AHLT Laboratory Project

The lab project consists in building a system for *SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts*. You can replicate some participant system(s), or design and implement your own system.

The DDIExtraction 2013 task concerned the recognition of drugs and extraction of drug-drug interactions that appear in biomedical literature.
Two subtasks were proposed for the challenge:

1) The recognition and classification of drug names.
2) The extraction and classification of their interactions.

Both tasks are independent and evaluated separately (that is, the second task is evaluated on the gold standard drugs, not on the output of the first task).

The participants were free to address either one the tasks, or both. However, for the lab project you are required to address **both** of them.

1. Task description and participant systems

The official site of the challenge is https://www.cs.york.ac.uk/semeval-2013/task9.html.

- A description of the results of the challenge can be found in [Segura-Bedmar at al, 2013]
- Also, papers describing each participant system are also available.

All papers can be found at https://aclanthology.coli.uni-saarland.de/events/semeval-2013 or in the `papers` folder provided with the lab project material.

2. Challenge data

The corpus used for the task is the DDI corpus [Herrero-Zazo et al, 2013].
A short description of the DDI corpus provided by *SemEval-2013 Task 9* organizers can be found at https://www.cs.york.ac.uk/semeval-2013/task9/data/uploads/the-corpus-ddi.pdf. and in the `papers` folder.

A copy of DDI corpus is included in the attached material. Please follow the license constraints regarding distribution and use.

3. Resources

To develop your systems you'll need language processing and machine learning toolkits.

You may also resort to ontologies and databases to obtain lists of drug names, or other medical terminology.

Some recommended tools and resources:

- **Language processing**
  - *NLTK*: http://www.nltk.org/
  - *FreeLing*: http://nlp.cs.upc.edu/freeling
  - *TextServer: http://textserver.cs.upc.edu/textserver*

- **Machine Learning**
  - *SciPy*: https://www.scipy.org/
  - *scikit-learn*: http://scikit-learn.org/
  - *Keras*: https://keras.io
  - *PyTorch:* https://pytorch.org/
  - *crfsuite:* http://www.chokkan.org/software/crfsuite/
    https://github.com/scrapinghub/python-crfsuite

- **External Knowledge**
  - *DrugBank*: https://www.drugbank.ca/
  - *HSDB*: https://sis.nlm.nih.gov/enviro/hsdbchemicalslist.html

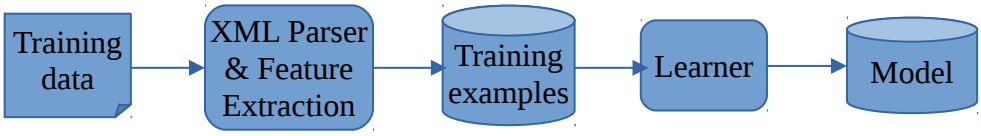
- **Utility tools**
  - *XML.dom*: https://docs.python.org/3.7/library/xml.dom.html

- **Word Embeddings**
  - *Word2vec:* https://code.google.com/archive/p/word2vec/
  - *FastText:* https://fasttext.cc/
  - *Glove:* https://nlp.stanford.edu/projects/glove/

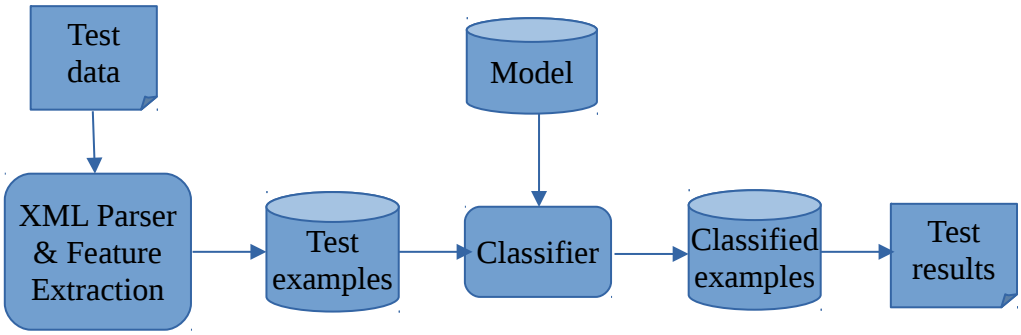
4. System architecture and Baseline

Your system will need to have two components: A training module and a test module.

The training module will read the training data from XML files, extract feature vectors (or any other representation suitable for the chosen ML method) and learn a model.



The test module will read the test data from XML files, extract feature vectors, run them through the learned model, convert the results to the appropriate task target, and output them in the required format.



A dummy rule-based baseline for each task is provided in `baseline` folder. They include the XML handling steps, so they can be used as an skeleton to develop your modules.

5. Contents of the Lab package

The package for the Lab project contains the following folders and files

`LabProjectAHLT.pdf`	- This file.
`data/`	- Folder containing the training and test corpus
`Train`	- Train corpus (the same training data are used for both tasks)
`Test-NER`	- Test corpus for task 9.1 (Drug name detection)
`Test-DDI`	- Test corpus for task 9.2 (DDI detection)
`papers/`	- Folder containing papers about the task
`Corpus/`	- Folder containing papers about the corpus
`SharedTask`	- Folder containing papers about the shared task and participant systems approaches.
`Evaluation`	- Folder containing papers about evaluation metrics, and formats for evaluation scripts.
`eval/`	- Folder containing evaluation scripts
`sessions/`	- Folder containing material for the first Lab sessions, where the baselines will be presented

6. Result Evaluation

To assess the performance of your systems, you must use the official *Semeval 2013 Task 9* evaluation scripts.

The scripts are included in folder `eval` in the lab package, and are used as follows:

Drug NER Task Evaluation (9.1)

To perform evaluation of the task 9.1, you should provide the script with the gold standard dataset directory and the output of your system (in a single file) formatted as described in the evaluation scripts documentation (and as the baseline program does):

```
java -jar eval/evaluateNER.jar goldDir submissionFile
```

where:

- `goldDir` is the directory where the gold standard dataset can be found (e.g. `data/Test-NER/DrugBank`)
- `submissionFile` is a single file containing all your predicted annotations for this dataset (for example, `task9.1_UC3M_1.txt`)

DDI Task Evaluation (9.2)

To perform evaluation of the task 9.2, you should provide the script with the gold standard dataset directory and the output of your system (in a single file) formatted as described in the evaluation scripts documentation (and as the baseline program does):

```
java -jar eval/evaluateDDI.jar goldDir submissionFile
```

where:

- `goldDir` is the directory where the gold standard dataset has been saved (e.g. `data/Test-DDI/DrugBank`)
- `submissionFile` is a single file containing all your predicted annotations for this dataset (for example, `task9.2_UC3M_1.txt`)

7. Required tasks

To carry out your project you will have to:

- Implement at least one baseline approach for each task (NER and DDI) based on a simple statistical or ML system for each task. You can extend the proposed baselines with some statistical information, or propose a different baseline (e.g. Naive Bayes).
- Implement at least one advanced ML approach for each task (NER and DDI). You can reproduce one of the proposals by the participants in the Task, or a variation of them, choose another approach (e.g. based on neural networks, such as [Raj et al. 2017, Asada et al. 2017, Lim et al. 2018]), or devise a new model.
- Evaluate your baselines and your systems using the official evaluation scripts. Self-built evaluation will **NOT** be accepted.
    - Note that building your system may involve using different tools, depending on the selected approach. For a classical classifier approaches such as SVM or CRF, you will need to extract features from the text, and thus to use NLP processing tools to get features such as the lemma, Part-of-Speech, or subject-verb relations. If you go for a neural network approach, you will need to use a word-embedding tool to create semantic representations of the words.
- Experimentally tune your system parameters to find the best configuration. Parameters to be tuned depend on the used algorithm, and should include: used features (or feature groups), threshold used to filter out low-frequency features, parameters of the algorithm (C, gamma, and kernel for SVM, learning rate for ANN, norm regularization for CRF, etc)
    - Tuning experiments should be done either on a development part extracted from the training set, or via cross-validation on the training set. Once the best configuration is selected, the resulting model can be applied to the Test set.
- Report experiments performed, conclusions extracted, and final results (both on development and test).

8. References

[Segura-Bedmar et al, 2013] I. Segura-Bedmar, P. Martínez, M. Herrero Zazo. **SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013)**. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pg 341--350, Atlanta, Georgia, USA, 2013.

[Herrero-Zazo et al, 2013] M. Herrero-Zazo, I. Segura-Bedmar, P. Martínez, T. Declerck: **The DDI corpus: An annotated corpus with pharmacological substances and drug-drug interactions**. *Journal of Biomedical Informatics* 46(5): 914-920 (2013)

[Raj et al, 2017] D. Raj, S. K. Sahu, A. Anand. **Learning local and global contexts using a convolutional recurrent network model for relation classification in biomedical text**. *Proceedings of 21th CoNLL 2017*, pages 311-321

[Lim et al, 2018] S. Lim, K. Lee, J. Kang. **Drug drug interaction extraction from the literature using a recursive neural network**. *PLOS ONE* 13(1): e0190926. 2018

[Asada et al, 2017] M. Asada, M. Miwa, Y. Sasaki. **Extracting Drug-Drug Interactions with Attention CNNs**. *Proceedings of BioNLP 2017*, pg 9-18. Vancouver, Canada, 2017