THE SHORT GATE – GOLD STANDARD TUTORIAL

What you need:

- Some documents extracted from the web
- Ontology
- Corpus – basically just a collection of documents (text, xml, etc.)
- Datastore that contains the corpus

Documents

Text documents extracted from the web, subtitles, books or any other types of files that can be exported to XML should do just fine.

Load plugins

Load the necessary plugins: 3 ontology plugins that are needed in order to work with the ontology.

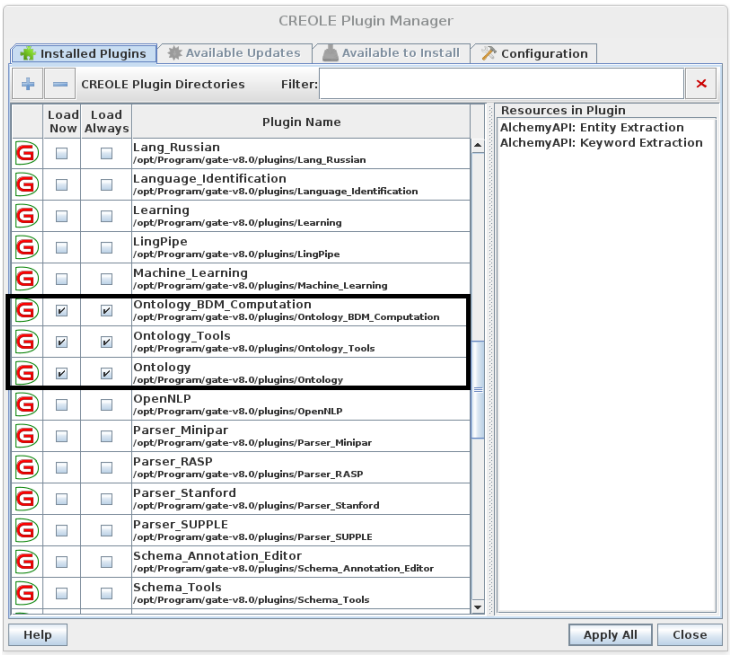

Figure 1. Ontology plugins.

Load an ontology

The current ontology (Recognyze) is really simple and uses only uses the following types of entities:

- Person (anything labelled as Person, including fictional characters)
- Organization (Companies, NPOs, NGOs, Governments, Sport Clubs, etc.)
- Location (anything that has something to do with Geopolitics)
- Product (Products)
- Event (only important events go here, otherwise everything that happens is an event)
- Work (usually works of art)
- Misc (in general Things that are not included in the previous categories, but also television shows or titles, etc.)

The only attribute you are interested to complete at each entity is the hasURI attribute – where you will fill the DBpedia link for the particular entity.

Links should be of the form:

http://de.dbpedia.org/resource/Allianz_Arena

Alternatively you can also complete the links as you copy paste them:

http://de.dbpedia.org/page/Allianz_Arena (this contains the page redirect instead of resource).

Please stick to just one of them: either **resource** or **page**.

Create OWLIM Ontology

For the OWLIM ONTOLOGY loading, please stick to the following settings if you plan to use the simple Recognyze ontology:

- *Name* - recognyze
- *Base URI* - http://semanticlab.net/ontology/recognyze#

Should load an owl xml ontology.

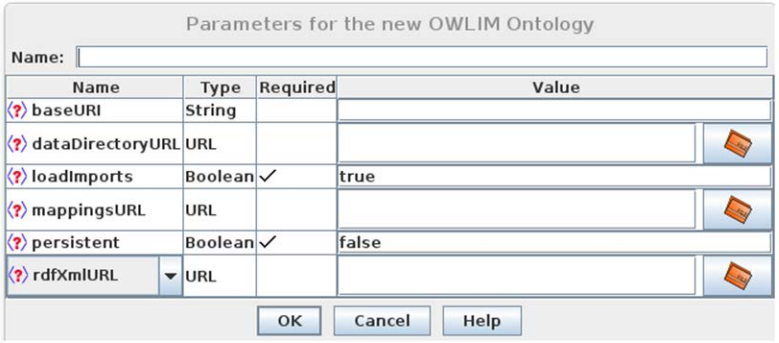

Figure 2. OWLIM Ontology parameters.

Please be careful with creating/loading ontologies as there is no easy method to save the **Ontology LR** (Ontology Language Resource in a Datastore), therefore you might need to repeat these steps each time you load an ontology in GATE.

Load OWLIM Ontology

Loading existing ontology (still needs Step 1 to be done before it, as far as I noticed)

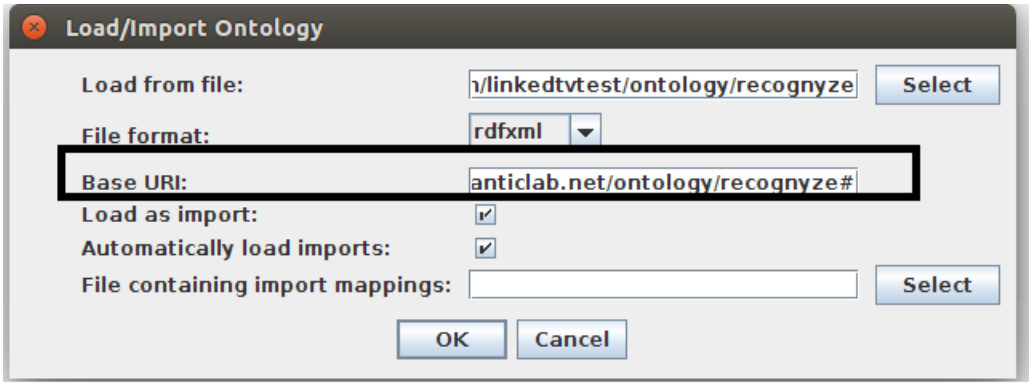

Figure 3. Load/Import Ontology from external file.

If you want to use an existing ontology (which you do!), use the **Load** option from the context menu.

You click on the created ontology and click **Load**!

In addition to loading the ontology from the file, you need to type the *Base URI* again.

Check if the Ontology was loaded OK

If your ontology was not loaded correctly, you will not see any annotation, therefore until you get used to it (or every time you don't see your annotations) it is good to check if the ontology was loaded ok. All you need to do is to double-click on its icon and it will be opened in a separate tab, as presented in the next figure.

The best clues regarding the correct loading is the fact that you can see and edit the ontology. It is also recommended to use this screen to familiarize yourself with the ontology, as you will have to use not just the main classes and subclasses of the ontology, but also some of the property types (especially *hasURI* and *comment*).

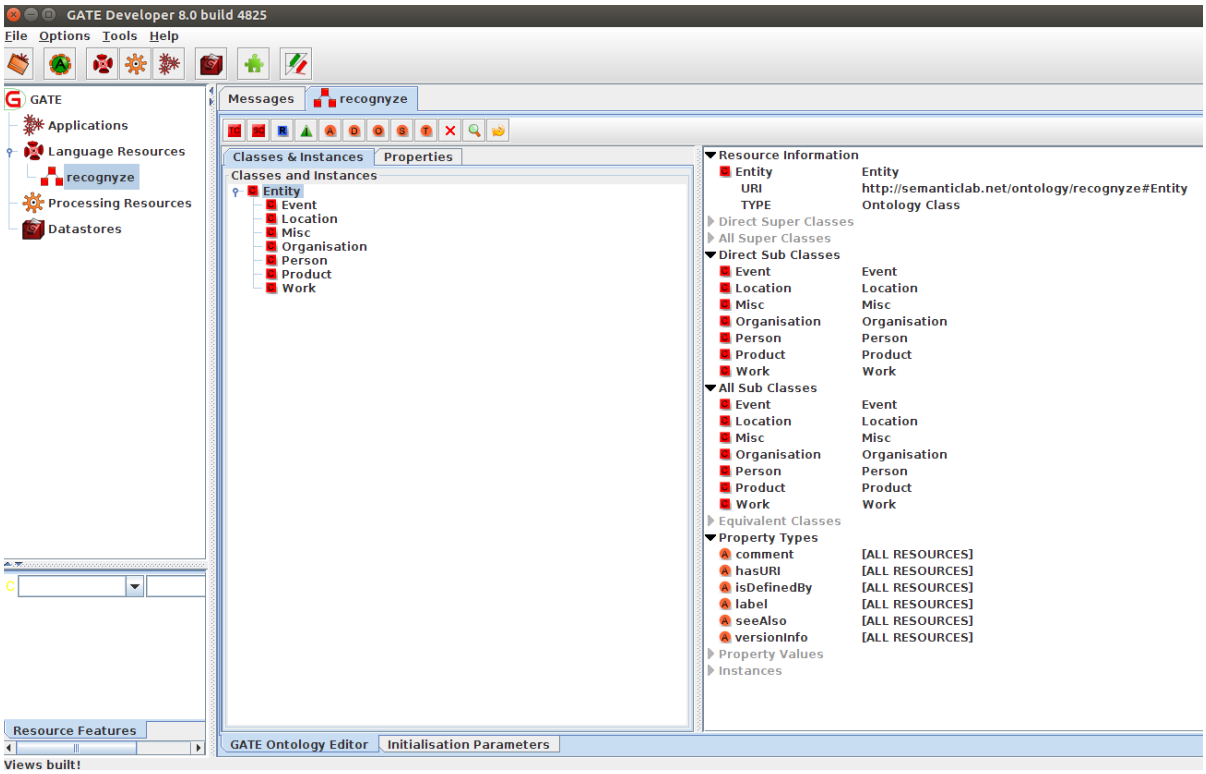

Figure 4. Loaded ontology. Classes and Instances can be seen on the screen.

Before getting into more details about creating annotations and gold standards (or ground truths) it is good to discuss a bit about datastores, since they are useful when you want process/save your work.

Another clue that you are ready to annotate is the following screen that you will see if you:

- Open a new document
- Open existing document from a corpus.

You need to do the following steps:

- In the **OAT** tab select recognyze as an ontology.
- Save your current configuration from the context menu.

In order to create gold standards it is however useful to know some things about several other concepts, like datastores, annotation sets, agreement measurements, etc.

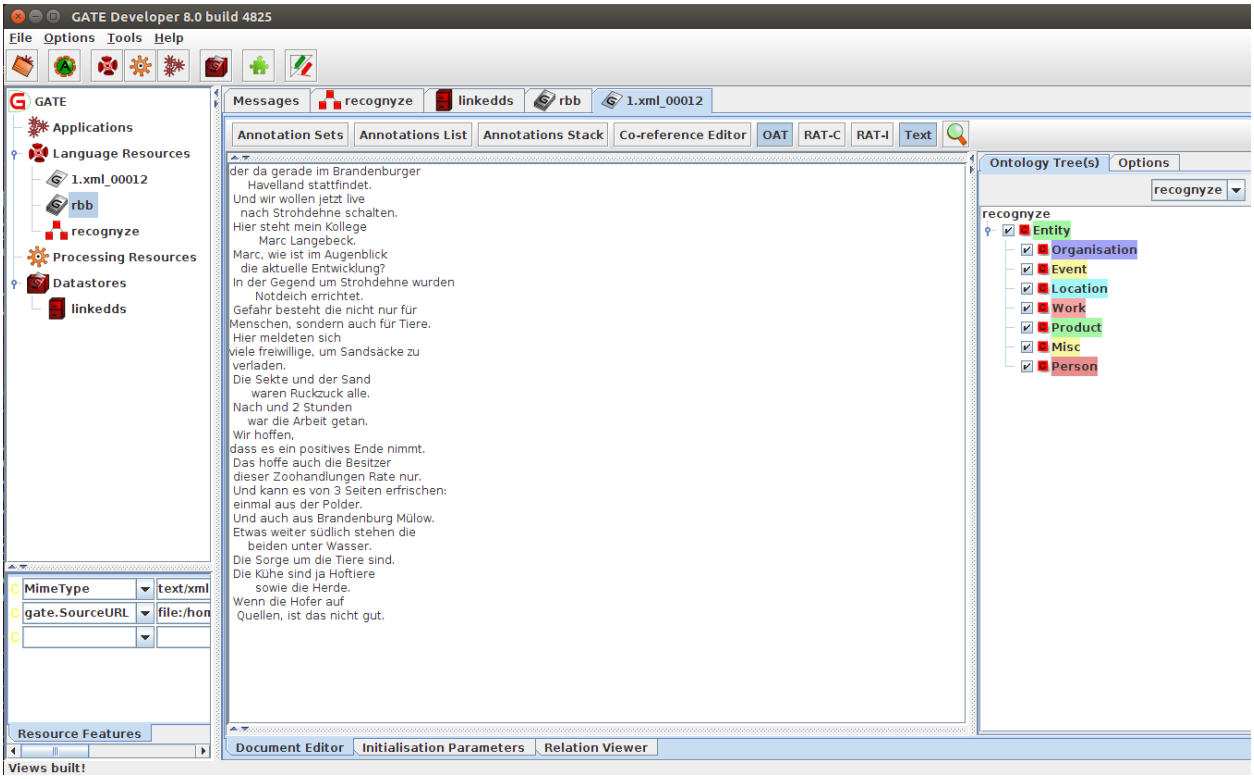

Figure 5. A loaded document and the **OAT** tab. The **Ontology Tree** should display a simple view of the ontology.

Creating and Loading datastores

To create a datastore it is enough to simply push *Create datastore* from the context menu that appears when you click Datastores.

For loading a datastore you need to go the same menu: *Datastores –> Load Datastores*.

You will see two types of resources in any new datastore:

- GATE Documents
- GATE Serial Corpus.

You need to use the GATE Serial Corpus option and click *Load* again.

If you use the Lucene-based option for creating datastores you will also be able to use your datastore for IR tasks (Information Retrieval tasks).

If you click on a corpus or document, the respective corpus or document will also be loaded in the Language Resource (LR) pane.

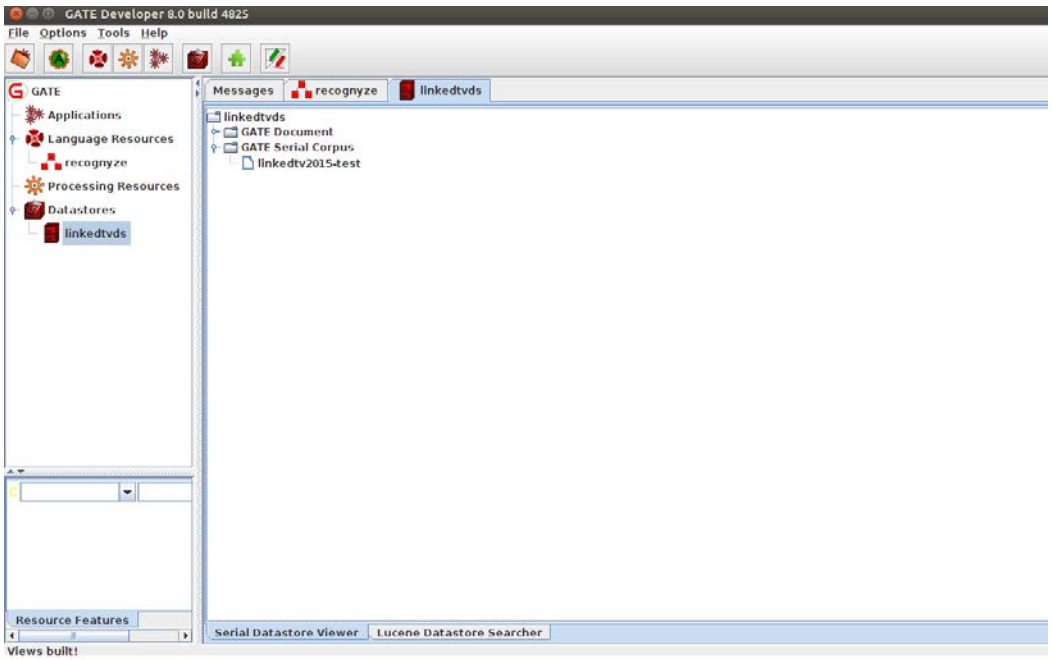

Figure 6. Loading a datastore. It should contain GATE Documents and a GATE Serial Corpus.

Creating annotations

Go to a specific document in the corpus.

You can start with the first one.

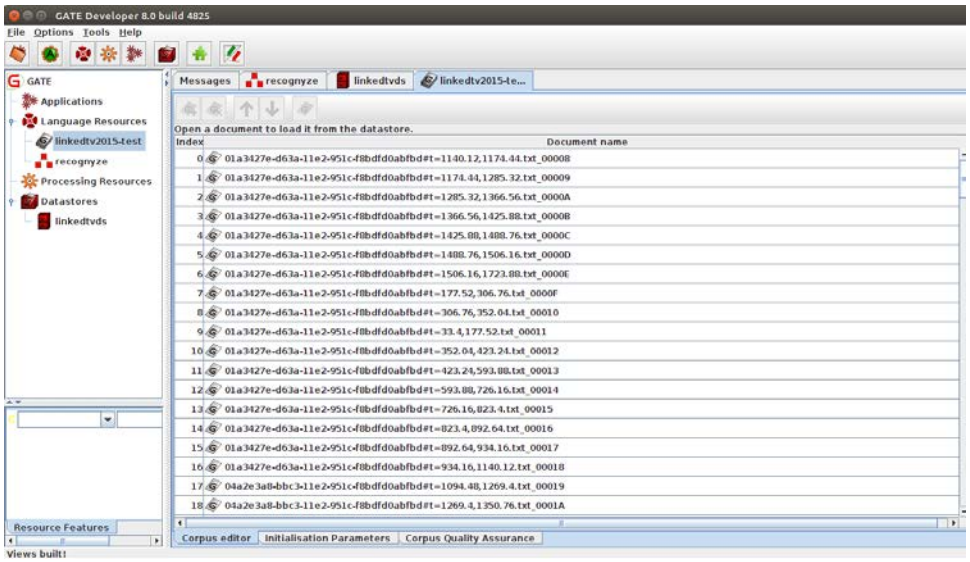

Figure 7. Loaded Corpus.

When you are in the respective dialogue you have to activate the OAT view so that you can access the Ontology Tree for the annotations.

To annotate you have to do 2 steps:

- First you set up the Entity type (Organisation, Person, Location, etc.)
- Then you need to add its DBpedia URI in the *hasURI* field as shown in the picture.

Both of the dialogues that you need can be found in the next picture:

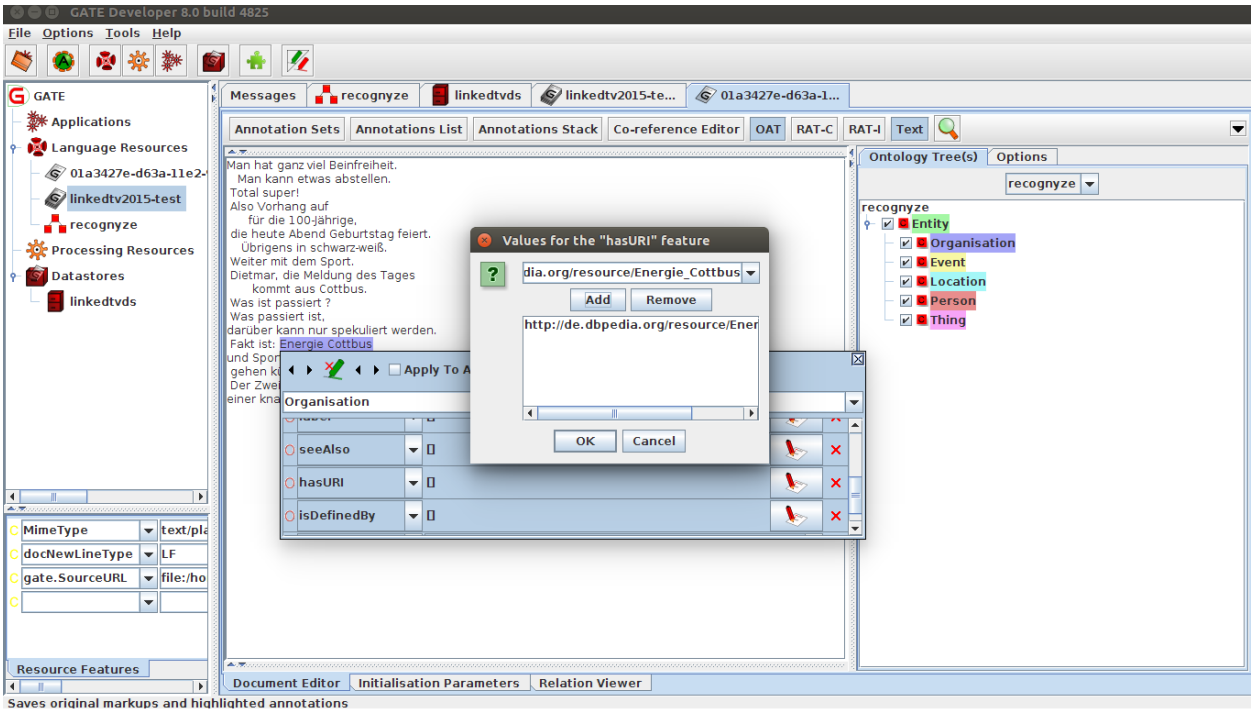

Figure 8. Adding annotations. Apply to All is disable in this picture.

WARNING! *If you select the **Apply to All** checkbox in the dialog that is displayed for annotating an entity, the respective entity will be automatically spotted in all the places where it appears in the respective text.*

Saving your work

You have several options:

- Save as XML
- Save to Datastore
- Save to its Datastore
- Save Preserving Document

We recommend **Save as XML** and **Save to Datastore** options in general, as Save Preserving Document usually reverts you to the original document (though with some minor paragraph formatting options since it is an XML document) and you risk the loss of your annotations.

A good strategy is to the first 3 saving options like this:

- **Save as XML** in a separate folder than the initial folder where you kept your texts/XML documents. This way, in case you corrupt your datastore somehow, you will still have the annotated document (very important if you have tight deadlines).
- **Save to Datastore / Save to its datastore** – so that you can load it back from the datastore.

Usually the fact that you do not use the right saving technique is the main cause for multiple errors in GATE.

For example you keep using just *Save as XML*, and you forget that this will not save it to the datastore!

If you have enough time, it is good to keep detailed notes about the entities you are in doubt, as additional text/csv files where you just recorded the name of the entity (its surface Form), and the basic problem you encountered (especially if you are not sure if the respective entity needs to be marked as a particular type of entity: perhaps you can't decide if The Syrian War is an Event or a Miscellaneous entity and you don't add it at all).

Keeping good records ensures that you create high-quality corpurses. It will also help you when you need to describe the corpus and detail the problems you encountered or the methodology you have used.

WARNING! *Save your work as often as you can!*

Even if you have 16-32 GB of RAM, GATE often crashes, therefore saving often is a must!

Annotation Sets

When multiple persons need to annotate the same documents, it is often a best practice to use Annotation Sets. In general you should create the following annotation sets:

- Ann-1 (or Annotator-1) – for the annotations saved by the first annotator.
- Ann-2 (or Annotator-2) – for the annotations saved by the second annotator.
- Ann-n (or Annotator-n) – for annotator n, obviously.
- Consensus – For the agreed annotations for the respective document. It should be created using the **Annotation Diff** dialog by a person who has some experience with these matters.

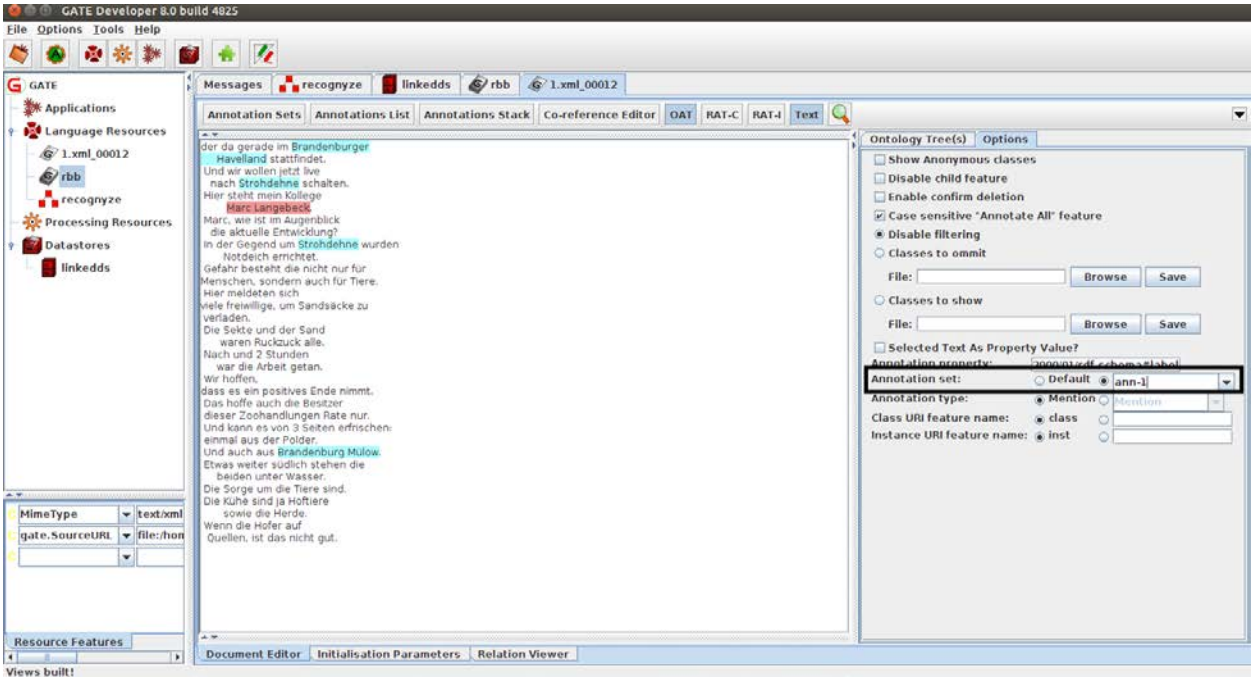

Figure 9. Creating Annotation Sets.

WARNING! *You can only connect one annotation set to the ontology, therefore you will only be able to see the annotations for a particular annotator on the screen. The default set however belongs to no annotator, therefore don't panic if when you load a document you thought you annotated already you see nothing on the screen!*

The **Options** panel from the **OAT** tab, allows you to:

- Select the annotation sets you are interested.
- Save the selected text as a property value.
- Mark the annotation type you are interested in. Since we are mainly interested in entities these annotation types will usually be of type Mention.

Agreement

In order to compare the annotations from various users you can use the Annotation Diff dialog (Tools -> Annotation Diff).

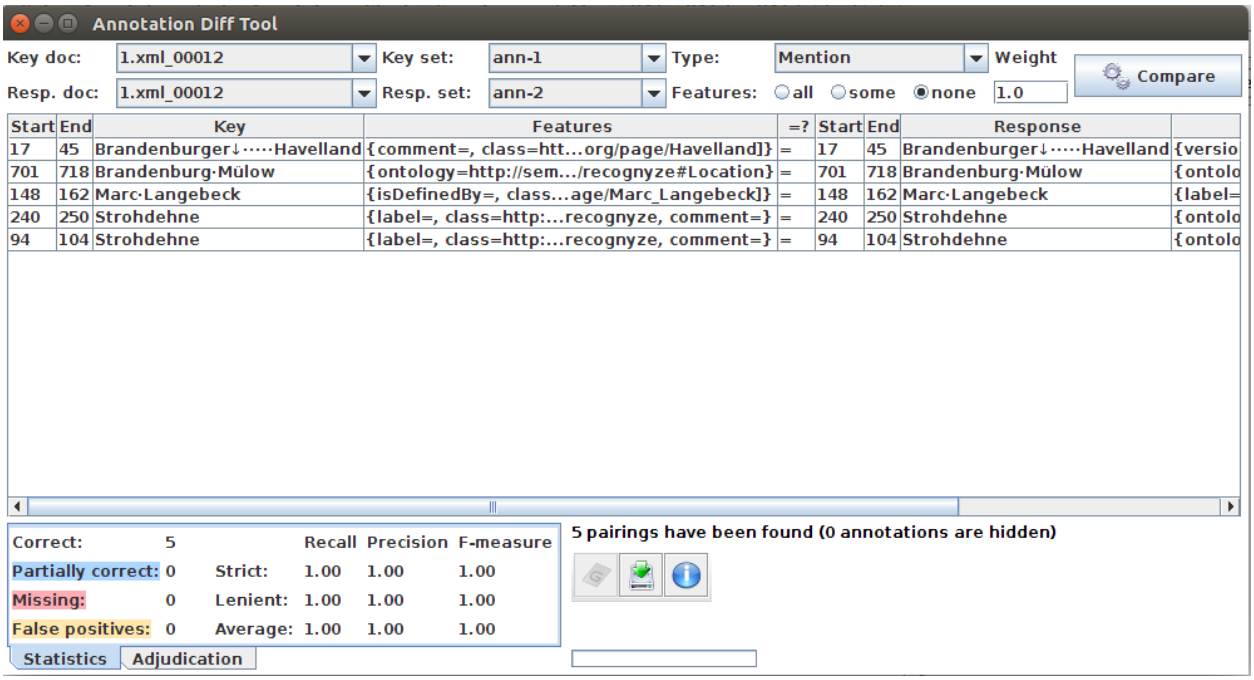

Figure 10. Annotation Diff Tool.

The **Annotation Diff** dialog has two panels:

- **Statistics** – Allows you to see the various measures that might be interesting in order to assess the quality of your annotations.
- **Adjudication** – Helps you copy annotations to the consensus annotation set.

As it can be seen in the picture, entities that were marked with Apply to All (*Strohdehne* in this case) appear multiple times with different POS tags (Start, End).

In order to create the consensus annotation set, all you need to do is to check the boxes for the entities that should go there.

All the entities that were agreed disappear from the dialog, leaving you only with those problematic entities that need to be fixed/agreed by a third party (namely the expert annotator).

Corpus Quality Assurance

Various quality measures need to be calculated in order to assess the quality of the whole corpus (Inter rater agreement, confusion matrices, etc.).

All these measures can be found in the Corpus Quality Assurance dialog, which has two tabs.

The **Document statistics** tab offers basic statistics like Observed Agreement, Cohen's Kappa (for Inter Rater Agreement), Pi's Kappa, etc.

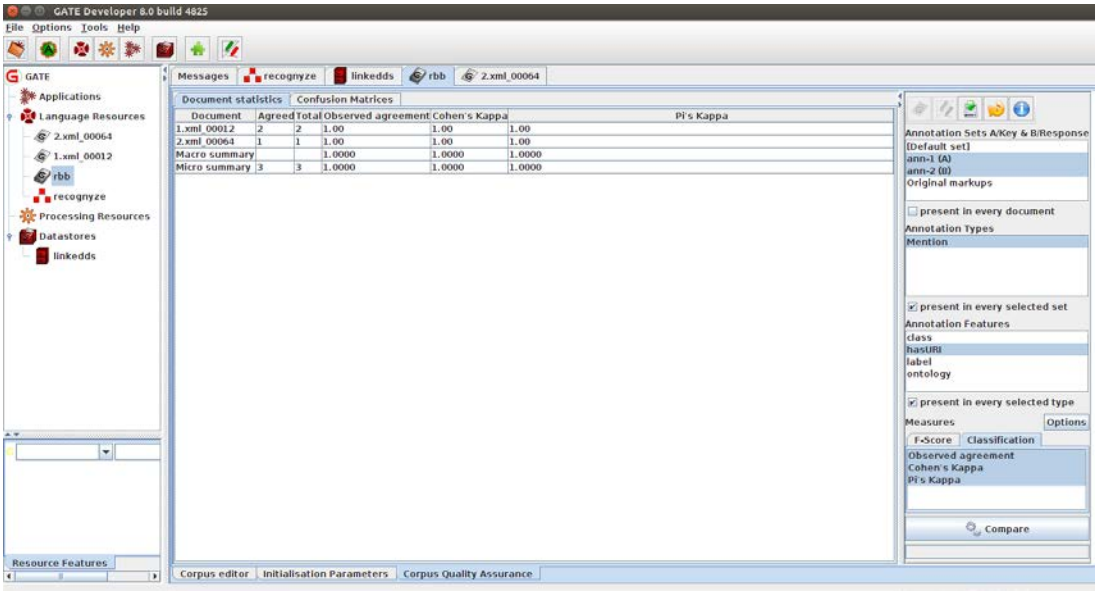

Figure 11. Document Statistics tab.

The **Confusion Matrices** tab shows the matched entities according to an annotation features (hasURI is typically the one we are interested because we want to see if the annotators have identified the same entity – geographic entities often refer to cities with the same name, therefore only the link tells us if two annotators have thought about the same entity).

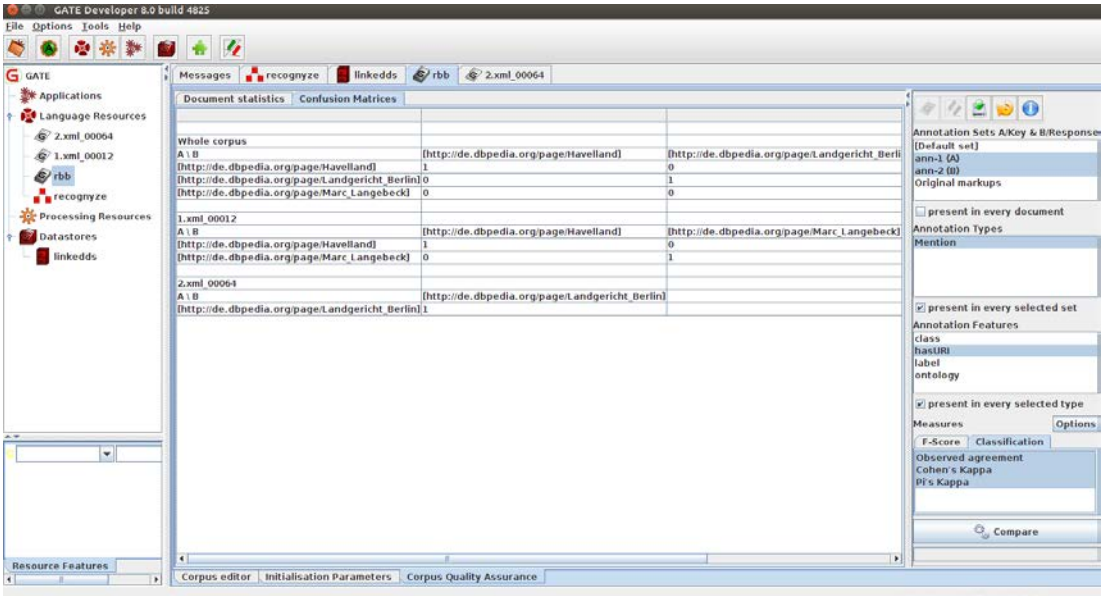

Figure 12. Confusion Matrices.

Processing the output

If you want to use the gold standard to create evaluations for Named Entity Resolution tools, what you need to do is parse the consensus annotation sets from the generated XML files.

Since these are just XML files you can parse them with a variety of programming languages and tools.

Be careful though, as often the XML files offer you only the POS tags (Start, End), and not the Surface Form (the actual string that corresponds to the entity you spotted in the text). You can of course, add it through a script/program written in Java or Python.

Bibliography

https://gate.ac.uk/gate/doc/