general architecture

GATE

for text engineering

Machine Learning--Chunking

Xingyi Song

Twin Karmakharm

Chunking for NER

- Chunking, means finding parts of text
  - Often used in Named Entity Recognition (NER)
    - E.g. person names in the text
  - Other tasks like
    - Negation range
    - Noun phrases

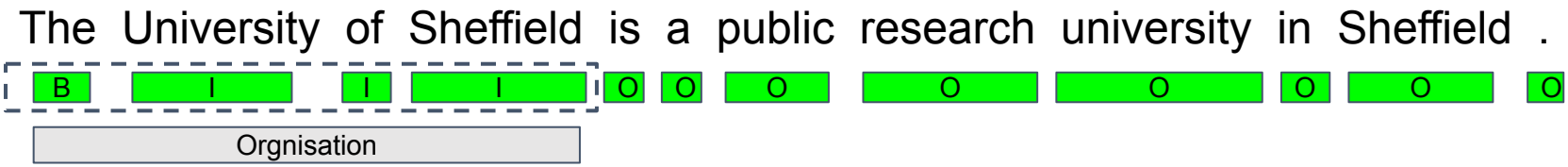
Chunking for NER

The University of Sheffield is a public research university in Sheffield .

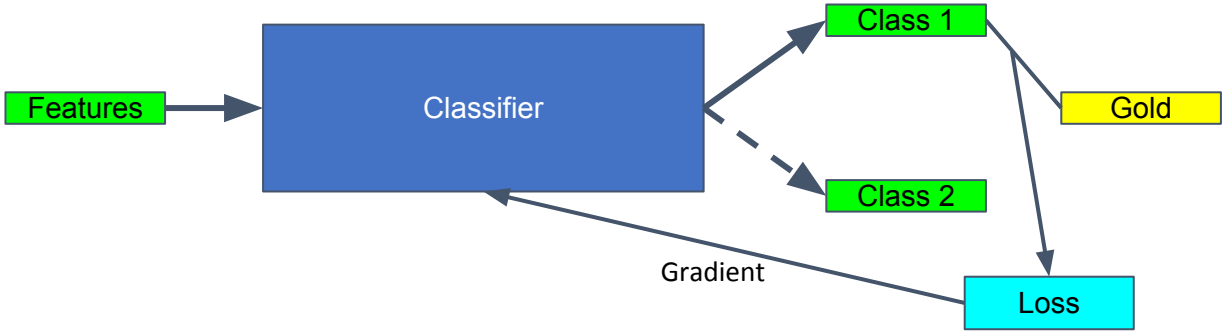
Orgnisation

Chunking for NER

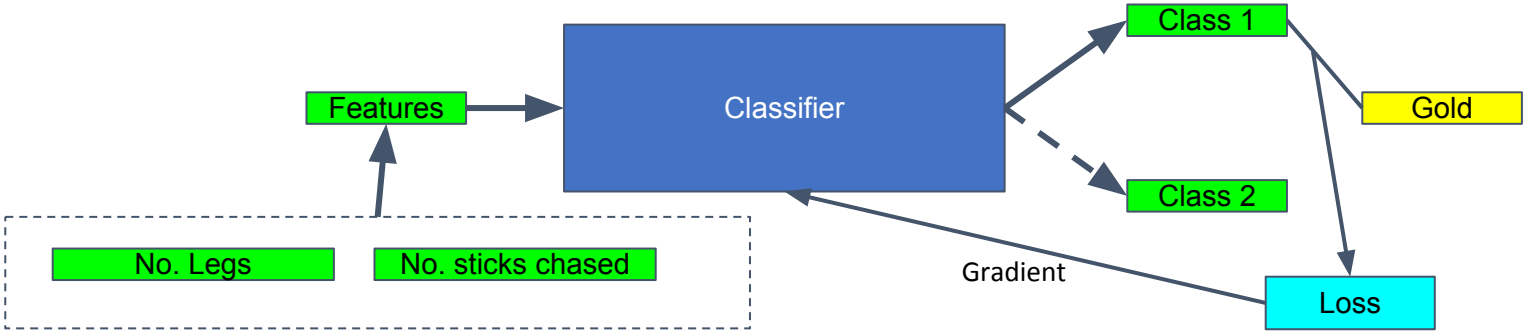
- Chunking = Special classification task
- Identify BIO labels of tokens
  - B= Beginning of the entity
  - I = Inside of the entity
  - O = Outside of the entity



Classification

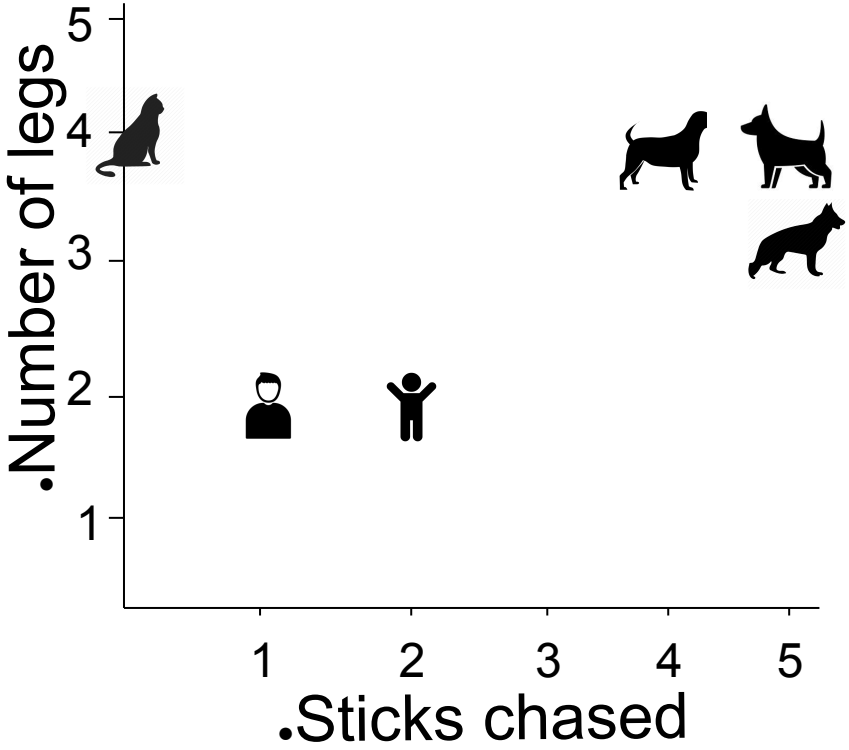


Dog Classification



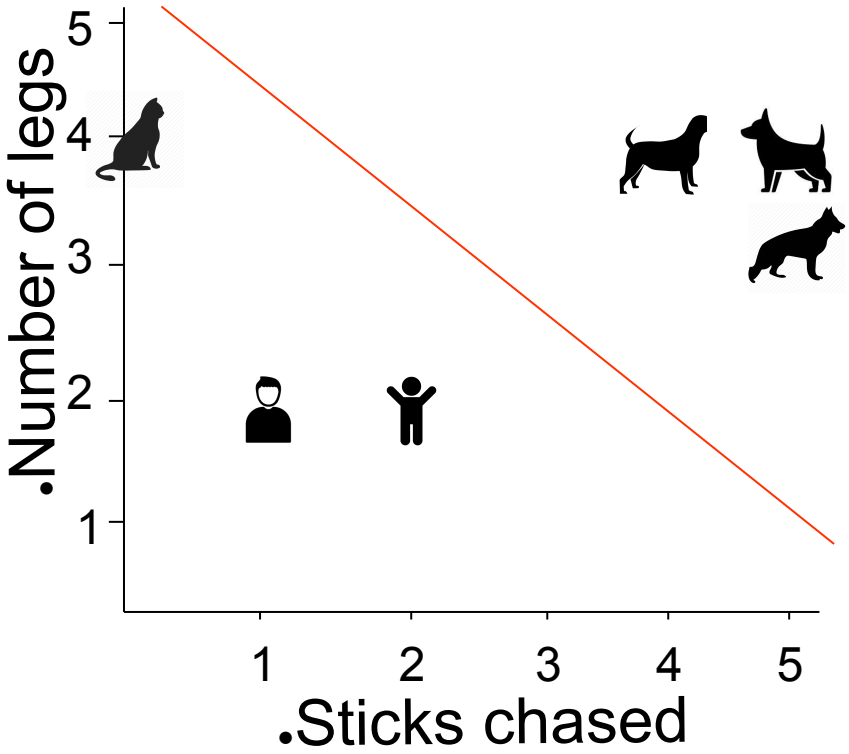
Dog Classification

Feature 1	Feature 2	Target
4	0	no
2	1	no
2	2	no
4	4	yes
4	5	yes
3	5	yes



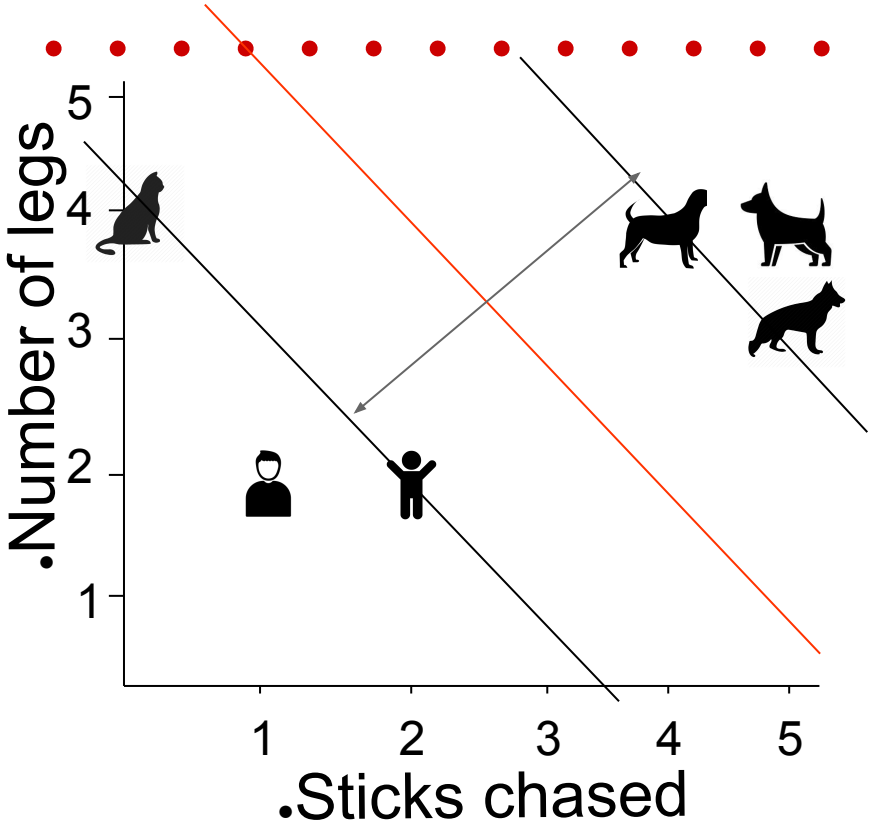
Dog Classification

Feature 1	Feature 2	Target
4	0	no
2	1	no
2	2	no
4	4	yes
4	5	yes
3	5	yes

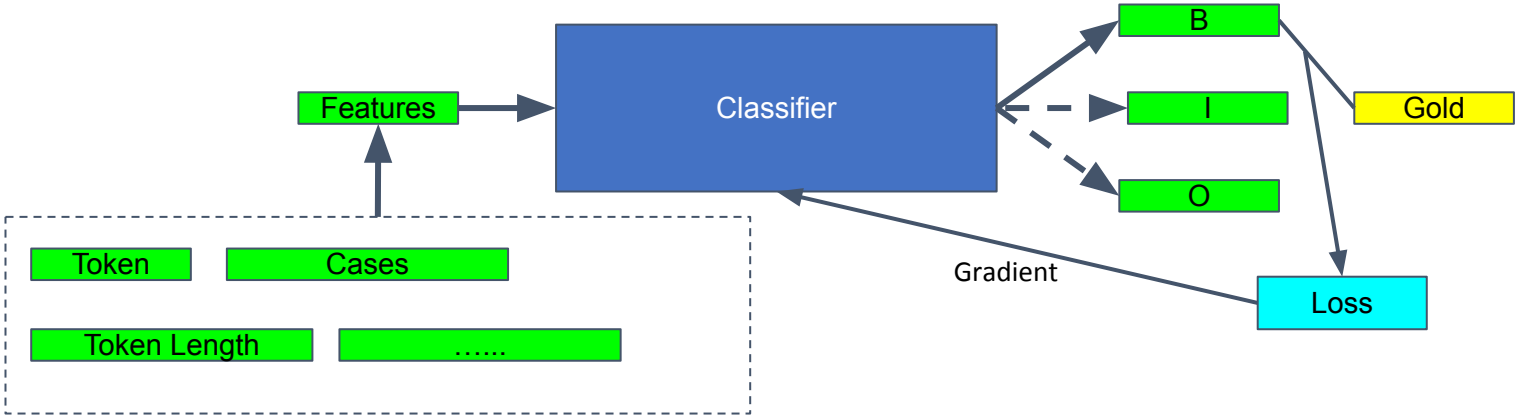


Dog Classification

Feature 1	Feature 2	Target
4	0	no
2	1	no
2	2	no
4	4	yes
4	5	yes
3	5	yes



BIO Classification



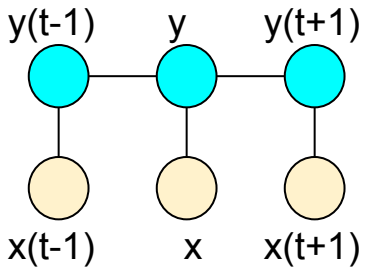
Sequence Classification

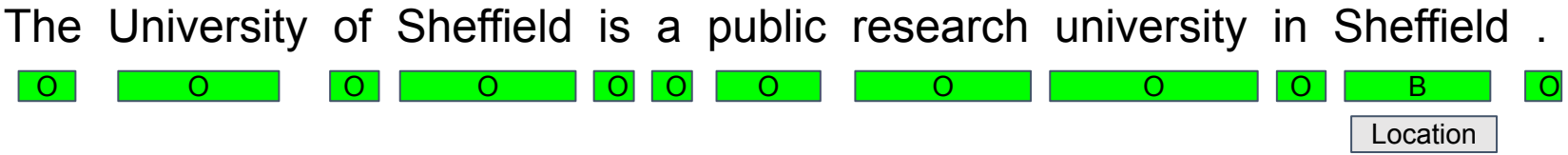
- Consider previous/after tokens as features

The	University	of	Sheffield	is	a	public	research	university	in	Sheffield	.
O	O	O	O	O	O	O	O	O	O	B	O
										Location	

Sequence Classification

- Consider previous/after tokens as features
- Use sequence classification algorithms
  - Conditional random fields
  - Recurrent Neural Network
  - Attention mechanism





Chunking Practical Exercise

- Materials for this exercise are in the folder called “chunking-hands-on”
- You might want to start by closing any applications and corpora from the previous exercise, so we have a fresh start
- Finding Person Mentions using Chunking Training and Application PRs

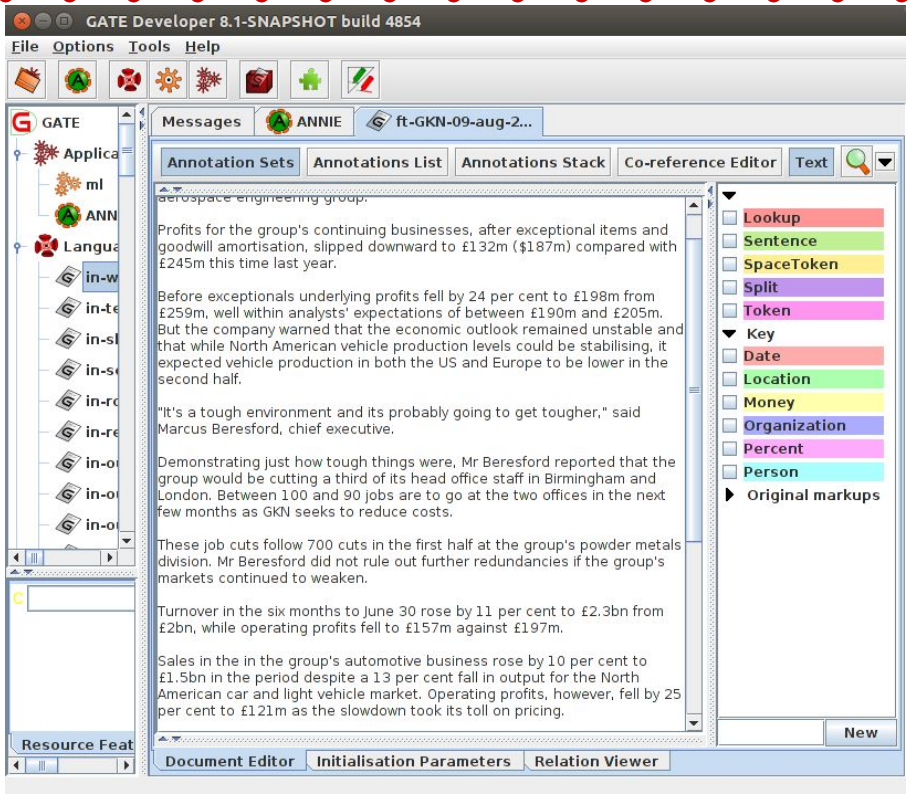
Load the corpus

- Create corpora for training and testing, with sensible names
- Populate them from the training and testing corpora you have in your chunking hands on materials
- Open a document and examine its annotation

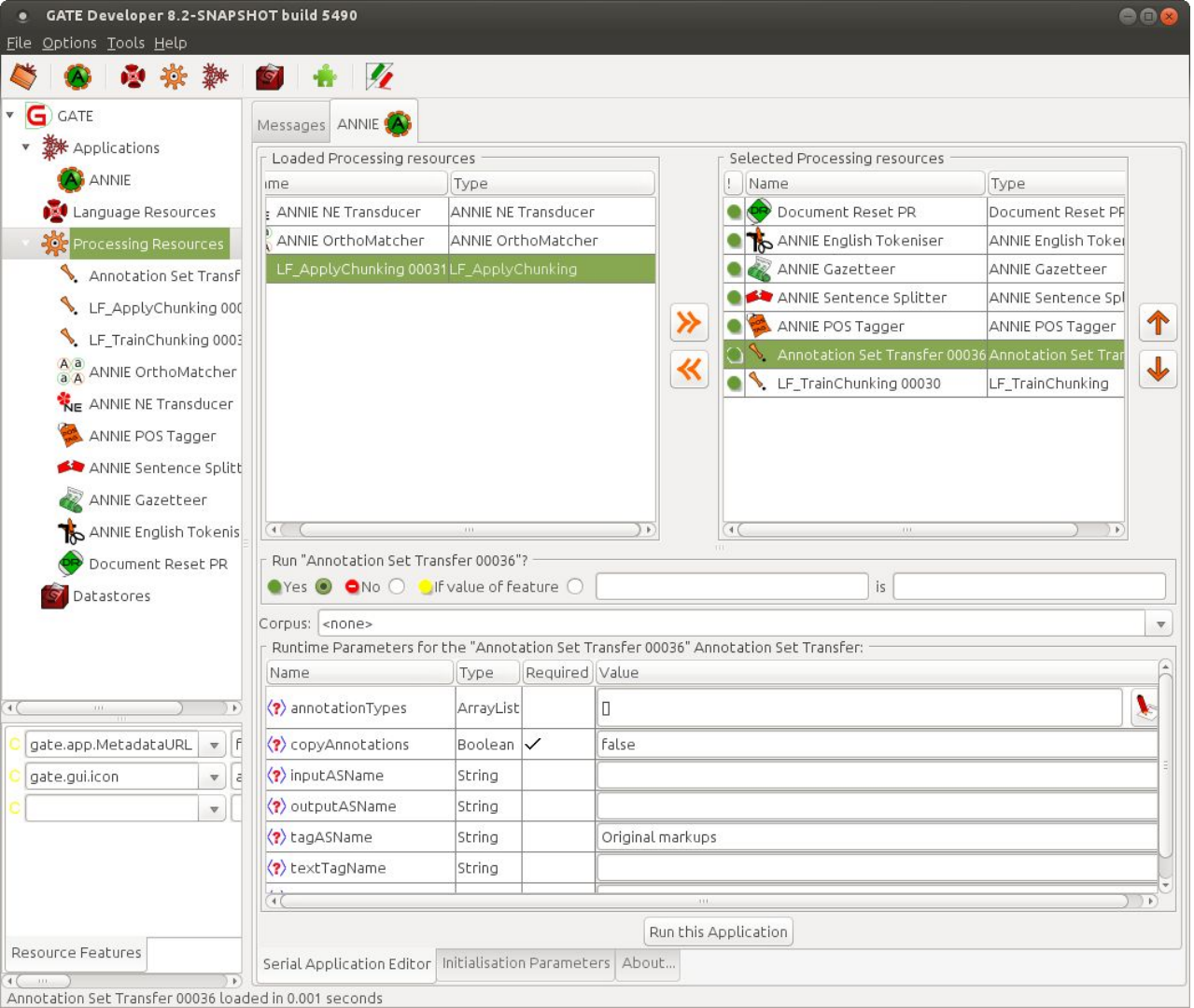
Examining the corpus

- 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

Creating the application



- As previously, if we run ANNIE on the corpus, we have more annotations to work with
- So start by loading ANNIE as the basis for your application
- Again, we don't need the NE transducer or orthomatcher



- Again, we need an Annotation Set Transfer, so create and add one
- Then create training chunking PR

Annotation Set Transfer

- We'll use the annotation set transfer to copy the Person and Organization annotations up to the default annotation set, where we can learn them
- Go ahead and set up your AST now
- Be sure to copy them, not move them!

GATE Developer 8.5-SNAPSHOT build ee930e5

File Options Tools Help

GATE
- Applications
  - ANNIE
- Language Resou
- Processing Resou
  - LF_TrainChunk
  - Annotation Se
  - ANNIE OrthoM
  - ANNIE NE Tran
  - ANNIE POS Tag
  - ANNIE Sentenc
  - ANNIE Gazette
  - ANNIE English
  - Document Res
- Datastores

gate.app.Metadata

gate.gui.icon

Resource Features

Messages | ft-bt-wireless-... | ANNIE

Loaded Processing resources

Name	Type
ANNIE NE Transducer	ANNIE NE Transducer
ANNIE OrthoMatcher	ANNIE OrthoMatcher

Selected Processing resources

!	Name	Type
	Document Reset PR	Document Reset
	ANNIE English Tokeniser	ANNIE English To
	ANNIE Gazetteer	ANNIE Gazetteer
	ANNIE Sentence Splitter	ANNIE Sentence
	ANNIE POS Tagger	ANNIE POS Tagge
	Annotation Set Transfer 0009F	Annotation Set T
	LF_TrainChunking 000A0	LF_TrainChunkin

Run "LF_TrainChunking 000A0"?

Yes ◉ No ○ If value of feature ○ is

Corpus: training

Runtime Parameters for the "LF_TrainChunking 000A0" LF_TrainChunking:

Name	Type	Required	Value
algorithmParameters	String		
classAnnotationTypes	List	✓	[Person, Organization]
dataDirectory	URL	✓	file:/home/genevieve/svn/sale/talks/gate-course-jun18/module-5-ml/chunki
featureSpecURL	URL	✓	file:/home/genevieve/svn/sale/talks/gate-course-jun18/module-5-ml/chunki
inputASName	String		
instanceType	String	✓	Token
scaleFeatures	ScalingMethod	✓	NONE
seqEncoder	SeqEncoderEnum		BIO
sequenceSpan	String		Sentence
trainingAlgorithm	AlgorithmClassification		MalletCRF_SEQ_MR

Run this Application

Serial Application Editor | Initialisation Parameters | About...

ANNIE run in 0.485 seconds

Instead of targetFeature,

we have classAnnotationTypes

Chunking training parameters

- For classification, the class to learn is in a feature on the instance, is specified to the PR in the targetFeature parameter
- But for chunking, the class or classes to learn take the form of an annotation type.

Chunking training parameters

- Set the classAnnotationTypes now
- Set the dataDirectory to where you want to save your model, and set the featureSpecURL (there's a feature spec to get you started in the hands on materials)
- Set instanceType. What do you think it should be?

Sequence Spans

- sequenceSpan is only relevant when using sequence learners
- Sequence learners classify each instance in the span by making use of the others
- For example, a noun phrase might be more likely to follow a determiner than a preposition, or a person name might be more likely to follow the word “Mrs”

Sequence Spans

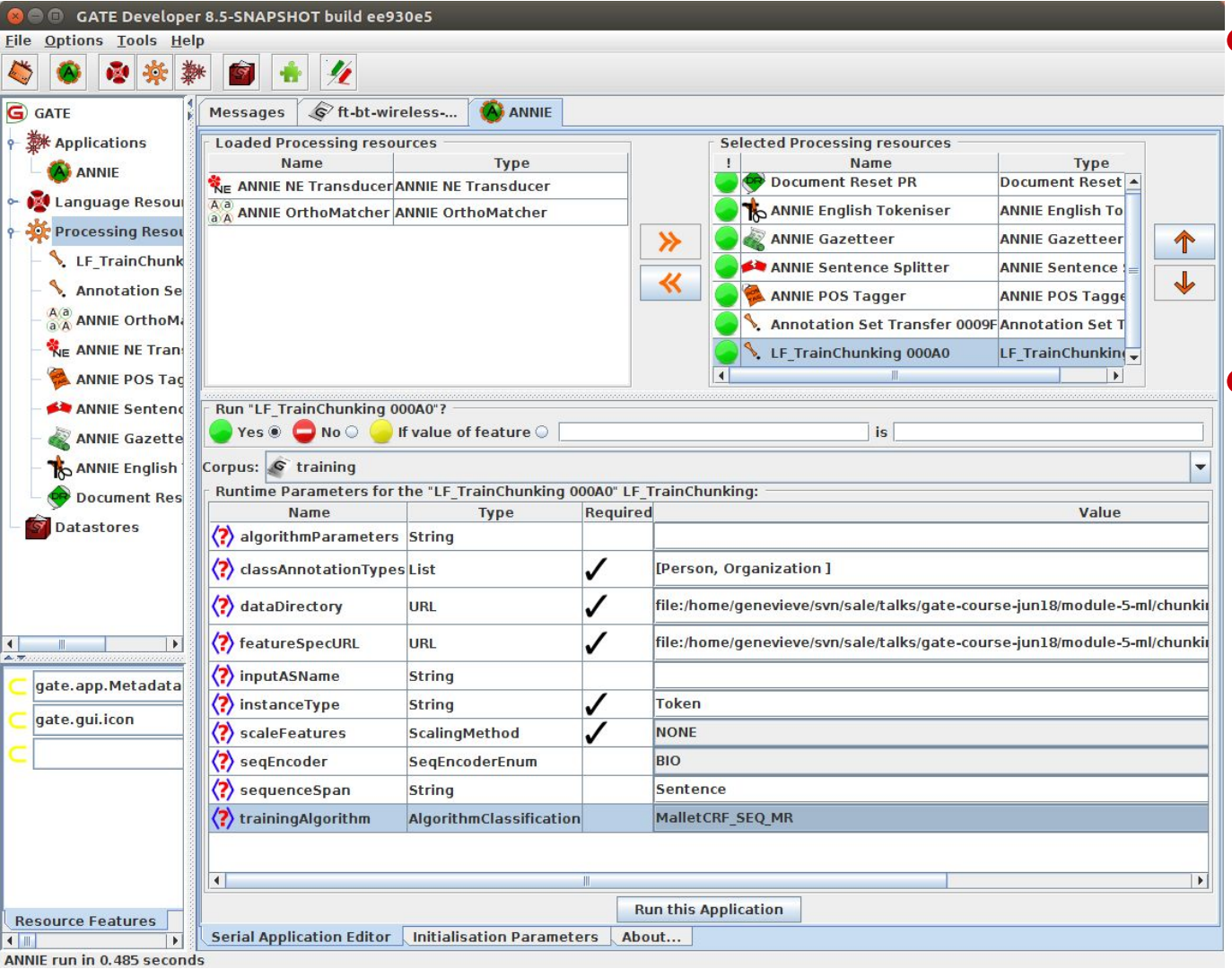
- We'll try the Conditional Random Fields sequence learner
  - You don't have to use a sequence learner for chunking though
  - What do you think would be a good sequence span?

Sequence Spans

- A sequence span shouldn't be longer than necessary
- Sentence would be a good span for our task
- Fortunately, ANNIE creates sentence annotations for us, so those are available to use

Set sequenceSpan to “Sentence”

```
<ML-CONFIG>

<ATTRIBUTE>
<TYPE>Token</TYPE>
<FEATURE>category</FEATURE>
<DATATYPE>nominal</DATATYPE>
</ATTRIBUTE>

<ATTRIBUTE>
<TYPE>Token</TYPE>
<FEATURE>kind</FEATURE>
<DATATYPE>nominal</DATATYPE>
</ATTRIBUTE>

<ATTRIBUTE>
<TYPE>Token</TYPE>
<FEATURE>length</FEATURE>
<DATATYPE>numeric</DATATYPE>
</ATTRIBUTE>

<ATTRIBUTE>
<TYPE>Token</TYPE>
<FEATURE>orth</FEATURE>
<DATATYPE>nominal</DATATYPE>
</ATTRIBUTE>

<ATTRIBUTE>
<TYPE>Token</TYPE>
<FEATURE>string</FEATURE>
<DATATYPE>nominal</DATATYPE>
</ATTRIBUTE>

</ML-CONFIG>
```



- Make sure you have selected the training corpus
- Run the application!

Chunking application

- Now switch off the training PR and create and add the chunking application PR
- (You can switch off the annotation set transfer too)
- It doesn't have a targetFeature parameter like the classification application PR did
- You don't need to tell it what type to create because the model knows it from training!

Chunking application

- Set dataDirectory to the location where you told the training PR to put the model
- Set the sequence span

Now run this on the test corpus

GATE Developer 8.2-SNAPSHOT build 5490

File Options Tools Help

ANNIE

Language Resources

- in-tesco-citywire-07
- in-scoot-10-aug-200
- in-reed-10-aug-2001
- in-outlook-10-aug-2
- in-oil-09-aug-2001.x
- in-german-bank-10-
- in-bayer-10-aug-200
- in-airlines-08-aug-20
- in-GKN-citywire-10-
- gu-w&d-10-aug-200
- gu-telewest-10-aug
- gu-synergie-10-aug
- gu-singtel-10-aug-2
- gu-scoot-10-aug-20
- gu-ryanair.xml_0008

gate.app.MetadataURL

gate.gui.icon

Resource Features

Messages | ANNIE

Loaded Processing resources

Name	Type
ANNIE NE Transducer	ANNIE NE Transducer
ANNIE OrthoMatcher	ANNIE OrthoMatcher
Annotation Set Transfer 00036	Annotation Set Trans
LF_TrainChunking 00030	LF_TrainChunking

Selected Processing resources

!	Name	Type
●	Document Reset PR	Document Reset PR
●	ANNIE English Tokeniser	ANNIE English Tokeniser
●	ANNIE Gazetteer	ANNIE Gazetteer
●	ANNIE Sentence Splitter	ANNIE Sentence Splitter
●	ANNIE POS Tagger	ANNIE POS Tagger
●	LF_ApplyChunking 00031	LF_ApplyChunking

Run "LF_ApplyChunking 00031"?

Yes ● No ○ If value of feature ○ _____ is _____

Corpus: training

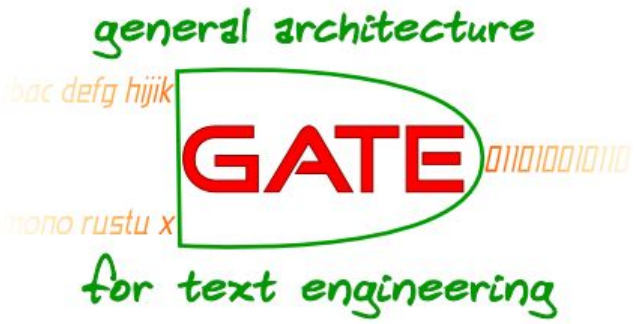
Runtime Parameters for the "LF_ApplyChunking 00031" LF_ApplyChunking:

Name	Type	Required	Value
algorithmParameters	String		
confidenceThreshold	Double	✓	0.0
dataDirectory	URL	✓	file:/home/genevieve/svn/sale/talks/gate-course-jun16/module-3-ml-barbour/chunkir
inputASName	String		
instanceType	String	✓	Token
outputASName	String		LearningFramework
sequenceSpan	String		Sentence

Run this Application

Serial Application Editor | Initialisation Parameters | About...

ANNIE run in 14.256 seconds



Chunking

Evaluation using Corpus QA

Chunking Evaluation

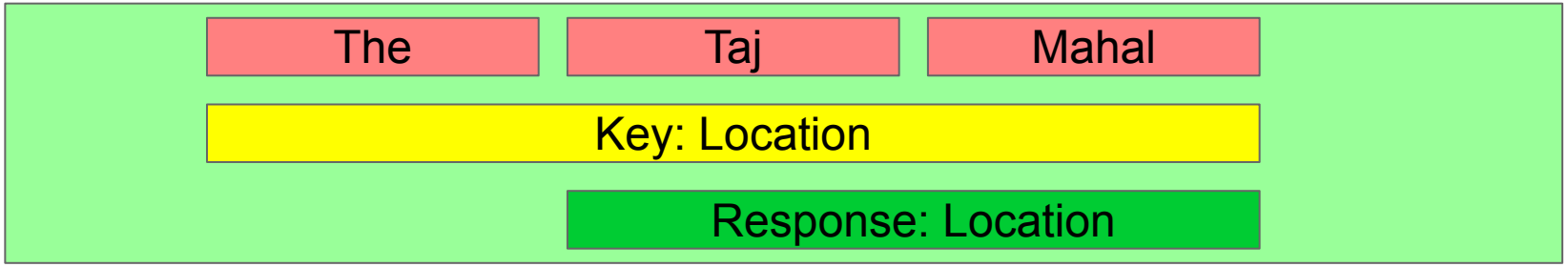
- We don't use a Learning Framework evaluation PR for this chunking task
  - No reason to obtain accuracy over BIOs
- More important measure how well finding named entities
  - there are more ways to be wrong

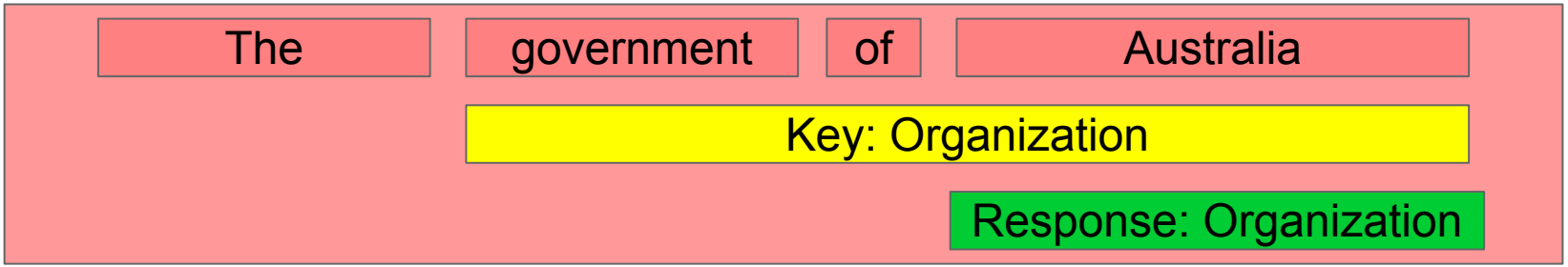
Strict and Lenient

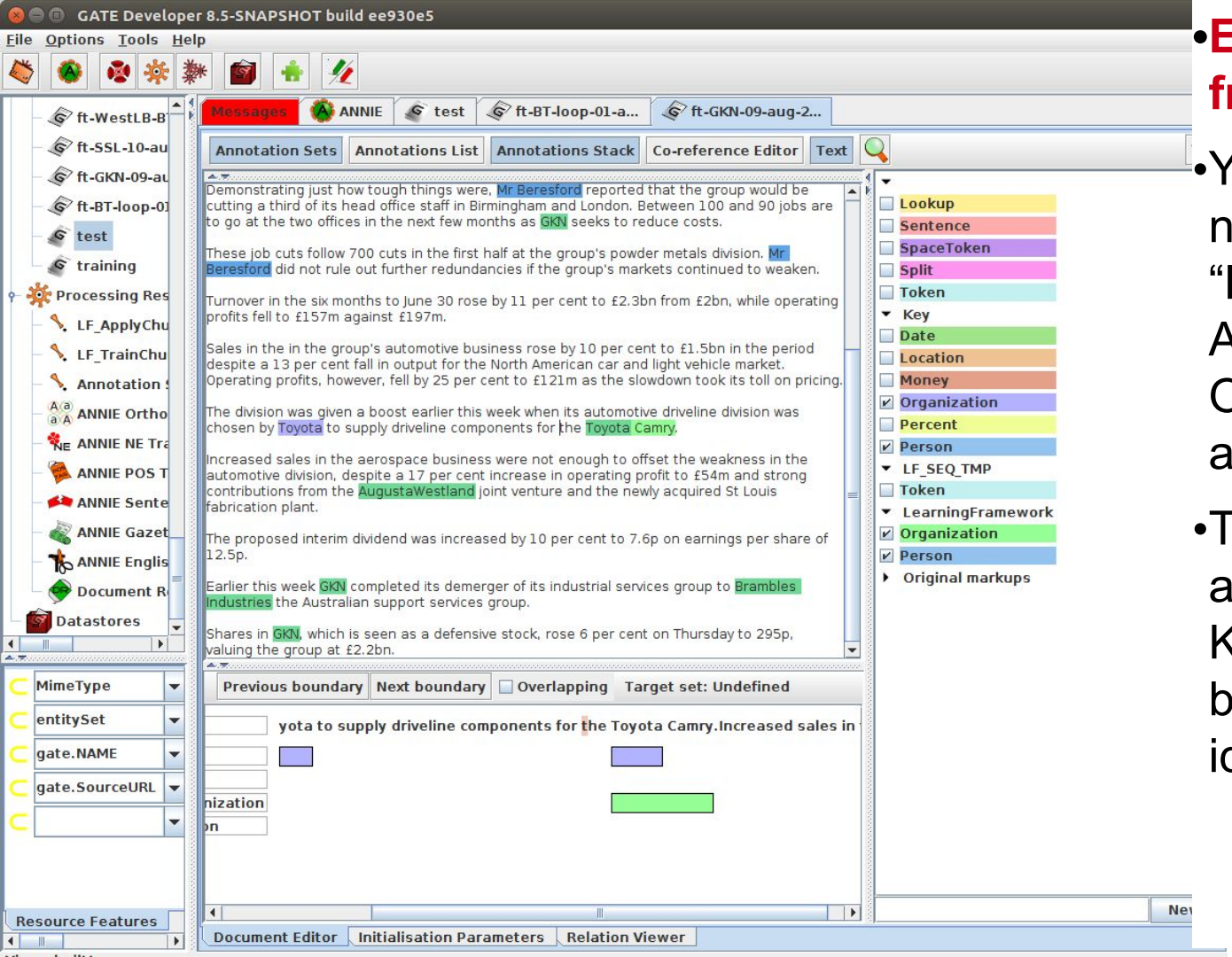
- “Strict” means we count an annotation as correct only if it has the same span as the gold standard annotation

- Lenient means we allow an annotation that overlaps to be correct, even if it isn't a perfect span match

Strict and Lenient



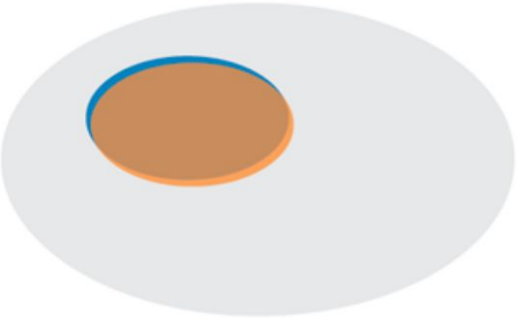


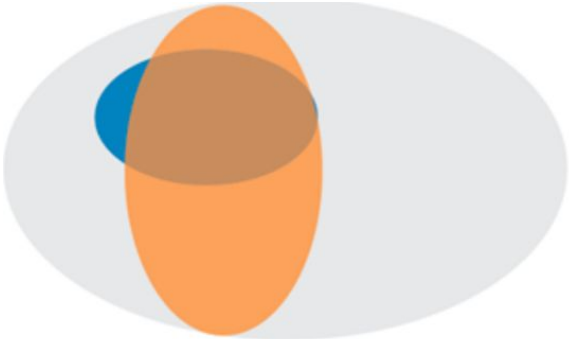


- **Examine a document from the test corpus**
- You should have a new "LearningFramework" AS with Person and Organization annotations
- The original annotations (in the Key AS) are similar but not always identical!

Precision and recall

- Precision: what proportion of our automatic annotations were correct?
- Recall: what proportion of the correct annotations did our automatic tool create?





F-measure

- 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
  - $F1 = 2PR / (P + R)$

GATE Developer 8.5-SNAPSHOT build ee930e5

File Options Tools Help

Messages | ANNIE | test | ft-BT-loop-01-a... | ft-GKN-09-aug-2...

Corpus statistics | Document statistics

Annotation	Match	Only A	Only B	Overlap	Prec.B/A	Rec.B/A	F1.0-a.
Organization	523	147	108	43	0.8079	0.7637	0.7851
Person	149	40	31	4	0.8207	0.7824	0.8011
Macro summary					0.8143	0.7731	0.7931
Micro summary	672	187	139	47	0.8106	0.7677	0.7885

- Select the test corpus and click on the Corpus Quality Assurance tab (it will take a few seconds to scan the documents)
- Select the Key and LearningFramework annotation sets as A and B, respectively
- Select the “Person” type
- Choose an F-measure
- Click on Compare

Annotation Sets A/Key & B/Response

[Default set]
Key (A)
LearningFramework (B)
LF_SEQ_TMP

present in every document

Annotation Types

Date
Location
Money
Organization

present in every selected set

Annotation Features

1
gender
kind
LF_confidence

present in every selected type

Measures | Options

F-Score | Classification

F1.0-score strict
F1.0-score lenient
F1.0-score average
F1.0-score strict BDM

Annotation Difference

Key doc: ft-claims-direct-10-a... Key set: Key Type: Person Weight

Resp. doc: ft-claims-direct-10-a... Resp. set: LearningFra... Features: all some none 1.0 Compare

Start	End	Key	Features	=?	Start	End	Response	Featur
1549	1557	Mr·Poole	{rule=PersonFinal, g...1=PersonTitleGender}	=	1549	1557	Mr·Poole	{LF_confidence=0.857
1534	1544	Mr·Sullman	{rule=PersonFinal, g...1=PersonTitleGender}	=	1534	1544	Mr·Sullman	{LF_confidence=0.804
1201	1211	Mr·Sullman	{rule=PersonFinal, g...1=PersonTitleGender}	=	1201	1211	Mr·Sullman	{LF_confidence=0.850
1188	1196	Mr·Poole	{rule=PersonFinal, g...1=PersonTitleGender}	=	1188	1196	Mr·Poole	{LF_confidence=0.848
916	924	Mr·Poole	{rule=PersonFinal, g...1=PersonTitleGender}	=	916	924	Mr·Poole	{LF_confidence=0.848
901	911	Mr·Sullman	{rule=PersonFinal, g...1=PersonTitleGender}	=	901	911	Mr·Sullman	{LF_confidence=0.842
710	721	Colin·Poole	{rule=PersonFinal, g...e, rule1=PersonFull}	=	710	721	Colin·Poole	{LF_confidence=0.545
809	824	Simon·Ware-Lane	{}	-?				
693	705	Tony·Sullman	{rule=PersonFinal, g...e, rule1=PersonFull}	-?				
1822	1829	Sullman	{}	-?				
1834	1839	Poole	{}	-?				
				?-	2569	2582	Claims·Direct	{LF_confidence=0.587
				?-	2073	2083	High·Court	{LF_confidence=0.536
				?-	2173	2186	Claims·Direct	{LF_confidence=0.476
				?-	602	615	Claims·Direct	{LF_confidence=0.628
				?-	0	13	Claims·Direct	{LF_confidence=0.677

			Recall	Precision	F-measure
Correct:	7				
Partially correct:	0	Strict:	0.64	0.58	0.61
Missing:	4	Lenient:	0.64	0.58	0.61
False positives:	5	Average:	0.64	0.58	0.61

Statistics Adjudication

16 pairings have been found (0 annotations are hidden)

Switch to the “Document statistics” tab

Choose a document

Click on the Annotation Diff icon

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”)

Deep Learning

- Gate support deep neural network
  - Require install python deep learning libraries
- Supported neural network architecture
  - CNN
  - RNN/LSTM
  - Pre-Trained word embedding
  - ELMO
  - BERT (in progress)

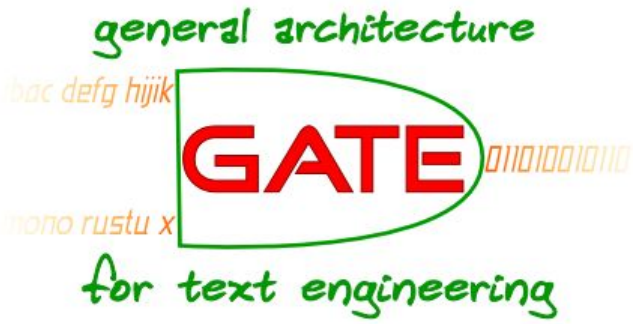
What different

- Still in development
  - Beta version available

- No different algorithms
  - Different architectures
  - Different loss functions, optimizers
  - Regularization, attention, CRF layer, GANs

Dummy Model

- Change trainingAlgorithm
  - PytorchWrapper_SEQ_DR
    - Using a simpe LSTM for sequence labelling
- If you are using pytorch
  - Customize your model
  - data_dir/FileJsonPyTorch/gate-lf-pytorch-json/gatelf pytorchjson/modules/
- We will support more options in future



Questions?