machine Learning PR**
 - Provides Weka integration, but is a little out of date and only supports token-based attributes – though you can get around this using JAPE
- **Third-party NLP components**
 - e.g. the OpenNLP PR can be used with any models, trained externally to GATE
- **Roll-your-own**
 - It is relatively straightforward to write a PR that will create learning instances from your text, and export them in a format suitable for your favourite ML toolkit. Or even to integrate more fully.
- **Forthcoming**: a new GATE machine learning PR called the **Learning Framework**
 - Integrates more libraries, including Mallet's CRF
 - Export to ARFF and compatible algorithm availability allows feature selection and parameter tuning in Weka



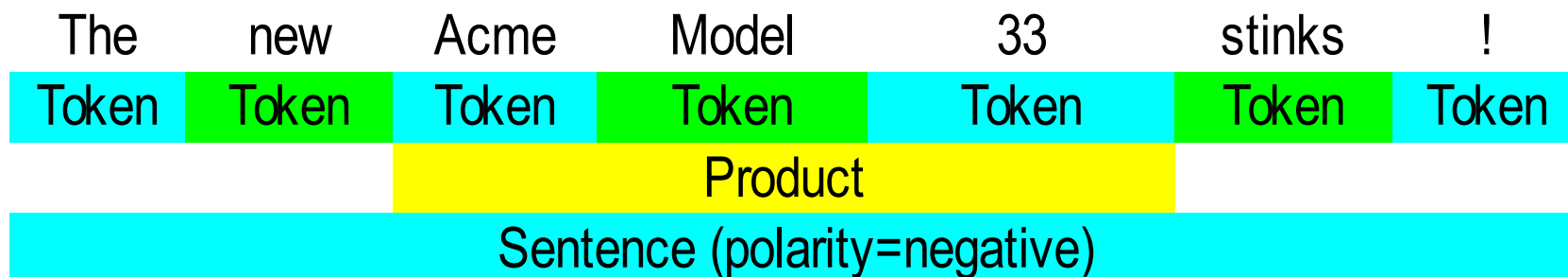
Learning PR

- We will not look at any of the above in this tutorial
- We will be looking at the **Learning PR**
- Advantages
 - Simple to use and configure
 - Good annotation to attribute mapping
 - Hides the mapping of NLP tasks to multiple binary classification problems
 - Provides its own evaluation framework
 - Support for common ML algorithms (SVM, Perceptron, some Weka algorithms)
- Disadvantages
 - Heavily biased to SVM
 - Inflexible: hard to dig in to the code
 - Hard to extend to more recent algorithms

ML Tasks in the Learning PR

- The Learning PR supports 3 types of ML tasks:
- chunk recognition (named entity recognition, NP chunking) as in the previous example
- text classification (sentiment classification, POS tagging) as in the following example
- relation annotation (this requires special techniques that are not covered in this module, although materials are available)

Example: text classification



- instance: Sentence annotation
- attributes: Token and Product annotations and their features (suppose that the Product annotations have been created earlier with gazetteers and rules)
- class: polarity= “negative”
- ML could learn that a Product close to the Token “stinks” expresses a negative sentiment, then add a polarity=“negative” feature to the Sentence.

Training

- Training involves presenting data to the ML algorithm from which it creates a model
- The training data (instances) have been annotated with class annotations as well as attributes
- Models are representations of decision-making processes that allow the machine learner to decide what class the instance has based on the attributes of the instance



Application

- When the machine learner is applied, it creates new class annotations on data using the model
- The corpus it is applied to must contain the required attribute annotations
- The machine learner will work best if the application data is similar to the training data

Evaluation

- We want to know how good our machine learner is before we use it for a real task
- Therefore we apply it to some data for which we already have class annotations
 - The “right answers”, sometimes called “gold standard”
- If the machine learner creates the same annotations as the gold standard, then we know it is performing well
- The test corpus must not be the same corpus as you trained on
 - This would give the machine learner an advantage, and would give a false idea of how good it is
- GATE's ML PR has a built-in evaluation mode that splits the corpus into training and test sets and cross-validates them

Setting up a Corpus

Load the corpus

- Create a corpus (any name is fine; you can even leave it blank)
- Populate it from `ner/corpus/*.xml` in the hands-on materials
 - Set the encoding to UTF-8
- You should get 93 documents (numbered 0 to 92 in the corpus)
- Open a document and examine its annotations



Examining the corpus

- The corpus contains an annotation set called “Key”, which has been manually prepared
- Within this annotation set are annotations of types “Date”, “Location”, “Money”, “Organization” and so forth
- There are also some annotations in the “Original markups” set (these represent HTML tags)

What are we going to do with this corpus?

- We are going to train a machine learner to annotate corpora with these entity types
- We need a training corpus and a test corpus
- The training corpus will be used by the machine learner to deduce relationships between attributes and entity types (classes)
- The test corpus will be used to find out how well it is working, by comparing annotations created by the learner with the correct annotations that are already there
- In *Evaluation* mode, which we will try first, the ML PR automatically splits one corpus up into training and test sets

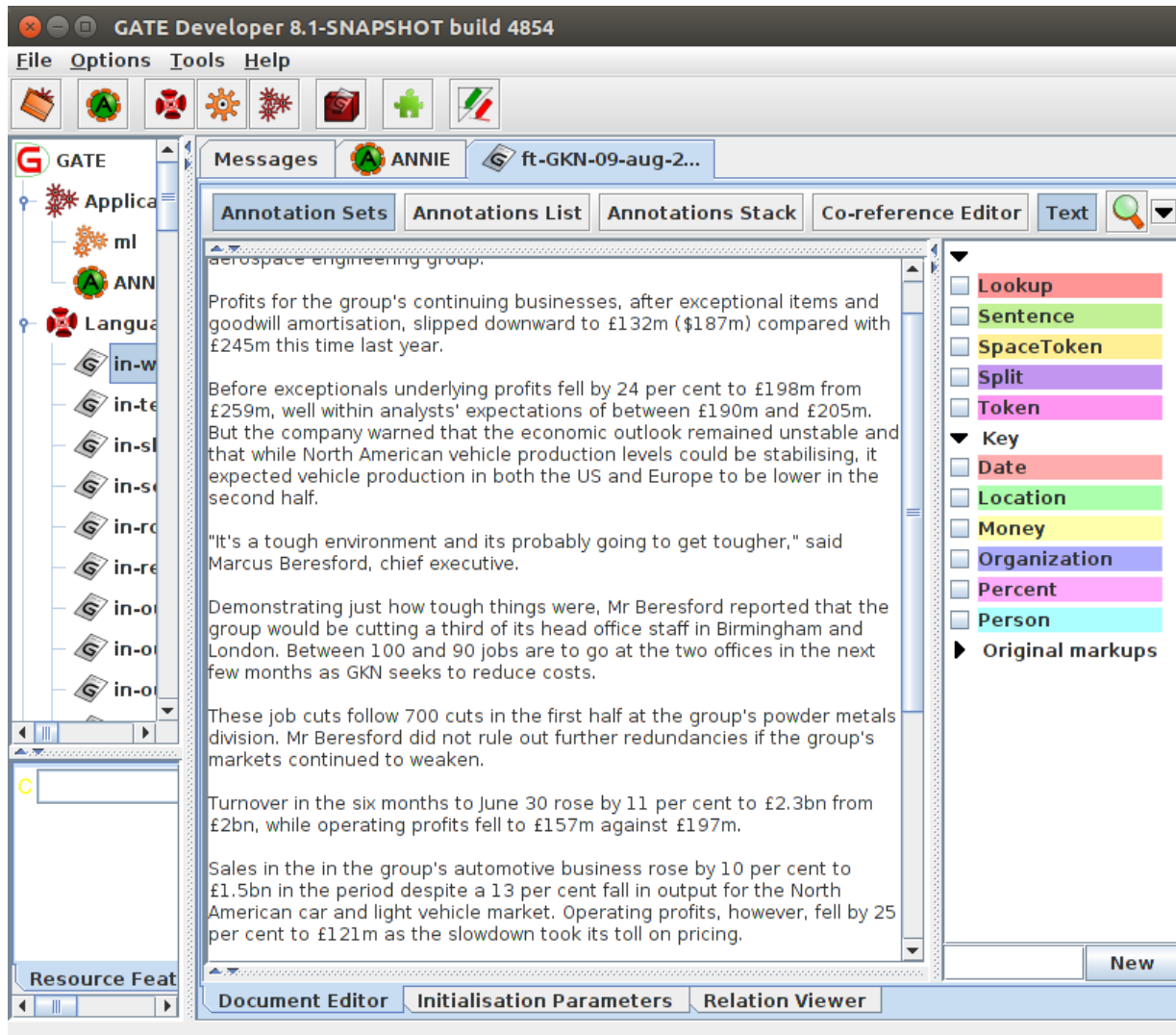
Instances and Attributes

- This corpus so far contains only the class annotations
- There is not much in this corpus to learn from
- What would our instances be?
- What would our attributes be?
- If we run parts of ANNIE over the corpus, then we can use:
 - Token annotations for instances
 - Token features for attributes
 - Gazetteer Lookups for attributes

Instances and Attributes

- **Load ANNIE**
- **We only want**
 - **Tokens and some basic features**
 - **Gazetteer Lookups**
- **So remove the last two Prs from the pipeline**
 - **ANNIE NE Transducer**
 - **ANNE Orthomatcher**
- **Check that the document reset PR's setsToKeep parameter includes “Key”!**
- **Run this cut-down ANNIE over your corpus**

Running ANNIE on the corpus



The screenshot shows the GATE Developer 8.1-SNAPSHOT build 4854 interface. The main window displays a document editor with the text of a news article. The ANNIE tool is running on the document, and the Annotations List pane on the right shows the following annotations:

- Lookup
- Sentence
- SpaceToken
- Split
- Token
- ▼ Key
 - Date
 - Location
 - Money
 - Organization
 - Percent
 - Person
- ▶ Original markups

The document text is as follows:

aerospace engineering group.

Profits for the group's continuing businesses, after exceptional items and goodwill amortisation, slipped downward to £132m (\$187m) compared with £245m this time last year.

Before exceptionals underlying profits fell by 24 per cent to £198m from £259m, well within analysts' expectations of between £190m and £205m. But the company warned that the economic outlook remained unstable and that while North American vehicle production levels could be stabilising, it expected vehicle production in both the US and Europe to be lower in the second half.

"It's a tough environment and its probably going to get tougher," said Marcus Beresford, chief executive.

Demonstrating just how tough things were, Mr Beresford reported that the group would be cutting a third of its head office staff in Birmingham and London. Between 100 and 90 jobs are to go at the two offices in the next few months as GKN seeks to reduce costs.

These job cuts follow 700 cuts in the first half at the group's powder metals division. Mr Beresford did not rule out further redundancies if the group's markets continued to weaken.

Turnover in the six months to June 30 rose by 11 per cent to £2.3bn from £2bn, while operating profits fell to £157m against £197m.

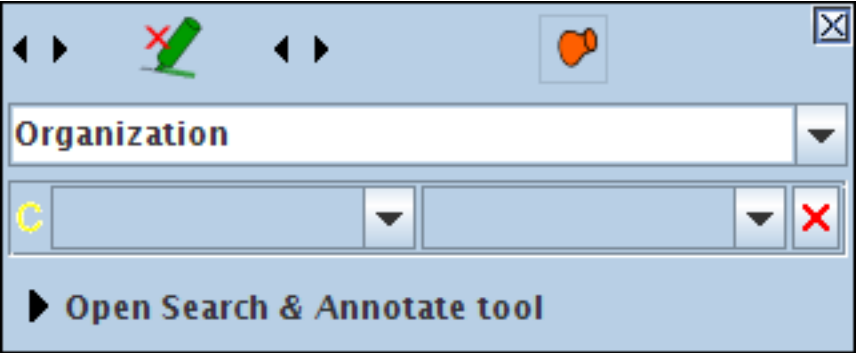
Sales in the in the group's automotive business rose by 10 per cent to £1.5bn in the period despite a 13 per cent fall in output for the North American car and light vehicle market. Operating profits, however, fell by 25 per cent to £121m as the slowdown took its toll on pricing.

- Having run ANNIE on the corpus, we have more annotations to work with

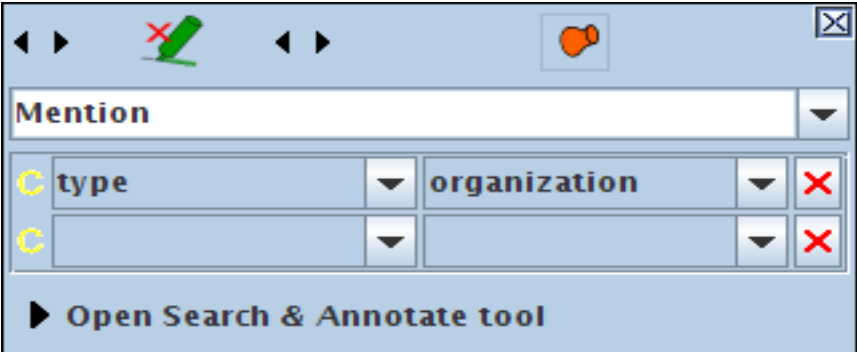


Preparing the corpus: Classes

- What we have:



- What we need:



Preparing the corpus: Classes

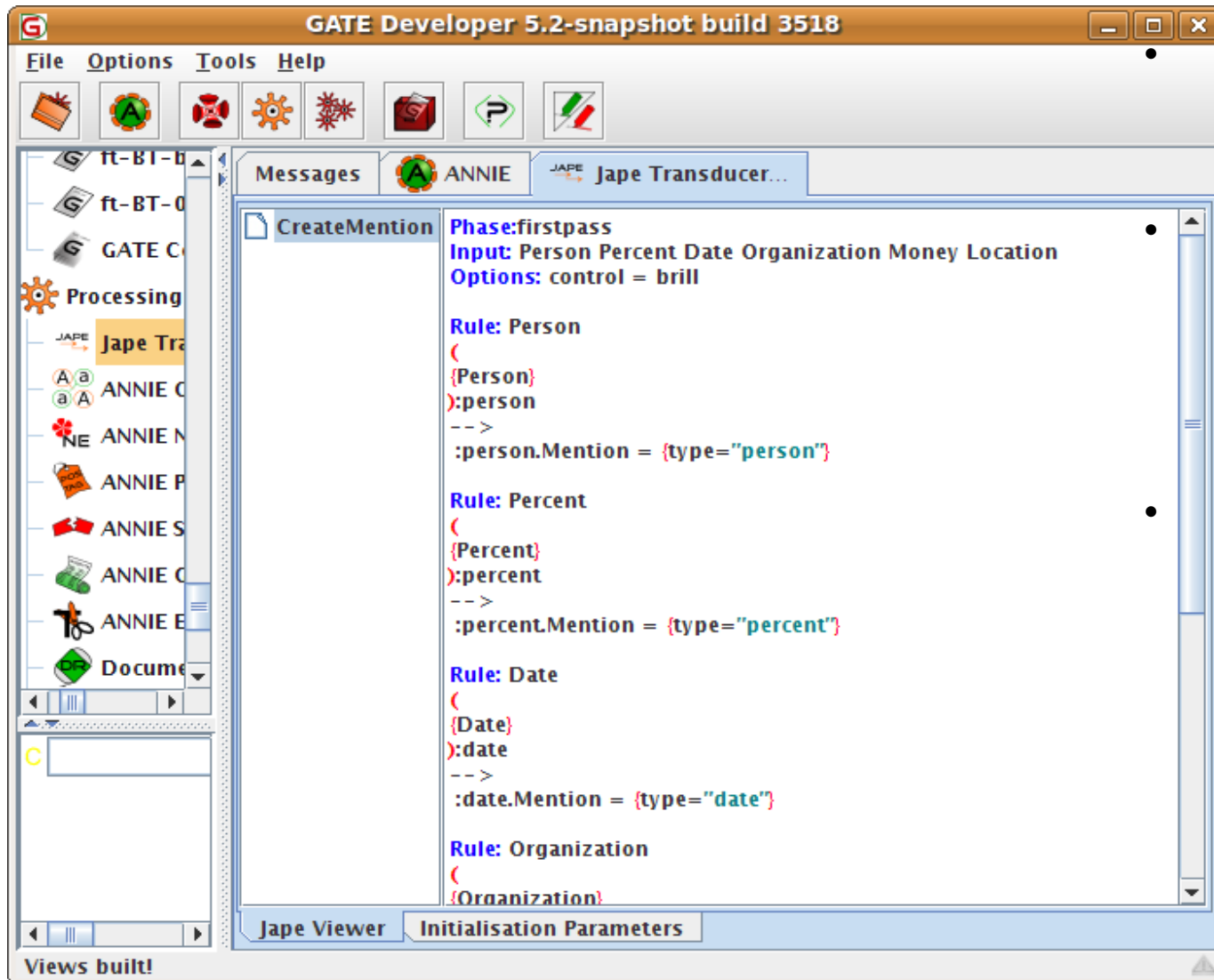
- Currently each class has its own annotation type (Date, Person, Percent etc.)
- But the ML PR expects the class (ML term) to be a feature value, not an annotation type
- So we need to make a new annotation type for the ML to learn from: “Mention” (it doesn't matter what it's called as long as we're consistent and configure the PR to match)



Making class annotations

- **Load a JAPE transducer from the ner/CreateMention.jape grammar**
- **Look at the grammar in GATE**

The CreateMention.jape grammar



The screenshot shows the GATE Developer 5.2 interface. The main window displays the 'CreateMention' grammar in the Jape Viewer. The grammar is defined as follows:

```

Phase: firstpass
Input: Person Percent Date Organization Money Location
Options: control = brill

Rule: Person
(
  {Person}
):person
-->
:person.Mention = {type="person"}

Rule: Percent
(
  {Percent}
):percent
-->
:percent.Mention = {type="percent"}

Rule: Date
(
  {Date}
):date
-->
:date.Mention = {type="date"}

Rule: Organization
(
  {Organization}

```

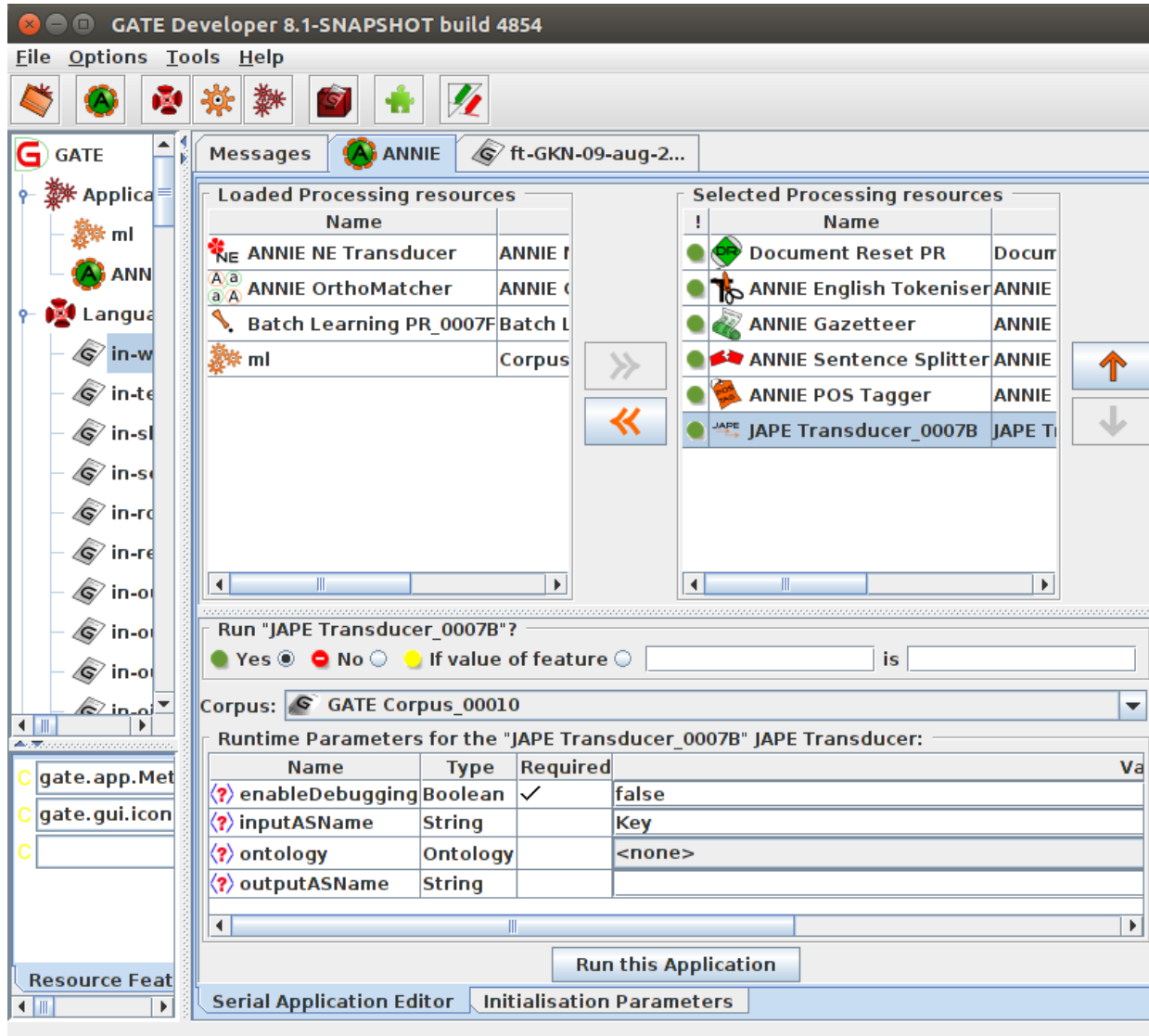
The interface also shows a sidebar with various components like 'Processing', 'Jape Transducer...', and 'ANNIE'. The status bar at the bottom indicates 'Views built!'.

This grammar makes a new annotation type called “Mention”

It makes the previous annotation type into a feature of the “Mention” annotation

Feature name is “type” because “class” is reserved for ontology use

Applying the grammar to the corpus



The screenshot shows the GATE Developer interface with the following components:

- Messages:** ANNIE, ft-GKN-09-aug-2...
- Loaded Processing resources:**

Name	
NE ANNIE NE Transducer	ANNIE I
ANNIE OrthoMatcher	ANNIE C
Batch Learning PR_0007F	Batch L
ml	Corpus
- Selected Processing resources:**

Name	
Document Reset PR	Docurr
ANNIE English Tokeniser	ANNIE
ANNIE Gazetteer	ANNIE
ANNIE Sentence Splitter	ANNIE
ANNIE POS Tagger	ANNIE
JAPE Transducer_0007B	JAPE T
- Run "JAPE Transducer_0007B"?:**

Yes No If value of feature is
- Corpus:** GATE Corpus_00010
- Runtime Parameters for the "JAPE Transducer_0007B" JAPE Transducer:**

Name	Type	Required	Value
enabledDebugging	Boolean	✓	false
inputASName	String		Key
ontology	Ontology		<none>
outputASName	String		
- Buttons:** Run this Application
- Serial Application Editor:** Initialisation Parameters

Add the JAPE transducer at the end of your application

Set the inputASName to "Key"

Leave the outputASName blank (default)



Check the “Mention” annotations

- **Check that you have some “Mention” annotations**
- **Check that they have a feature “type” and that the values look right**
- **These Mention annotations are derived from the named entities in the “Key” AS, and will form our instances**

The Configuration File

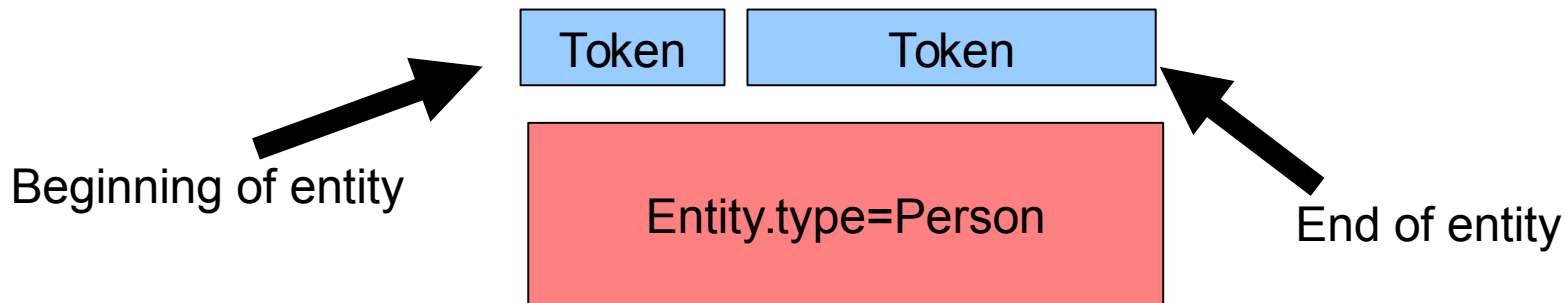


Looking at the configuration file

- In the configuration file, we tell the machine learning PR what we want it to do
- You will find a configuration file in your hands-on materials, called ner/ner-config-file.xml
- **Open it using a text editor**

<SURROUND value="true"/>

California Governor Arnold Schwarzenegger proposes deep cuts.



- The class to be learned covers more than one instance (chunking problem)
- The PR has to learn the boundaries (chunking problem)
- So we tell the PR to use *surround mode*

Confidence Thresholds

```
<PARAMETER name="thresholdProbabilityEntity" value="0.2"/>  
<PARAMETER name="thresholdProbabilityBoundary" value="0.4"/>
```

- Classifiers provide confidence ratings—how likely a result is to be correct
- We must determine how certain is good enough
- Depending on the application we might prefer to include or exclude annotations for which the learner is not too sure
- `thresholdProbabilityBoundary` is a threshold for the beginning and end instances
- `thresholdProbabilityEntity` is a threshold for beginning and end instances combined



California Governor Arnold Schwarzenegger proposes deep cuts.

Entity.type
=Location

Entity.type=Person

- Many algorithms are binary classifiers (e.g. yes/no)
- We have several classes (Person, Location, Organization etc.)
- Therefore the problem must be converted to a set of binary problems, so we can use binary algorithms
- **one-vs-others**
 - LOC vs PERS+ORG / PERS vs LOC+ORG / ORG vs LOC+PERS
- **one-vs-another**
 - LOC vs PERS / LOC vs ORG / PERS vs ORG

<multiClassification2Binary method="one-vs-others"/>

- With more than 3 classes, **one-vs-another** becomes very computationally expensive!
- **one-vs-others**: N classes => N classifiers
 - A vs B+C+D, B vs A+C+D, C vs A+B+D, D vs A+B+C
- **one-vs-another**: N classes => $N \times (N-1) / 2$ classifiers
 - A vs B, A vs C, A vs D, B vs C, B vs D, C vs D

```
<EVALUATION method="holdout"  
ratio="0.66"/>
```



- We are going to evaluate our application in two ways today
 - The ML PR can automatically evaluate for us
 - We will also run our own evaluation
- This parameter dictates how the ML PR will work in evaluation mode
- The PR ignores this part of the config file in training and application modes

Evaluation

`<EVALUATION method="kfold" runs="4"/>`

OR

`<EVALUATION method="holdout" ratio="0.66"/>`

- Holdout randomly picks *ratio* documents for training and uses the rest for testing; this is faster than k-fold because it only runs once
- k-fold cross-validation will give you more reliable results and lets you “stretch” your corpus

K-Fold Cross-Validation

- In k-fold cross-validation, the corpus is split into k equal parts, and the learner is trained k times on k-1 parts and evaluated on 1; the results are averaged
- For example, if k=4, the documents are split into groups A, B, C, & D, then:
 - train on A+B+C, test on D;
 - train on A+B+D, test on C;
 - train on A+C+D, test on B;
 - train on B+C+D, test on A;
 - average these 4 results
- This maximises the use of the training data without losing testing accuracy, but takes 4 times as long



<ENGINE nickname="SVM" ..

- Next we specify what machine learning algorithm we wish to use
- Today we are using the SVM (“SVM”)
- We will use the following options: options="-t 0 -m 100 -tau 0.4"
 - Challenge: find out what these options do! (Hint: user guide §19.2)

<INSTANCE-TYPE>...

- Next, we tell the ML PR what our instance annotation is
- The goal of the ML PR is to try to learn how the attributes of every instance relate to its class, so the instance is an important choice
- We have decided that the “Token” is our instance annotation type
 - We have run the tokenizer to ensure we have Token annotations in our corpus
 - The POS tagger adds category features to the Token annotations

Specifying Attributes

```
<ATTRIBUTELIST>  
  <NAME>POS</NAME>  
  <SEMTYPE>NOMINAL</SEMTYPE>  
  <TYPE>Token</TYPE>  
  <FEATURE>category</FEATURE>  
  <RANGE from="-2" to="2"/>  
</ATTRIBUTELIST>
```

- For every attribute, we create a specification like the one above
- This is the information from which the PR will learn, so it is important to give it some good data
- You can see in the configuration file that there are several attributes (including Lookup.majorType), providing a good range of information
- However, if you have too many attributes it can take a very long time to learn!

Breaking down the attribute specification

- `<NAME>POS</NAME>`
 - This is the name that we choose for this attribute. It can be anything we want, but it will help us later if we make it something sensible!
- `<SEMTYPE>NOMINAL</SEMTYPE>`
 - Is the value of this attribute a number or a name?

Breaking down the attribute specification

- `<TYPE>Token</TYPE>`
 - The value of the attribute will be taken from the “Token” annotation
- `<FEATURE>category</FEATURE>`
 - The value of the attribute will be taken from the “category” feature

Breaking down the attribute specification

```
<ATTRIBUTELIST>  
:  
  <RANGE from="-2" to="2"/>  
</ATTRIBUTELIST>
```

- Because this is an “ATTRIBUTELIST” specification, we can specify a “RANGE”
- In this case, we will gather attributes from the current instance and also the preceding and following two; i.e., a window of 5 Token annotations centred on the one in question

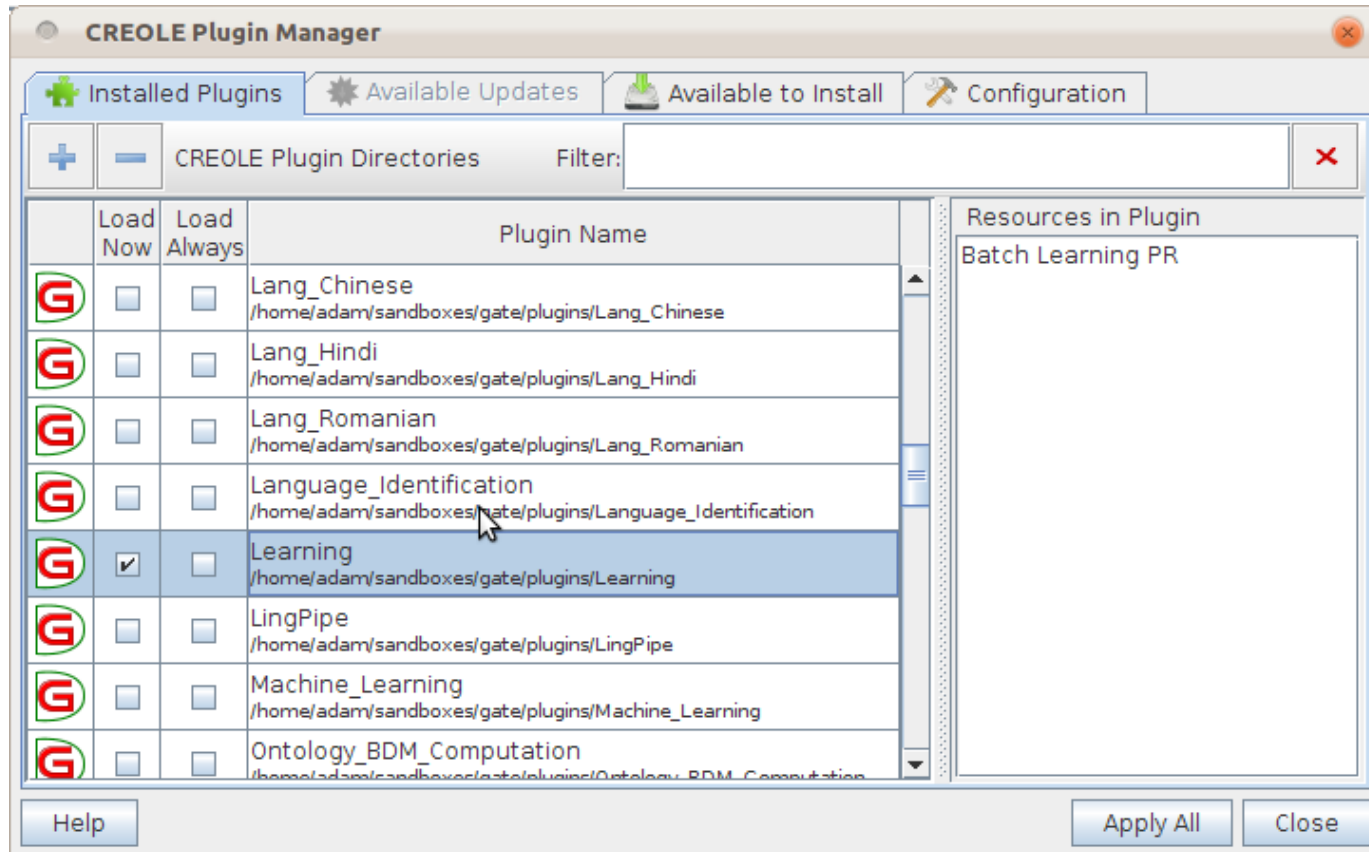
Specifying the Class Attribute

```
<ATTRIBUTE>
  <NAME>Class</NAME>
  <SEMTYPE>NOMINAL</SEMTYPE>
  <TYPE>Mention</TYPE>
  <FEATURE>type</FEATURE>
  <POSITION>0</POSITION>
  <CLASS/>
</ATTRIBUTE>
```

- You can call the class attribute whatever you want, but “Class” is a sensible choice
- Remember that our class attribute is the “type” feature of the “Mention” annotation
- This is an ATTRIBUTE, not an ATTRIBUTELIST, so we have “position”, not “range”
- The <CLASS/> element tells the Batch Learning PR that this is the class attribute to learn.

Running the ML PR in evaluation mode

Loading the Learning plugin



- **Load the “Learning” plugin**
- (We are **not** going to use the “Machine Learning” plugin, as explained earlier)

Creating a learning application

- **Create a “Batch Learning PR” with ner/ner-config.xml as the the configFileURL parameter**
- **Make a new corpus pipeline and put this PR (only!) in it**

Running the application in evaluation mode



GATE Developer 7.2-SNAPSHOT build 4620

File Options Tools Help

GATE

- Applications
 - Corpus Pipeline_00...
 - ANNIE
- Language Resources
- Processing Resources
 - Batch Learning PR...
 - JAPE createMention
 - ANNIE OrthoMatch...
 - ANNIE NE Transdu...
 - ANNIE POS Tagger
 - ANNIE Sentence Sp...
 - ANNIE Gazetteer
 - ANNIE English Tok...

Messages ANNIE in-outlook-ba-0... Corpus Pipeline...

Loaded Processing resources

Name	
ANNIE English Tokeniser	ANNIE English T...
ANNIE Gazetteer	ANNIE Gazetteer
ANNIE NE Transducer	ANNIE NE Trans...
ANNIE OrthoMatcher	ANNIE OrthoMa...
ANNIE POS Tagger	ANNIE POS Tagg...
ANNIE Sentence Splitter	ANNIE Sentence...
Document Reset PR	Document Rese...
JAPE createMention	JAPE Transduce...

Selected Processing resources

Name
Batch Learning PR_000...

Corpus: GATE Corpus_00011

Runtime Parameters for the "Batch Learning PR_0008D" Batch Learning PR:

Name	Type	Required	
inputASName	String		
learningMode	RunMode	✓	EVALUATION
outputASName	String		
runProtocolDir	URL		

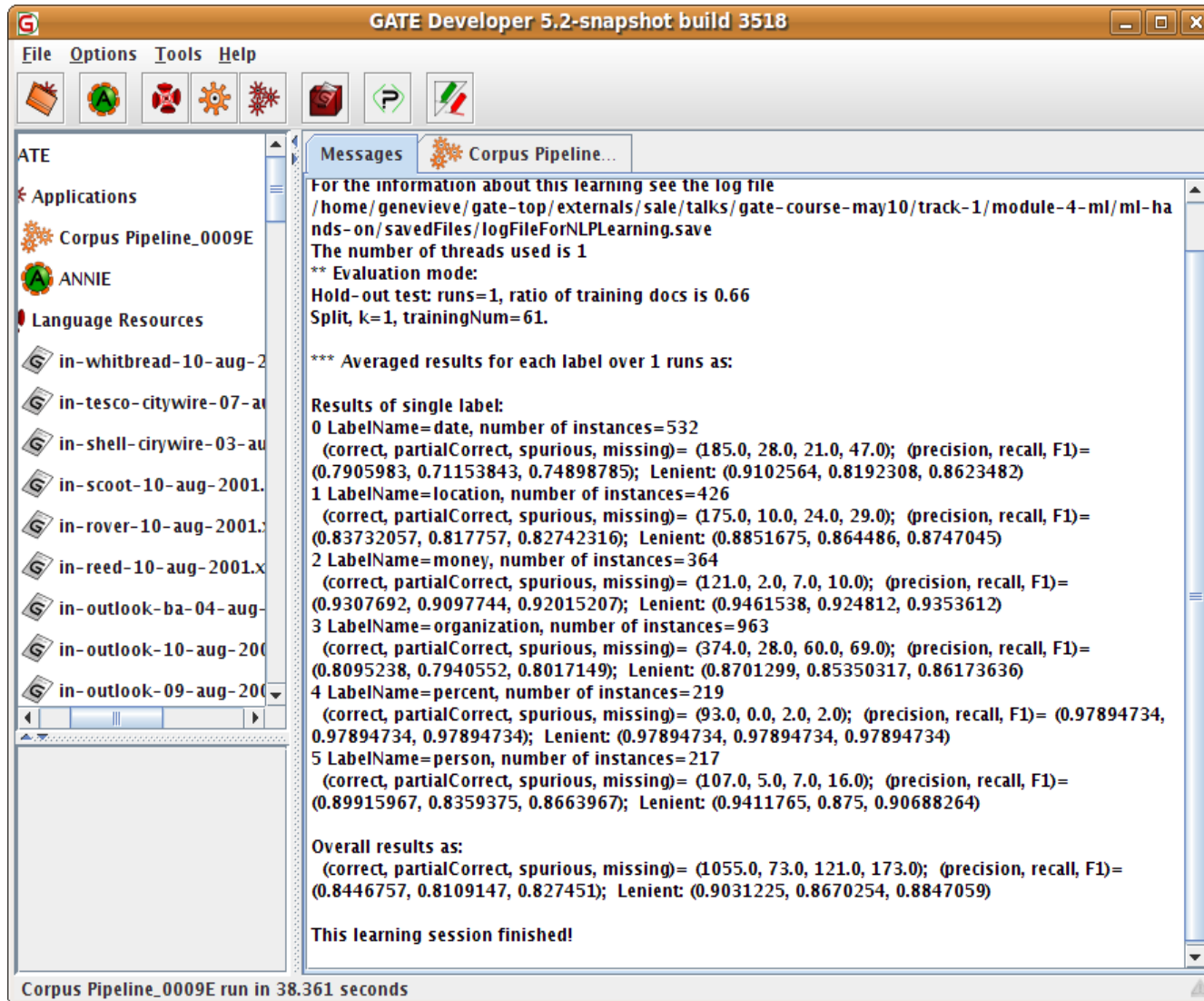
Run this Application

Serial Application Editor Initialisation Parameters

Corpus Pipeline_0008E run in 80.858 seconds

- **Make sure the corpus is selected**
- **The inputASName is blank because the attributes and classes are in the default annotation set**
- **Select "EVALUATION" for the learningMode**
- **OutputASName should be the same as inputASName in evaluation mode**
- **Run the application!**

Inspecting the results



GATE Developer 5.2-snapshot build 3518

File Options Tools Help

Messages **Corpus Pipeline...**

For the information about this learning see the log file
 /home/genevieve/gate-top/externals/sale/talks/gate-course-may10/track-1/module-4-ml/ml-hands-on/savedFiles/logFileForNLPLearning.save
 The number of threads used is 1
 ** Evaluation mode:
 Hold-out test: runs=1, ratio of training docs is 0.66
 Split, k=1, trainingNum=61.

*** Averaged results for each label over 1 runs as:

Results of single label:

0 LabelName=date, number of instances=532
 (correct, partialCorrect, spurious, missing)= (185.0, 28.0, 21.0, 47.0); (precision, recall, F1)=
 (0.7905983, 0.71153843, 0.74898785); Lenient: (0.9102564, 0.8192308, 0.8623482)

1 LabelName=location, number of instances=426
 (correct, partialCorrect, spurious, missing)= (175.0, 10.0, 24.0, 29.0); (precision, recall, F1)=
 (0.83732057, 0.817757, 0.82742316); Lenient: (0.8851675, 0.864486, 0.8747045)

2 LabelName=money, number of instances=364
 (correct, partialCorrect, spurious, missing)= (121.0, 2.0, 7.0, 10.0); (precision, recall, F1)=
 (0.9307692, 0.9097744, 0.92015207); Lenient: (0.9461538, 0.924812, 0.9353612)

3 LabelName=organization, number of instances=963
 (correct, partialCorrect, spurious, missing)= (374.0, 28.0, 60.0, 69.0); (precision, recall, F1)=
 (0.8095238, 0.7940552, 0.8017149); Lenient: (0.8701299, 0.85350317, 0.86173636)

4 LabelName=percent, number of instances=219
 (correct, partialCorrect, spurious, missing)= (93.0, 0.0, 2.0, 2.0); (precision, recall, F1)= (0.97894734,
 0.97894734, 0.97894734); Lenient: (0.97894734, 0.97894734, 0.97894734)

5 LabelName=person, number of instances=217
 (correct, partialCorrect, spurious, missing)= (107.0, 5.0, 7.0, 16.0); (precision, recall, F1)=
 (0.89915967, 0.8359375, 0.8663967); Lenient: (0.9411765, 0.875, 0.90688264)

Overall results as:
 (correct, partialCorrect, spurious, missing)= (1055.0, 73.0, 121.0, 173.0); (precision, recall, F1)=
 (0.8446757, 0.8109147, 0.827451); Lenient: (0.9031225, 0.8670254, 0.8847059)

This learning session finished!

Corpus Pipeline_0009E run in 38.361 seconds

- The application may take a few minutes to run
- **When it is finished, switch to the “Messages” tab to examine the results**

How well did we do?

- Here is my previous result:

(precision, recall, F1) = (0.89, 0.75, 0.82)

- These figures look pretty good, but what do they mean?
- Next we will discuss evaluation measures
- Then we will run the PR in different modes
- Then we will see if we can improve these numbers



Evaluation in Machine Learning



Recap of Evaluation in GATE

- Evaluation is an important part of information extraction work
 - We need to find out how good our application is by comparing its annotations to the “right answers” (manually prepared or corrected annotations)
 - Sometimes we need to compare the work of different human annotators, to see how consistent they are
- We use similar functions for both types of evaluation tasks



Evaluation Mode

- We ran the machine learning PR in evaluation mode earlier
- We specified how the PR should run evaluation in the configuration file
- Once we had run the application, we obtained evaluation statistics in the “Messages” tab

What are precision, recall and F1?

- Precision: what proportion of our automatic annotations were correct?
- Recall: what proportion of the correct annotations did our automatic tool create?
- $P = \text{correct} / (\text{correct} + \text{spurious}) = \text{tp} / (\text{tp} + \text{fp})$
- $R = \text{correct} / (\text{correct} + \text{missing}) = \text{tp} / (\text{tp} + \text{fn})$
- where tp = true positives, fp = false positives, fn = false negatives

What are precision, recall and F1?

- F-score is an amalgam of the two measures
- $F_{\beta} = (1+\beta^2)PR / (\beta^2 P + R)$
 - The equally balanced F1 ($\beta = 1$) is the most common F-measure
 - $F1 = 2PR / (P + R)$
- We can also run our own ML evaluation using the Corpus QA tool—let's do that now



Splitting into training and test corpora

- To tell how well a machine learner is performing, you need to train it and test it on different sets of data
- Evaluation mode does this automatically over “folds” of the corpus
- To see a detailed evaluation, we need to split our corpus into two parts: the training corpus and the test corpus; we will train and apply in separate runs

Saving and splitting the corpus

ml-hands-on

Name	Size	Type
▶ corpus	93 items	folder
▶ test	0 items	folder
▶ training	4 items	folder
CreateMention.jape	571 bytes	plain text
ml-config-file.xml	1.8 KB	XML document

- **Create new “training” and “test” directories on your computer (somewhere easy to find)**
- **Right click on your corpus, select “Save as XML”, and save the whole corpus in the “training” directory**
- **Use your file manager to move roughly half the documents from “training” into “test” (try to randomise them a little)**



Tidying up

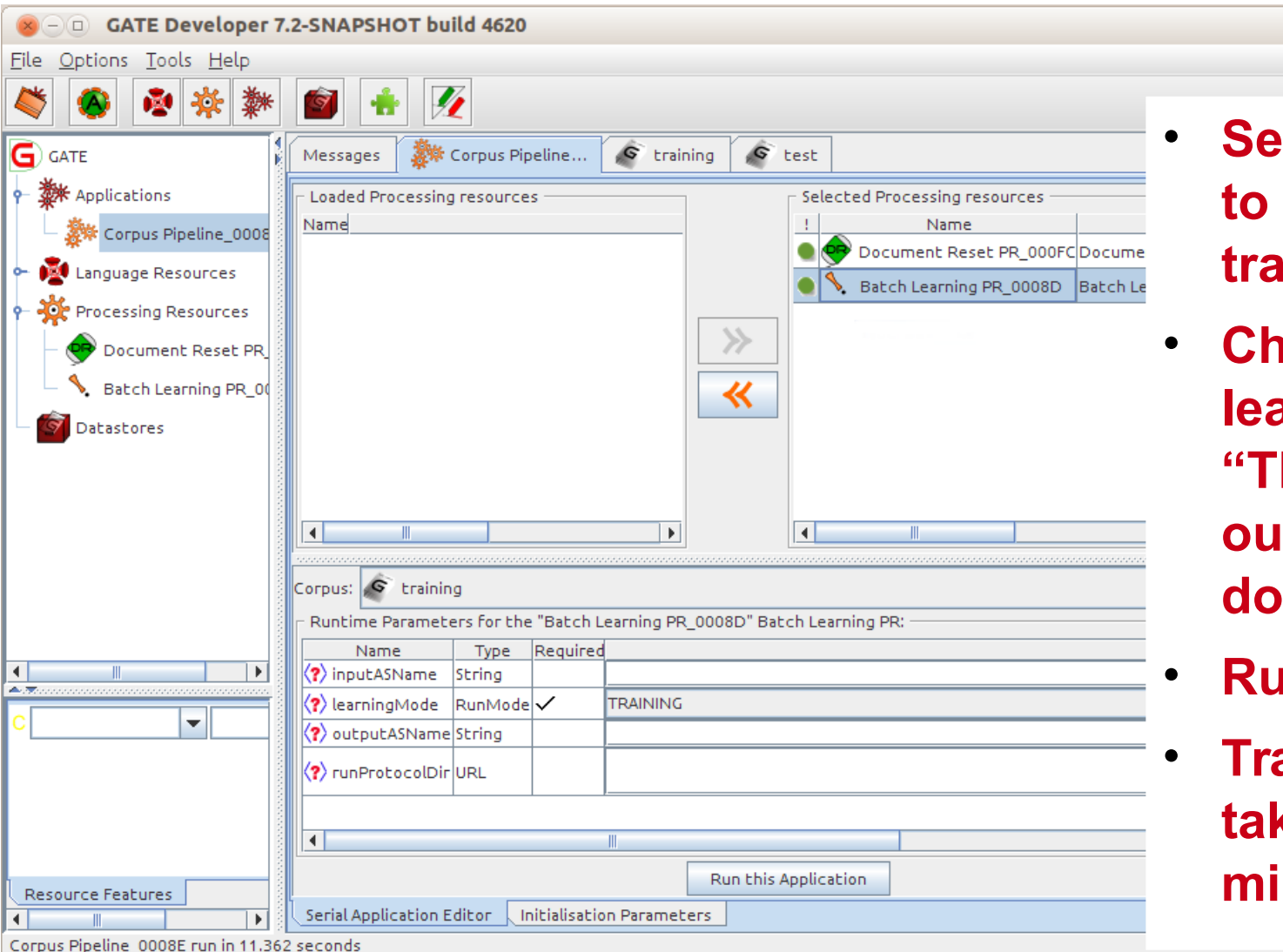
- **Do not close the Batch Learning PR and its corpus pipeline! (We are going to keep using them.)**
- **Close all your open documents and corpora in GATE Developer**
- **Close the modified ANNIE application recursively**
- **Create new GATE corpora called “training” and “test”**
- **Populate each corpus from the appropriate directory (as before, set the encoding to UTF-8!)**

Setting up the application

- **Create a Document Reset PR**
- **Add it to the ML pipeline before the Batch Learning PR**
- **Edit the Document Reset PR's setsToRemove parameter to include just “ML”**
- **Edit the setsToKeep parameter to be an empty list**

Running the ML PR in Training Mode

- **Set your pipeline to run on the training corpus**
- **Change the PR's learningMode to "TRAINING" (the outputASName doesn't matter)**
- **Run the pipeline**
- **Training may take a few minutes**

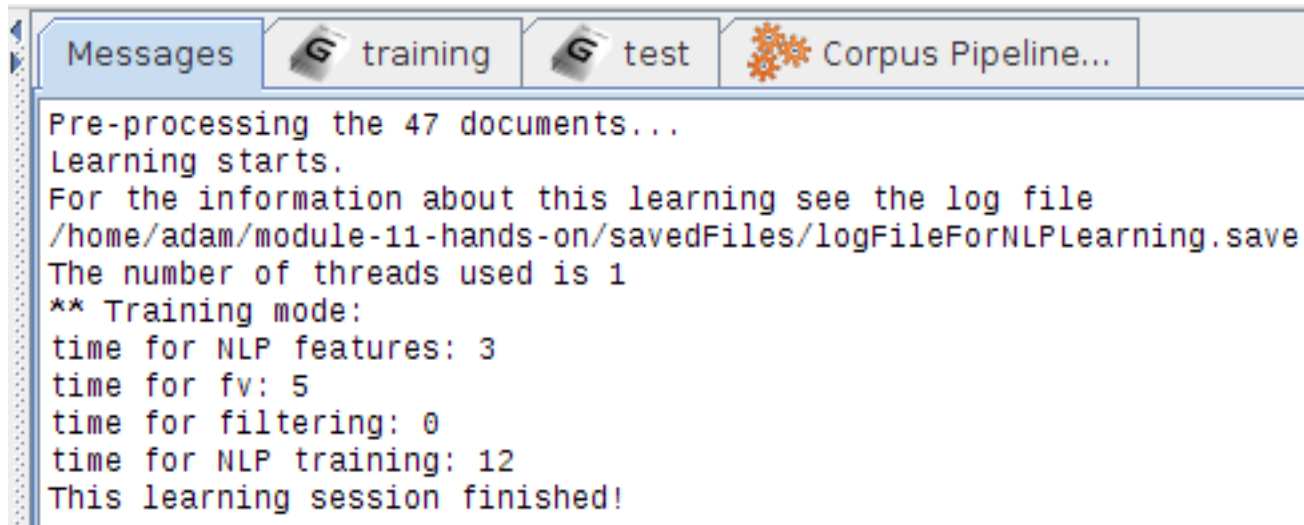


The screenshot shows the GATE Developer interface. The left sidebar displays a tree view with 'Applications' expanded to 'Corpus Pipeline_0008E'. The main window shows the 'Corpus Pipeline...' configuration. The 'Loaded Processing resources' list is empty, and the 'Selected Processing resources' list contains 'Document Reset PR_000FC' and 'Batch Learning PR_0008D'. The 'Corpus' is set to 'training'. Below, the 'Runtime Parameters for the "Batch Learning PR_0008D" Batch Learning PR:' table is visible.

Name	Type	Required	Value
inputASName	String		
learningMode	RunMode	✓	TRAINING
outputASName	String		
runProtocolDir	URL		

At the bottom, there is a 'Run this Application' button and a 'Serial Application Editor' window showing 'Initialisation Parameters'. The status bar at the bottom indicates 'Corpus Pipeline_0008E run in 11.362 seconds'.

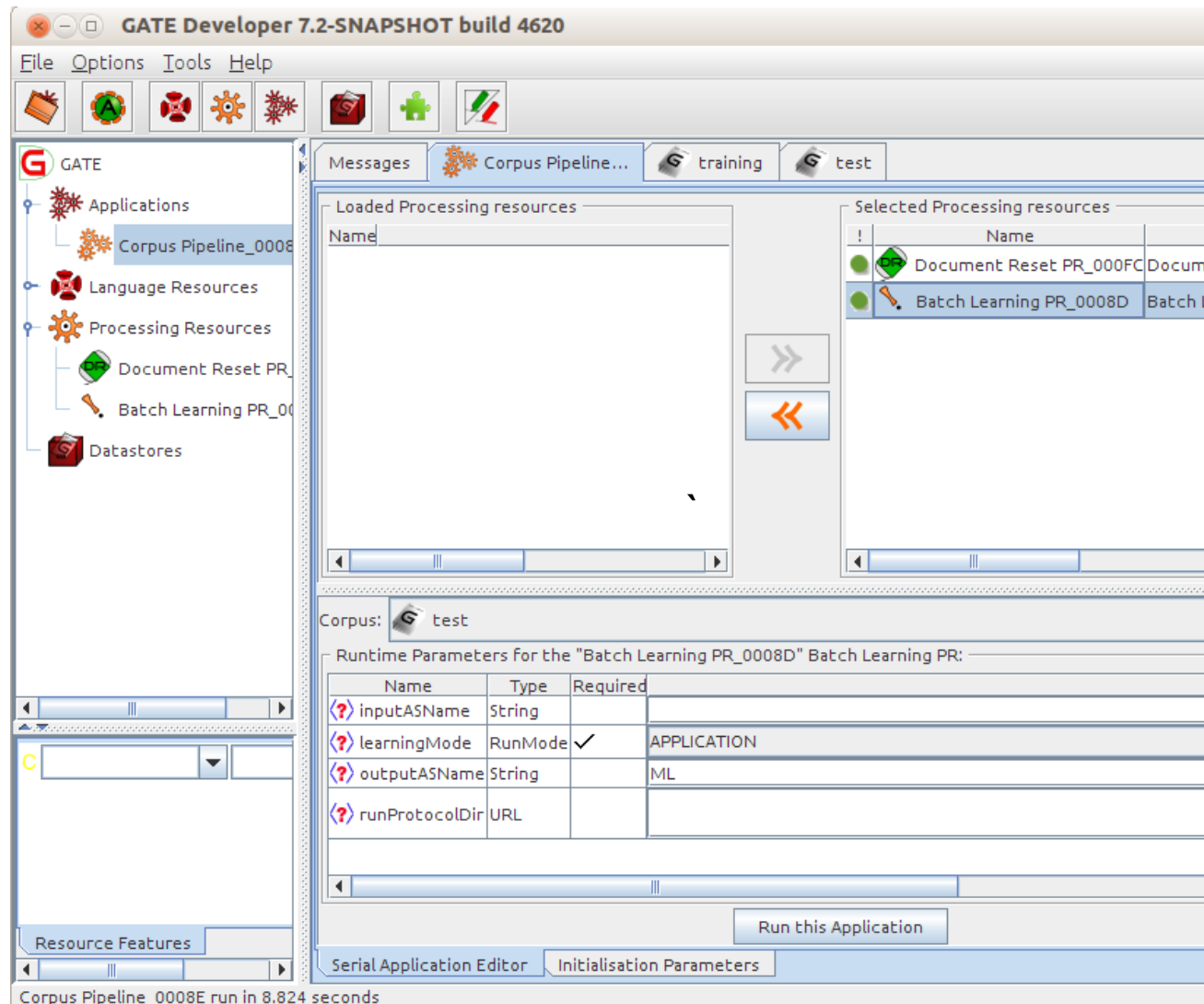
Finished Training!

A screenshot of the GATE software's Messages window. The window has a title bar with tabs for "Messages", "training", "test", and "Corpus Pipeline...". The "Messages" tab is active, showing a log of training progress. The text in the window is as follows:

```
Pre-processing the 47 documents...
Learning starts.
For the information about this learning see the log file
/home/adam/module-11-hands-on/savedFiles/logFileForNLPlearning.save
The number of threads used is 1
** Training mode:
time for NLP features: 3
time for fv: 5
time for filtering: 0
time for NLP training: 12
This learning session finished!
```

- Because we saved the documents after running our modified ANNIE, they already have the instances and attributes for ML
- This time there are no evaluation results in the messages tab (because we are only training the model)
- Note the “savedFiles” directory beside the XML configuration file
 - Training mode saves the model there
 - Application mode reads it from there
 - The runProtocolDir parameter (in recent versions of GATE) can change the location

Running the ML PR in Application Mode



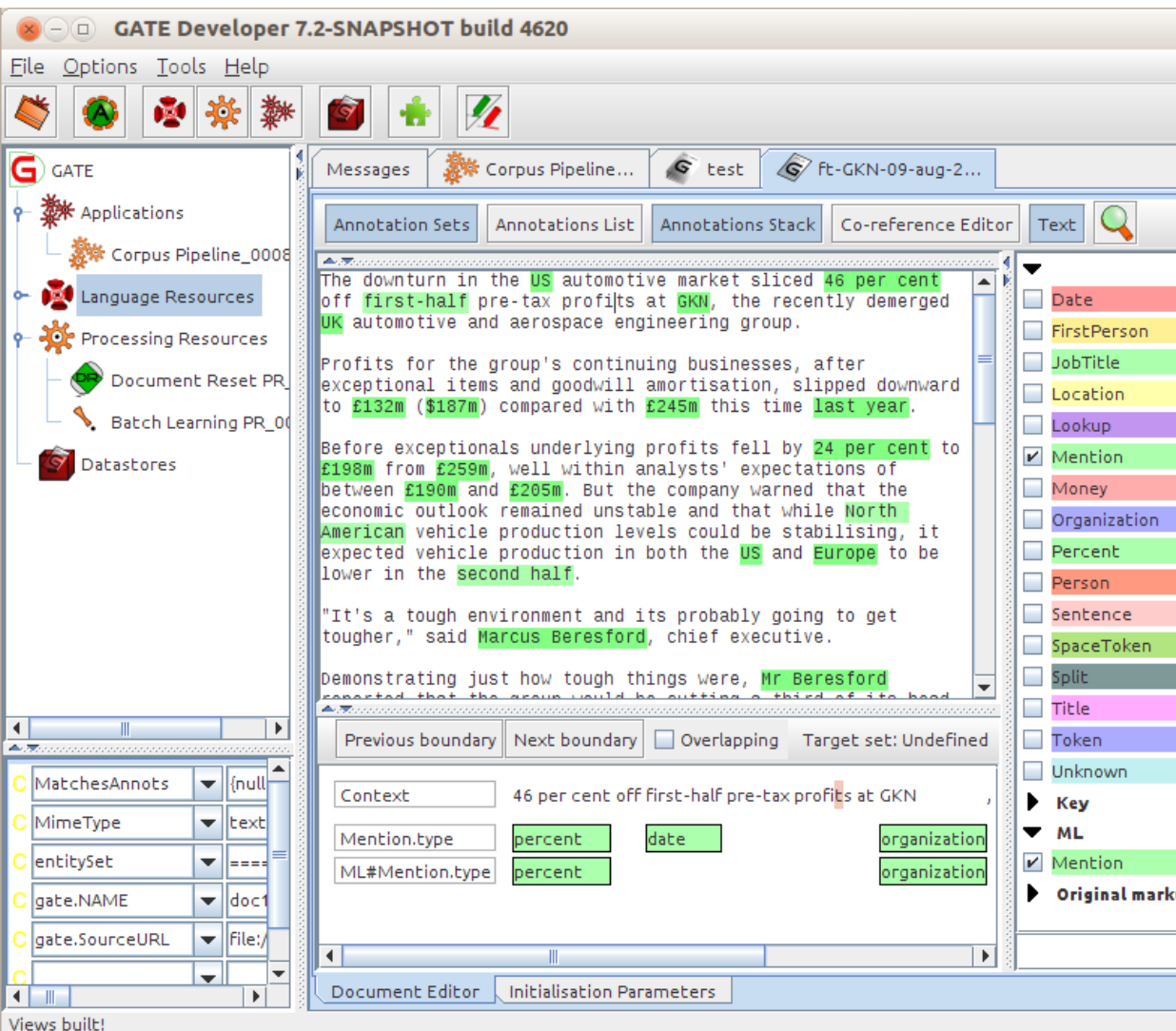
The screenshot shows the GATE Developer interface with the following configuration:

- Corpus:** test
- Runtime Parameters for the "Batch Learning PR_0008D" Batch Learning PR:**

Name	Type	Required	Value
inputASName	String		
learningMode	RunMode	<input checked="" type="checkbox"/>	APPLICATION
outputASName	String		ML
runProtocolDir	URL		
- Buttons:** Run this Application
- Bottom Bar:** Serial Application Editor, Initialisation Parameters
- Status Bar:** Corpus Pipeline_0008E run in 8.824 seconds

- **Change corpus to "test"**
- **Change learningMode to "APPLICATION"**
- **Set outputASName to "ML": your new (automatic) annotations will go here so they don't get mixed up with the existing ones**
- **Application mode is faster than training mode**

Examining the results of application

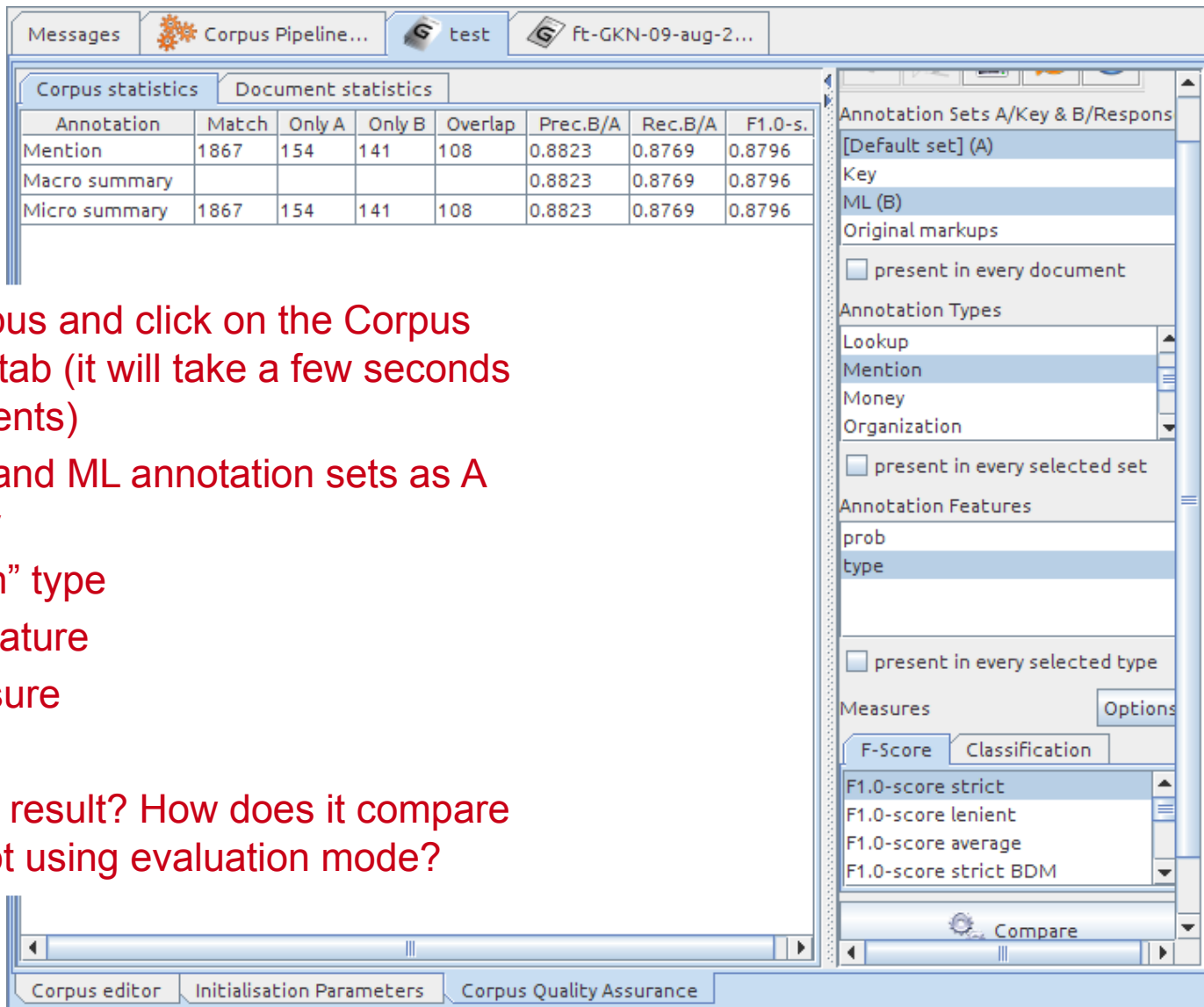


The screenshot shows the GATE Developer interface. The main window displays a document with several annotations. The Annotations Stack on the right shows a list of annotation types, with 'Mention' checked. The Annotations List at the bottom shows the following table:

Context	46 per cent	off first-half	pre-tax profits	at GKN
Mention.type	percent	date		organization
ML#Mention.type	percent			organization

- **Examine a document from the test corpus**
- You should have a new “ML” AS with Mention annotations
- The original Mention annotations (in the default AS) are similar but not always identical!
- The Annotations Stack is good for comparing them
- How similar do they appear to be? Do you think you will get a good result?

Comparing the Sets with Corpus QA



The screenshot shows the GATE Corpus QA interface. The main window displays a table of statistics for the 'test' corpus. The table has columns for Annotation, Match, Only A, Only B, Overlap, Prec.B/A, Rec.B/A, and F1.0-s. The rows are Mention, Macro summary, and Micro summary. The statistics for Mention are: Match: 1867, Only A: 154, Only B: 141, Overlap: 108, Prec.B/A: 0.8823, Rec.B/A: 0.8769, F1.0-s: 0.8796. The Macro summary and Micro summary rows have the same values for Prec.B/A, Rec.B/A, and F1.0-s, but no values for Match, Only A, Only B, or Overlap.

Annotation	Match	Only A	Only B	Overlap	Prec.B/A	Rec.B/A	F1.0-s.
Mention	1867	154	141	108	0.8823	0.8769	0.8796
Macro summary					0.8823	0.8769	0.8796
Micro summary	1867	154	141	108	0.8823	0.8769	0.8796

The right-hand panel shows the configuration for the 'Annotation Sets A/Key & B/Respons' comparison. The 'Default set' (A) is selected, and the 'ML (B)' set is also selected. The 'Mention' type is selected under 'Annotation Types', and the 'type' feature is selected under 'Annotation Features'. The 'F-Score' measure is selected under 'Measures', and the 'F1.0-score strict' option is selected. The 'Compare' button is visible at the bottom of the panel.

- Select the test corpus and click on the Corpus Quality Assurance tab (it will take a few seconds to scan the documents)
- Select the Default and ML annotation sets as A and B, respectively
- Select the “Mention” type
- Select the “type” feature
- Choose an F-measure
- Click on Compare
- Did you get a good result? How does it compare to the result you got using evaluation mode?



Using Annotation Diff to examine performance

Annotation Difference

Key doc: ft-BT-briefing-02-a... Key set: [Default set] Type: Mention Weight:

Resp. doc: ft-BT-briefing-02-a... Resp. set: ML-results Features: all some none 1.0

Start	End	Key	Features	=?	Start	End	Response	Features
1517	1519	BT	{class=organization}	=	1517	1519	BT	{class=organization, prob=1.0}
171	173	2p	{class=money}	=	171	173	2p	{class=money, prob=1.0}
1956	1972	Deutsche · Telekom	{class=organization}	=	1956	1972	Deutsche · Telekom	{class=organization, prob=1.0}
46	55	yesterday	{class=date}	=	46	55	yesterday	{class=date, prob=1.0}
1322	1327	Oftel	{class=organization}	=	1322	1327	Oftel	{class=organization, prob=1.0}
867	882	January · 22 · 2001	{class=date}	=	867	882	January · 22 · 2001	{class=date, prob=1.0}
1198	1203	Scoot	{class=organization}	=	1198	1203	Scoot	{class=organization, prob=1.0}
514	524	Amazon.com	{class=organization}	~	514	520	Amazon	{class=organization, prob=1.0}
1753	1761	Scoot · UK	{class=organization}	-?				
1181	1195	late · last · year	{class=date}	-?				
1007	1017	Air · Canada	{class=organization}	-?				
1924	1926	DT	{class=organization}	-?				
				?-	1499	1511	0800 · 192 · 192	{class=money, prob=1.0}
482	488	Amazon	{class=organization}	<>	482	488	Amazon	{class=location, prob=0.99999946}
800	806	Amazon	{class=organization}	<>	800	806	Amazon	{class=location, prob=0.99999905}
756	762	Amazon	{class=organization}	<>	756	762	Amazon	{class=location, prob=1.0}

Correct: 36 Recall Precision F-measure
 Partially correct: 1 Strict: 0.82 0.88 0.85
 Missing: 7 Lenient: 0.84 0.90 0.87
 False positives: 4 Average: 0.83 0.89 0.86

93 documents loaded

- Switch to the “Document statistics” tab
- Choose a document
- Click on the Annotation Diff icon
- What kind of mistakes did your application make?

Using Annotation Diff...

- “Correct”: the response annotation has the right feature and span
- “Partially correct”: response has the right feature and overlapping but not exactly matched span; this counts as correct in the “lenient” scoring
- “Missing”: key annotation+feature is missing from the response (a.k.a. “false negative”)
- “False positive”: response annotation+feature shouldn't be there (a.k.a. “spurious”)



Varying the configuration file

- Now we are going to experiment with varying the configuration file to see if we can produce varied results
- You can edit the configuration file in your favourite text editor
- Make sure you save your changes then **reinitialise the PR** (this reads the file again and updates the configuration used inside GATE)

Confidence Thresholds

```
<PARAMETER name="thresholdProbabilityEntity" value="0.2"/>  
<PARAMETER name="thresholdProbabilityBoundary" value="0.42"/>  
<PARAMETER name="thresholdProbabilityClassification" value="0.5"/>
```

- Each classifier will provide confidence ratings—how likely is a result to be correct; we must determine how certain is good enough
- Depending on the application we might prefer to include or exclude annotations for which the learner is not too sure
- `thresholdProbabilityBoundary` and `thresholdProbabilityEntity` are thresholds for chunk learning
- `thresholdProbabilityClassification` applies to classification tasks, such as sentiment or genre detection, author identification, language identification

Classification tasks

- Opinion mining
 - Example: the documents contain spans of text (such as individual sentences or longer consumer reviews) which you want to classify as positive, neutral, or negative
- Genre detection: classify each document or section as a type of news
- Author identification

Classification tasks

- `thresholdProbabilityClassification`: the “pickiness” of the classifiers
- increasing this generally raises precision and reduces recall
- decreasing this generally increases recall and reduces precision
- `thresholdProbabilityBoundary` and `thresholdProbabilityEntity`: ignored

Classification tasks

- `<SURROUND VALUE="FALSE"/>`
- the class boundaries are known
- INSTANCE-TYPE: type of annotation that covers each span of text to classify (Sentence, p (paragraph), etc.)
- We typically use NGRAM elements as attributes
- The GATE user guide gives examples



Hands-on: text classification

- Close open applications, PRs, and LRs in GATE
- If you've closed GATE since the last exercise, you need the ANNIE, Tools, and Learning plugins for this exercise
- If you haven't closed GATE, load the Tools plugin
- Create new empty “training” and “test” corpora
- Populate them from “language/training-corpus” and “language/test-corpus” directories in the hands-on material
- Set the encoding to UTF-8 before you click OK
- Inspect the documents: the Key AS contains Sentence annotations with a lang feature
- Very few documents, but many instances (Sentence annotations)
- Task: language identification

Text classification

- Create a new **Conditional** Corpus Pipeline and add the following PRs:
- Document Reset
- ANNIE English Tokenizer
- ANNIE Sentence Splitter
- Annotation Set Transfer
- Batch Learning PR with “language/ml-language.xml” as the config file
- Examine this config file in an editor and notice how it differs from the NER file

Text classification config file

- Note the changes for text classification:
- `<SURROUND value="false"/>`
- `thresholdProbabilityClassification` is used
- `INSTANCE-TYPE` is `Sentence`

Text classification example

- Training
- We use the Sentence annotations as instances, lang features as ML classes, and the tokenizer's output as attributes
- Check that Document Reset will keep the “Key” AS
- Switch the Sentence Splitter off (red signal light)
- Configure the AS Transfer PR to copy all annotations from “Key” to the default AS
- Set the Batch Learning PR to TRAINING mode
- Set the pipeline to run on the training corpus
- Run the pipeline

Text classification example

- Testing
- Here we create our own Sentence annotations and use ML to classify them
- Switch the Sentence Splitter on (green light)
- Switch the AS Transfer PR off (red light)
- Set the Batch Learning PR to APPLICATION mode
- Leave the inputASName blank (default AS)
- Set the Batch Learning PR's output AS to "Output"
- Set the pipeline to run on the test corpus
- Run the pipeline

Text classification example

- Inspect the test corpus with Corpus QA:
- A = Key, B = Output
- select “Sentence” annotations and the “lang” feature

Text classification example

- In Corpus QA, try Classification → Observed Agreement, click Compare, and look at the “Confusion Matrices” tab
- I get a table like this:

	de	en	fr
de	217	0	0
en	6	72	0
fr	6	0	27

- This shows that 6 English & 6 French sentences were misclassified as German



Further tinkering

- Try lower or higher threshold values
- Try different combinations of attributes

Learning Framework Preview

- We want:
 - More, and more up to date, algorithms
 - Feature selection
 - Faster parameter tuning cycle
- Learning Framework offers:
 - Several Weka algorithms, Mallet algorithms including the sequence learning algo CRF, and LibSVM
 - Easier to add more algorithms as requested
 - GATE isn't really set up for complex ML tuning, but export to ARFF allows Weka to be used for that
 - Feature selection can improve performance for many algos and speeds up SVM—Weka supports this
 - For this to work, Weka and Learning Framework algos must be the same—LF includes many Weka algos and Weka integrates LibSVM
 - Parameter tuning is faster off the ARFF because every time you evaluate a new variant inside of GATE, it scrapes the features off the docs again, which is quite slow
- When will it be available? Hopefully later in the year. Currently being extended to include semantic modeling in the same plugin.