Annex I. BioTML tutorial using @Note2 platform for NER and RE tasks.

The tutorial aims to show all steps necessary to use the BioTML plug-in integrated in the @Note2 platform. Two main sections are presented in this annex, a section dedicated for a NER task and another for a RE task.

BioTML usage for a NER task: BioCreative V CHEMDNER-patents

The training and development sets of CEMP corpus can be download at `http://www.biocreative.org/media/store/files/2015/cemp_training_set.tar.gz` and `http://www.biocreative.org/media/store/files/2015/cemp_development_set_v03.tar.gz` provided by BioCreative Team.

Unpack the files and each folder contains four files shown in Figure 1.

The document files (chemdner_patents_train_text.txt) and the annotations file (chemdner_cemp_gold_standard_train.tsv) can be loaded into the @Note2 platform by the CHEMDNER Corpus Loader.

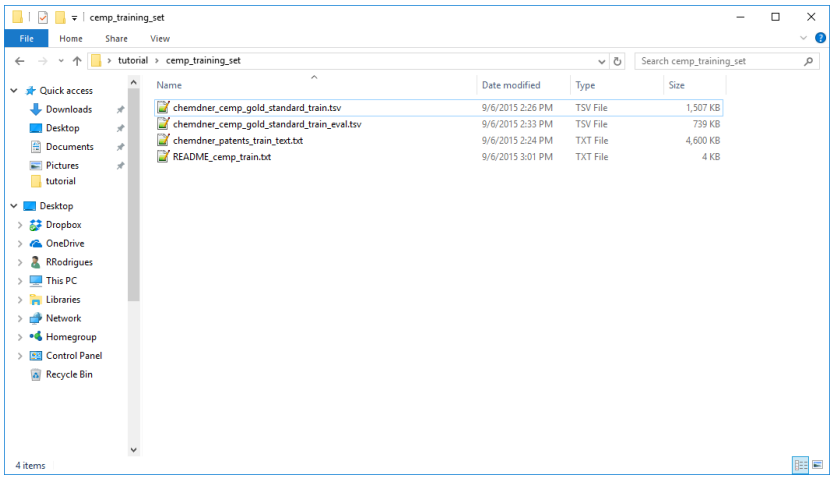

Figure 1: Unpacked files from training set CEMP Corpus tag.gz file (a README txt file, a txt file that contains sentences from patents, a tsv file with curated chemical annotations and a tsv file for evaluation of further predictions).

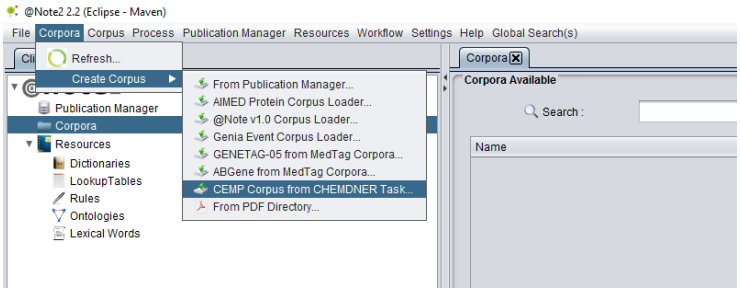

Figure 2: @Note2 drop down menu with CHEMDNER Corpus Loader selected.

To open the loader, access the Corpora on @Note2 dropdown menu, click on Create Corpus and select CEMP Corpus from CHEMDNER Task as Figure 2 shows.

A GUI (Figure 3) appears, allowing to select the document and annotations files to be loaded into @Note2 platform.

A corpus loader report GUI appears, after the task finishes, showing the number of documents and processes loaded into the @Note2 platform (Figure 4).

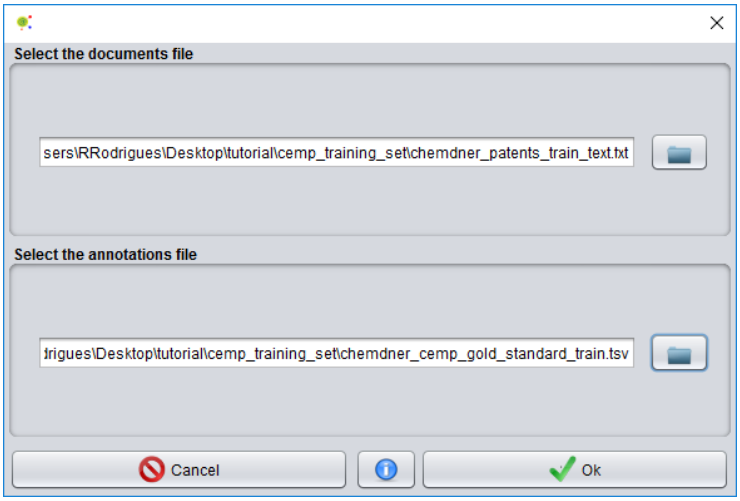

Figure 3: CHEMDNER Corpus Loader GUI with selected files to be loaded on @Note2 platform.

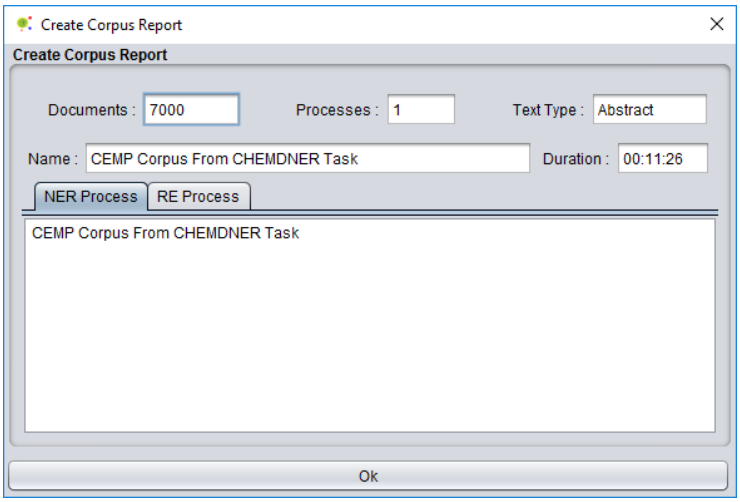

Figure 4: CHEMDNER Corpus Loader report GUI with selected files to be loaded on the @Note2 platform.

Perform the loading of the training and development sets to the @Note2 platform, as shown below.

Note: After the corpus loading, the default corpus name contains the name of the loader and the time at which it was executed. In this case, we renamed each corpus to "Training Set CEMP Corpus from CHEMDNER Task" and to "Development Set CEMP Corpus from CHEMDNER Task" as shown in Figure

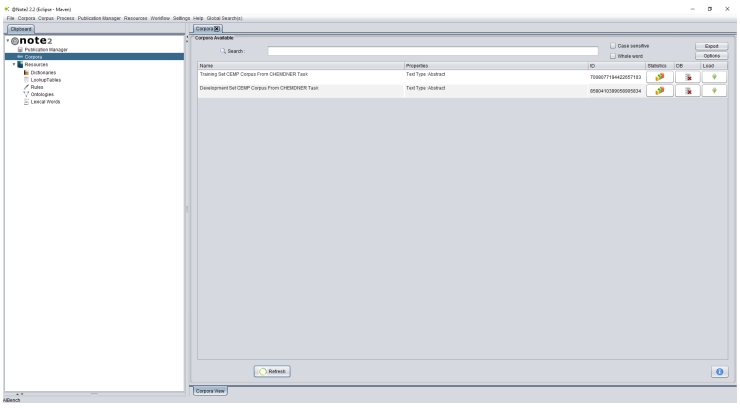

Figure 5: Training set and develpment set CEMP Corpus loaded on the @Note2 platform.

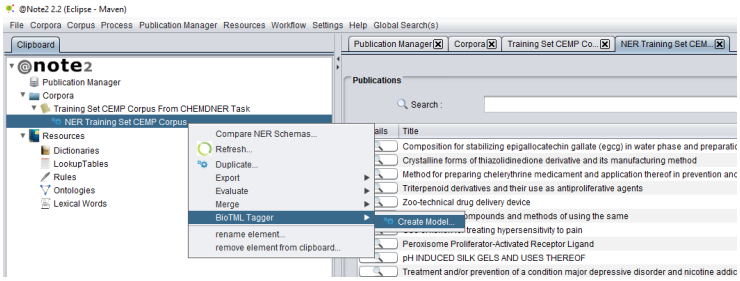

Figure 6: Activation of BioTML Tagger using the @Note2 platform.

5.

A ML model based in the training set can be created to test the BioTML framework for NER tasks. Afterwards, the model can be applied to the development set to generate new annotations that can be compared against curated annotations.

To create the ML model, select the training set process with the right mouse button and create the model. The BioTML Tagger option must be clicked with left mouse button (Figure 6).

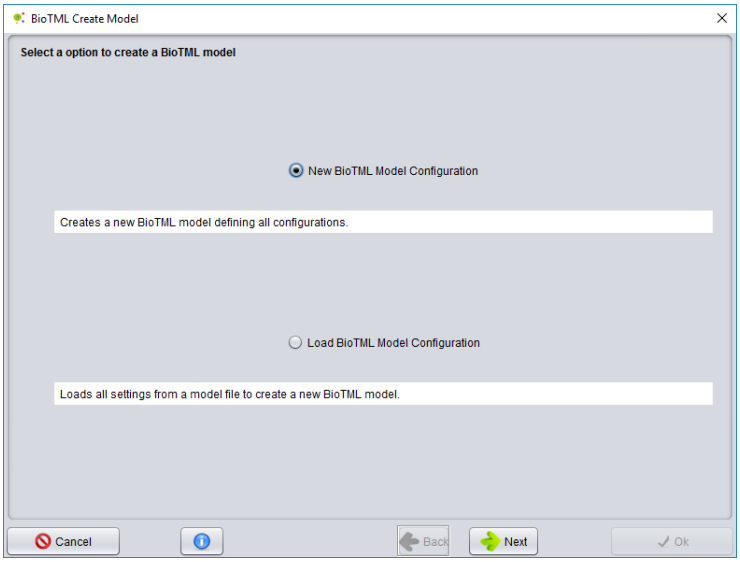

Figure 7: BioTML Tagger Wizard to create ML model using the @Note2 platform.

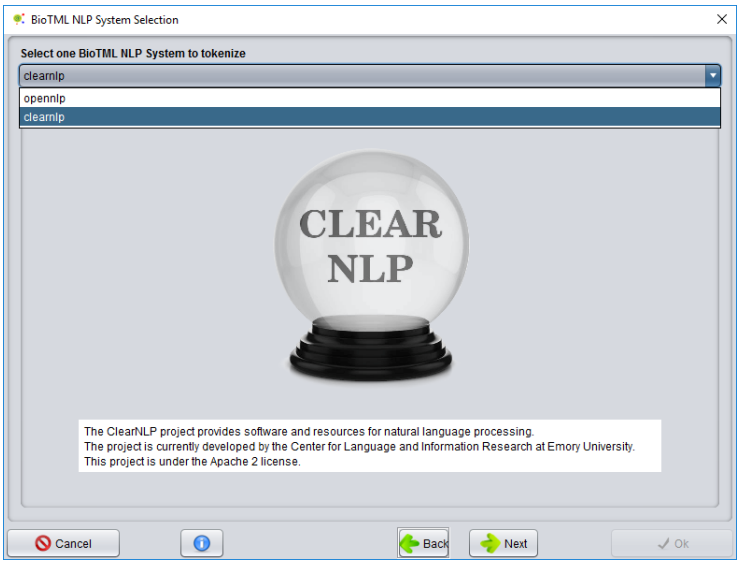

Figure 8: NLP tokenizer selection page on the BioTML configuration wizard.

A configuration wizard opens to select all configurations to create the ML model (Figure 7). In the first wizard page, we can select a new configuration or use a configuration from a model created previously. As we want to create a new model, select the option "New BioTML Model Configuration" and click on the Next button.

The second wizard step (Figure 8) allows to define which NLP system will be used to tokenize the text streams submitted to the training process. In this tutorial, we selected the *Clearnlp* system to perform all tokenization tasks and clicked in the Next button.

The features selection wizard (Figure 9) appears to enable/disable which features we want to use on the ML model training to fit the provided annotation data. Each feature contains a description that is shown by moving the mouse pointer over the feature line. By default, a set of features is selected, which in this tutorial will be used to showcase the ML training.

The ML algorithm used to train the ML model is shown in the next wizard page. We selected the CRF algorithm, first order, using 7 CPU threads (90%

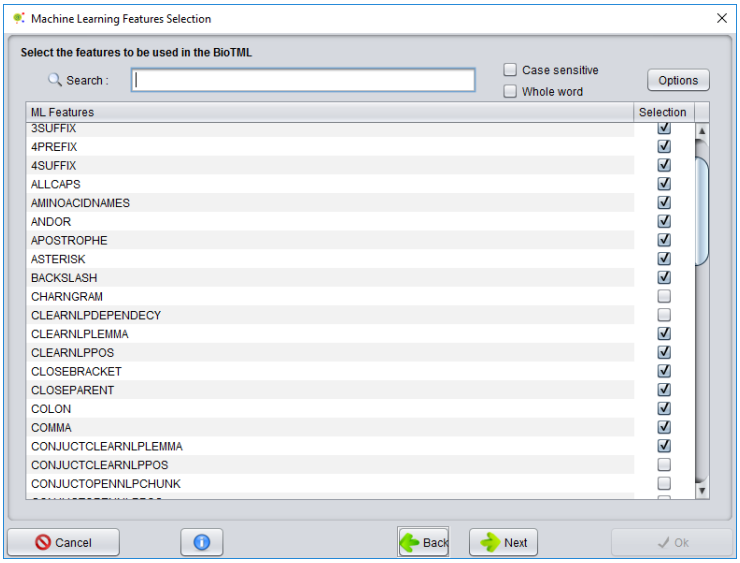

Figure 9: @Note2 features selection configuration wizard of the BioTML Tagger.

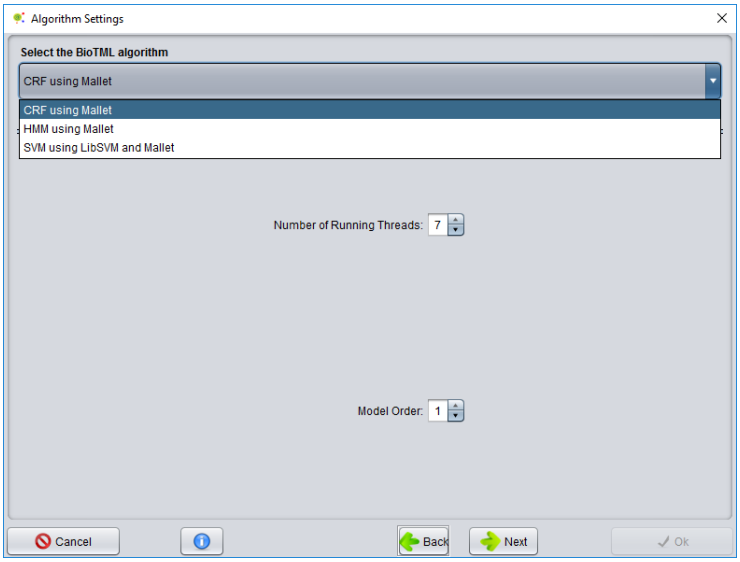

Figure 10: ML algorithm selection to be used in model training.

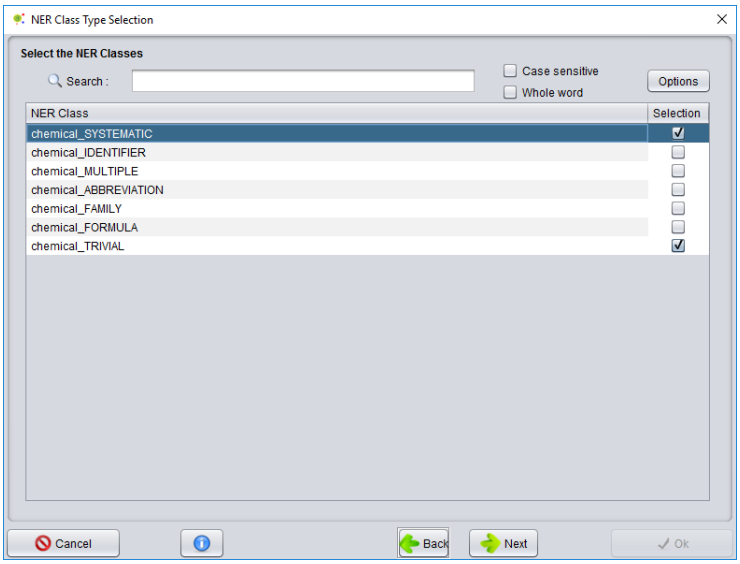

Figure 11: ML algorithm selection to be used in model training.

of our CPU capacity). Other algorithms like SVM or HMM can be selected in alternative. The configuration of all these algorithms is described in the @Note2 Wiki page.

The annotation class selection wizard is shown in the next page. The training set contains 6 different classes shown in Figure 11. We selected only the *chemical_SYSTEMATIC* and *chemical_TRIVIAL* classes to be used on the ML model. This page enables the possibility to annotate multiple classes with a single ML training configuration or use only one class for each features selection and ML algorithm configuration.

In the final wizard step 12, we can select where the ML model file will be saved. A Zip file is created when the ML model training is finished. In this tutorial, a NER_MODEL_FROM_TRAINING_SET.zip file will be created in the computer desktop. To finalize the configuration and start the training ML model process click on the Ok button. The process takes some time depending on the training set used, CPU usage, features selected and algorithm defined on configuration steps.

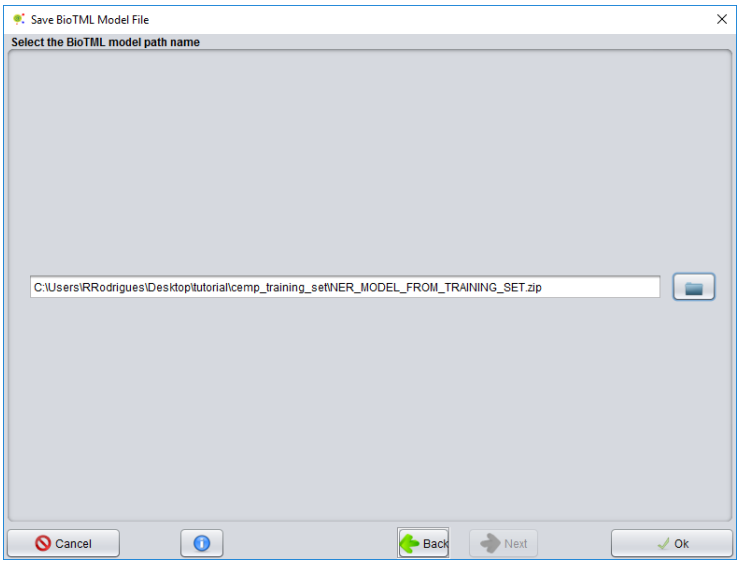

Figure 12: ML model file path to export selection wizard page.

The resulting ML model file can now be applied into any @Note2 corpus. As we want to test the ML model's annotation performance, in this tutorial, we will apply the model to the development set. Afterwards, we will compare the annotations generated by the ML model against the provided CEMP annotations with @Note2's evaluation tool.

To apply the ML model to the development set, we must load the development set on @Note2 corpora view. Next, the development set corpus will appear in the @Note2 clipboard, click with the right mouse button and select NER- BioTML Tagger - Annotate with model (Figure 13). A wizard appears

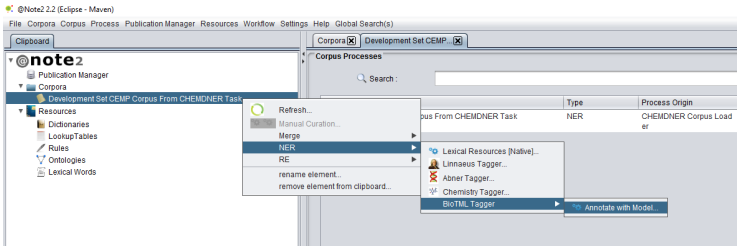

Figure 13: Selection of Corpus to annotate with BioTML Tagger on @Note2 clipboard.

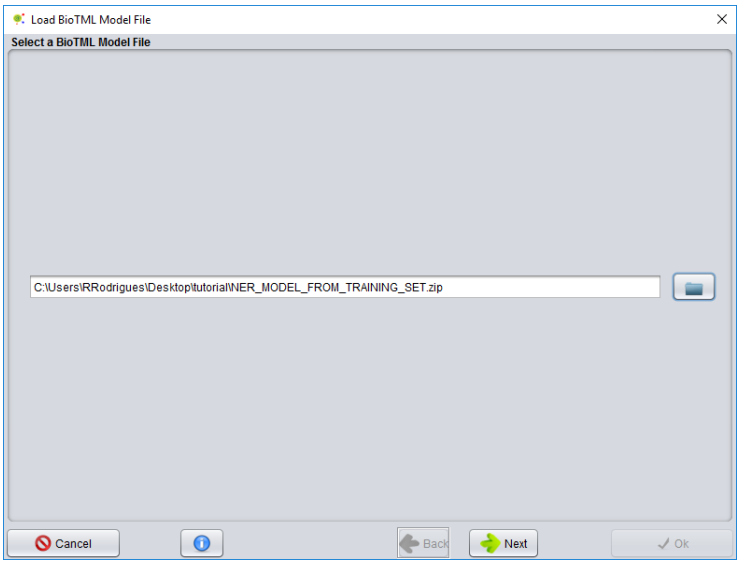

Figure 14: ML model file path to import selection wizard page.

to select a ML model to be loaded into the @Note2 platform (Figure 14).

In the next page, the wizard shows (Figure 15) the possible classes that the

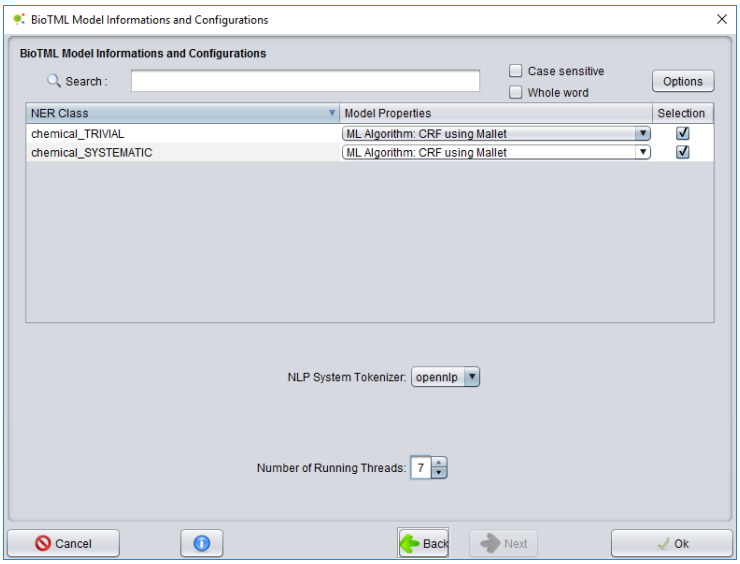

Figure 15: Wizard informations about ML model on @Note2 platform

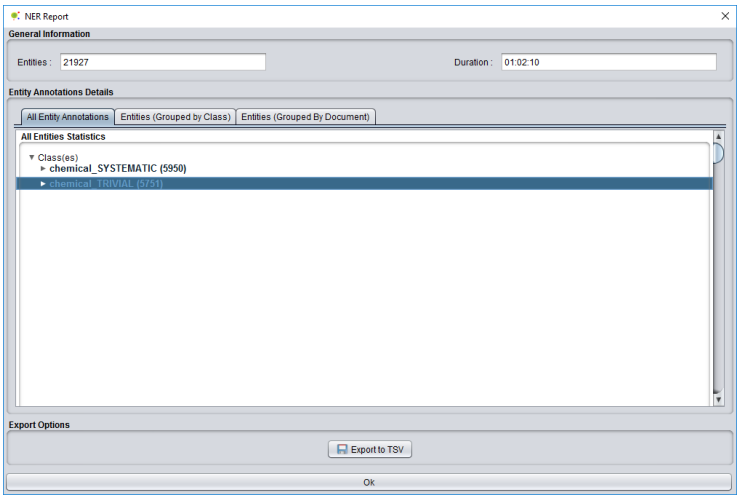

Figure 16: Report of BioTML tagger annotation process finished.

ML model is able to annotate (on our tutorial *chemical_SYSTEMATIC* and *chemical_TRIVIAL* classes are the only ones able to be annotated). In the same page, we can select which NLP tokenizer and number of CPU threads to be used in the annotation process. To start the annotation process, click on the Ok button. The process takes a time to annotate the corpus, and in the end a report is shown to the user (Figure 16).

We can evaluate the annotations generated by the ML using the evaluation tool. To perform a comparison of the development set gold annotations ver-

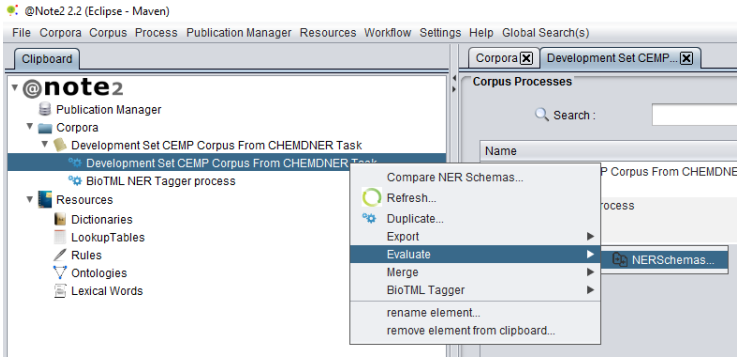

Figure 17: Process selection to execute a NER evaluation.

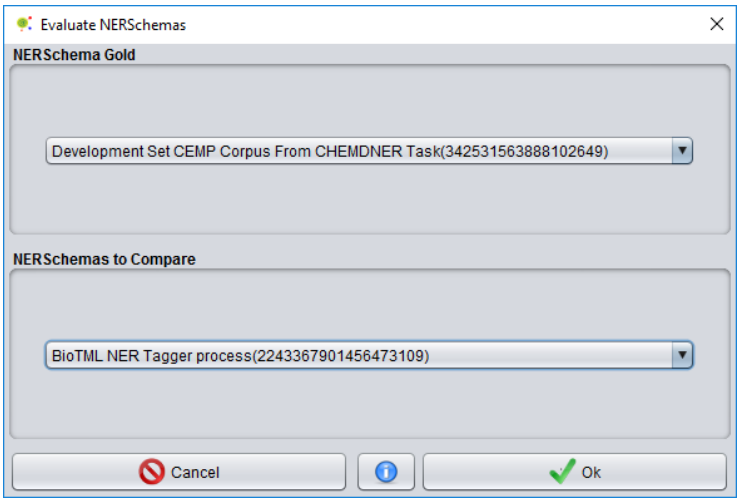

Figure 18: @Note2 evaluation GUI view.

sus generated ML annotations, load the new process and the development set process into @Note2 clipboard. Afterwards, click on right mouse button on the development set process and select evaluate - NER schemas as shown in Figure 17.

An evaluation GUI appears to select the NER processes to be used in the

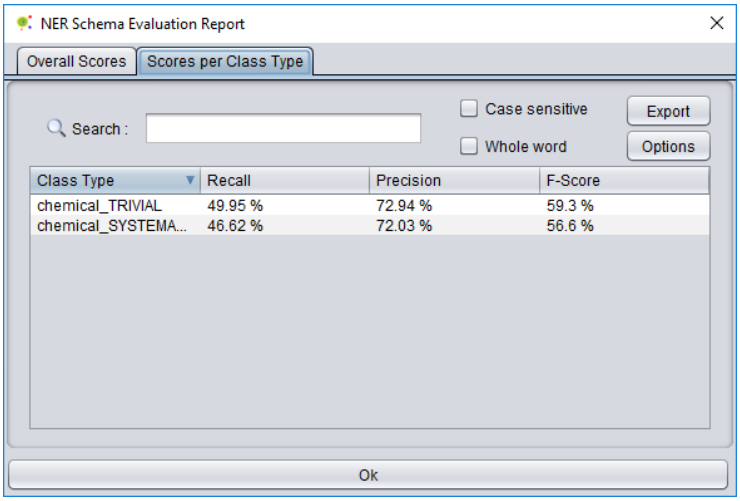

Figure 19: Evaluation NER schemas scores result GUI.

evaluation. In our tutorial, the gold standard is the process from development set and the BioTML tagger process is the process to compare to (Figure 18). To start the evaluation click on the Ok button.

As result of the evaluation process, a report is shown (Figure 19). The mean of all scores are presented in the overall scores tab and the score of each annotation class is presented in the scores per class type tab.

BioTML for a RE task: BioNLP-ST 2011 Genia task

The training and development sets of BioNLP-ST 2011 corpus can be downloaded at `http://weaver.nlplab.org/~bionlp-st/BioNLP-ST/downloads/files/BioNLP-ST_2011_genia_train_data_rev1.tar.gz` and `http://weaver.nlplab.org/~bionlp-st/BioNLP-ST/downloads/files/BioNLP-ST_2011_genia_devel_data_rev1.tar.gz` provided by BioNLP team.

Unpack the files and each folder contains three types of files (.txt, .a1 and .a2).

Open the @Note2 platform and select the BioNLP Corpus Loader in the menu Corpora - create corpus - BioNLP Corpus from A1 A2 format (Figure 20).

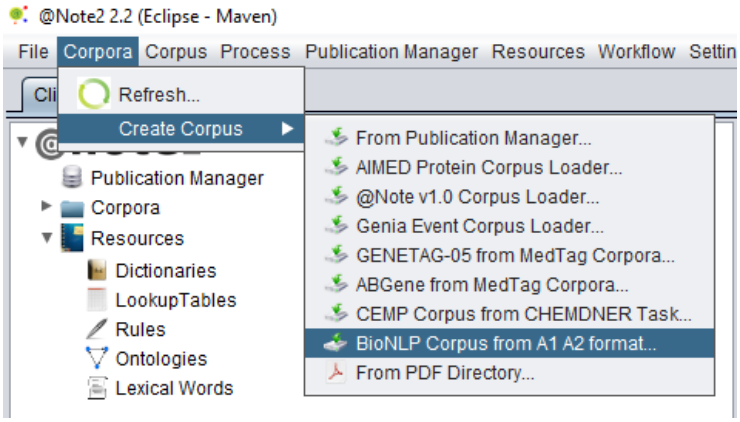

Figure 20: BioNLP Corpus loader on @Note2 platform.

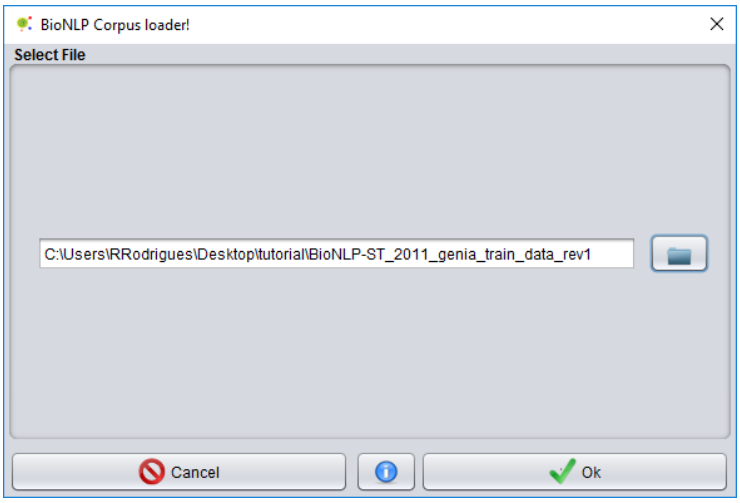

Figure 21: BioNLP Corpus loader directory selection GUI.

A GUI (Figure 21) appears to select the directory of unpacked BioNLP corpus files to be loaded into the @Note2 platform. The process takes few minutes to insert all documents and annotations in the platform.

To execute the creation of a BioTML model for RE tasks, select the RE process from BioTML corpus in the training set with the right mouse button click and choose BioTML Tagger - create model (Figure 22).

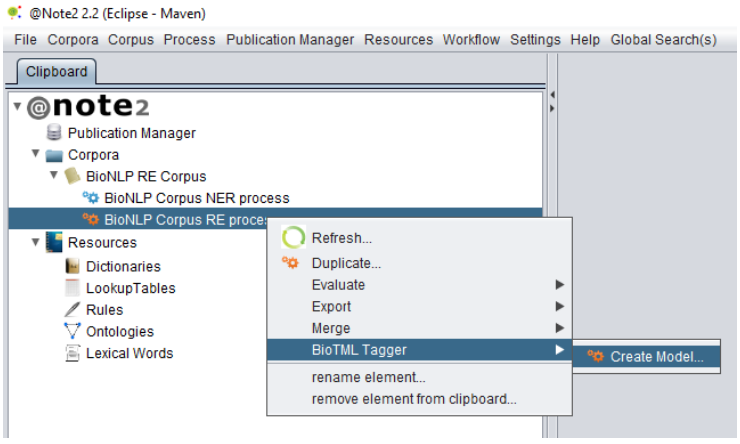

Figure 22: BioTML tagger for RE task selection on @Note2.

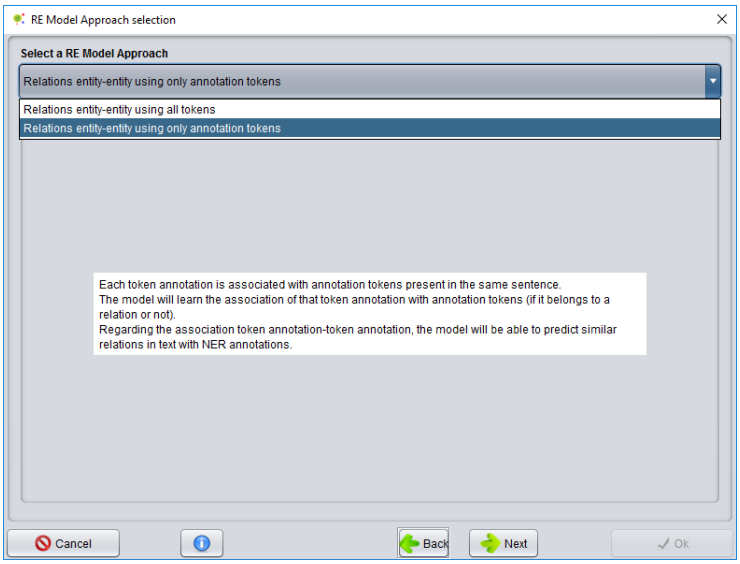

Figure 23: RE approach selection for ML model creation on BioTML.

The BioTML tagger contains steps that are similar to an NER process. The selection of a new or previous model configuration set (Figure 7), the selection

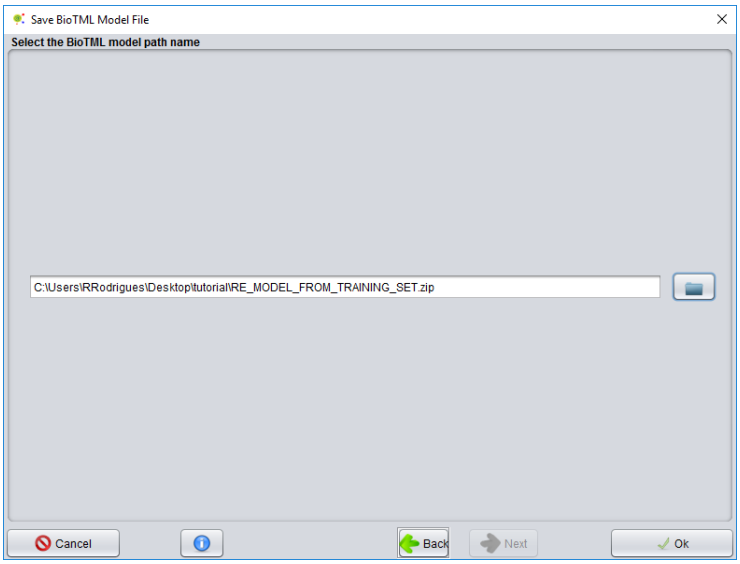

Figure 24: BioTML ML file result wizard on @Note2.

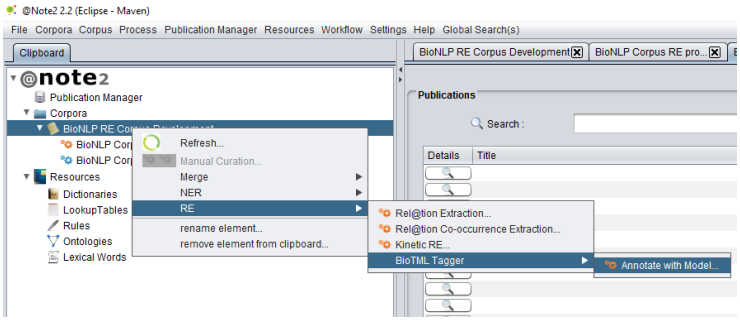

Figure 25: Application of BioTML tagger on BioNLP development set Corpus

of a NLP tokenizer (Figure 8), the features selection to train the model (Figure 9) and the ML model algorithm configuration (Figure 10) are steps in common to the NER process and we used the same settings as executed in the previous NER task.

After those steps, a wizard page appears to select which RE approach will be used to create the ML model. In this tutorial, we selected relations entity-entity using only annotation tokens, as shown in Figure 23 (the description of each approach appears in the GUI). Finally, the last step of the RE model creation is to select a directory and file name for the resulting zip model file (figure 24).

The ML model for the RE task is created after some time and can be used in the development set to predict relations. To load the model, click in the BioNLP development set corpus on @Note2 with the mouse right button and select RE - BioTML tagger - Annotated with model (Figure 25).

A wizard appears to select which NER process will be used to perform the RE annotation, as shown in Figure 26 (entities from the NER process are used to be associated with the ML model). Clicking in the Next button, the model file selection wizard page appears (Figure 27). Finally, the last wizard page with all model configuration information, NLP tokenizer and number of CPU threads selection is shown to the user (Figure 28).

An RE process is created after clicking in the Ok button and can be used to

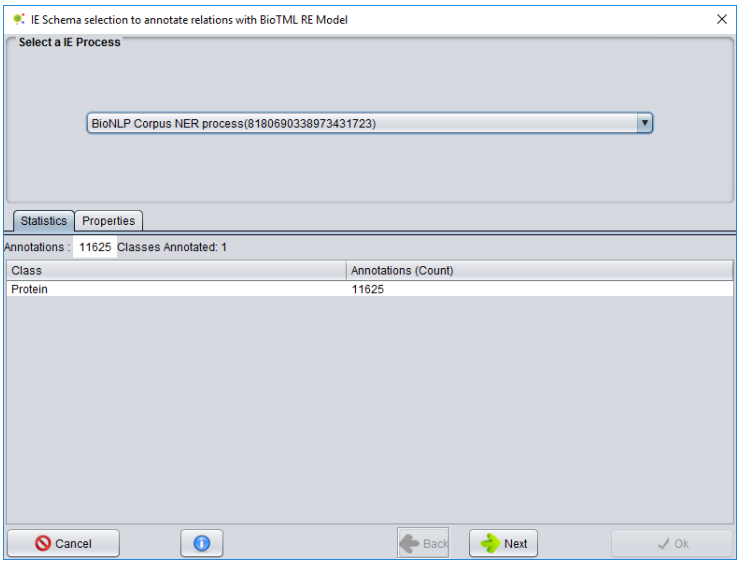

Figure 26: NER process selection to apply BioTML ML RE wizard.

evaluate against the loaded BioNLP RE development set process. The evaluation steps are similar to the previously shown in the NER task.

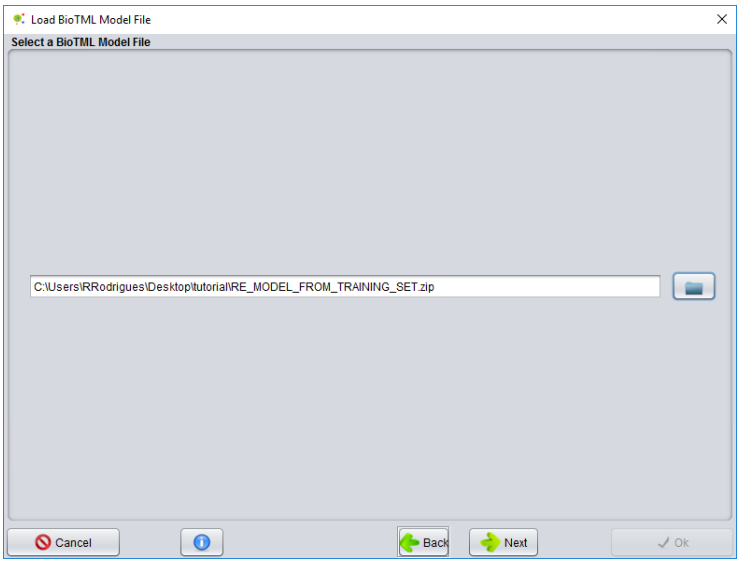

Figure 27: BioTML ML RE file selection wizard.

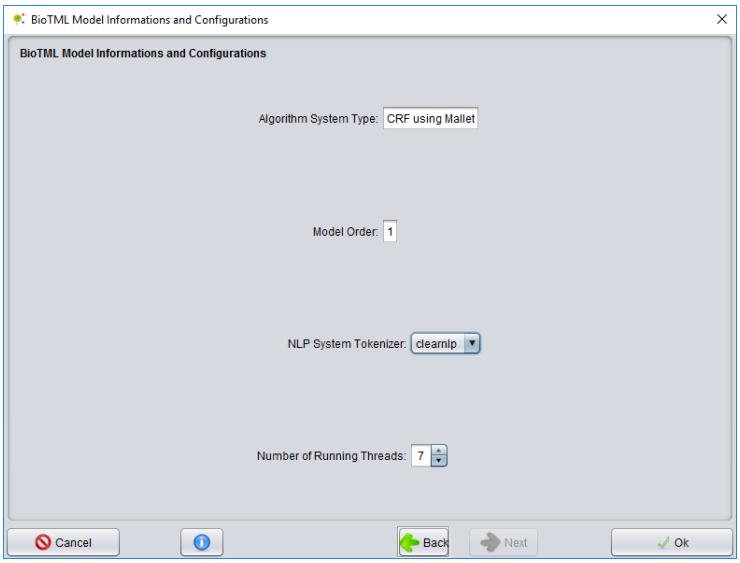

Figure 28: ML model information, NLP and CPU threads selection wizard