merpy

Feb 07, 2023

Contents:

This software provides a Python interface with MER - Minimal Entity Recognition. MER is a Named-Entity Recognition tool which given any lexicon and any input text returns the list of terms recognized in the text, including their exact location (annotations).

Given an ontology (owl file) MER is also able to link the entities to their classes.

MER is a Named-Entity Recognition tool which given any lexicon and any input text returns the list of terms recognized in the text, including their exact location (annotations).

Given an ontology (owl file) MER is also able to link the entities to their classes.

CHAPTER 1

Dependencies

1.1 awk

MER was developed and tested using the GNU awk (gawk) and grep. If you have another awk interpreter in your machine, there's no assurance that the program will work.

For example, to install GNU awk on Ubuntu:

```
sudo apt-get install gawk
```

Currently, merpy will not run unless gawk is available.

1.2 ssmpy

To calculate similarities between the recognized entities

```
pip install ssmpy
```

CHAPTER 2

Installation

```
pip install merpy
```

or

```
python setup.py install
```

Then you might want to update the MER scripts and download preprocessed data:

```
>>> import merpy
>>> merpy.download_mer()
>>> merpy.download_lexicons()
```

CHAPTER 3

Basic Usage

```
>>> import merpy
>>> merpy.download_lexicons()
>>> lexicons = merpy.get_lexicons()
>>> merpy.show_lexicons()
lexicons preloaded:
['cl', 'osci', 'lexicon', 'bireme_decs_por2020', 'bireme_decs_eng2020', 'ecto', 'go',
 ↪'hp', 'wordnet-hyponym', 'doid', 'bireme_decs_spa2020', 'radlex', 'envo', 'chebi_
 ↪lite']

lexicons loaded ready to use:
['osci', 'bireme_decs_por2020', 'radlex', 'go', 'envo', 'doid', 'chebi_lite', 'ecto',
 ↪'bireme_decs_spa2020', 'bireme_decs_eng2020', 'wordnet-hyponym', 'cl', 'hp',
 ↪'lexicon']

lexicons with linked concepts:
['doid', 'bireme_decs_por2020', 'lexicon', 'bireme_decs_spa2020', 'osci', 'bireme_
 ↪decs_eng2020', 'go', 'hp', 'cl', 'radlex', 'chebi_lite', 'ecto', 'envo']

>>> document = 'Influenza, commonly known as "the flu", is an infectious disease␣
 ↪caused by an influenza virus. Symptoms can be mild to severe. The most common␣
 ↪symptoms include: a high fever, runny nose, sore throat, muscle pains, headache,␣
 ↪coughing, and feeling tired ... Acetylcysteine for reducing the oxygen transport␣
 ↪and caffeine to stimulate ... fever, tachypnea ... fiebre, taquipnea ... febre,␣
 ↪taquipneia, ... neuronal stem cell, water vapour saturated air'
>>> entities = merpy.get_entities(document, "hp") # get_entities_mp uses␣
 ↪multiprocessing (set n_cores param)
>>> print(entities)
[['111', '115', 'mild', 'http://purl.obolibrary.org/obo/HP_0012825'], ['119', '125',
 ↪'severe', 'http://purl.obolibrary.org/obo/HP_0012828'], ['168', '173', 'fever',
 ↪'http://purl.obolibrary.org/obo/HP_0001945'], ['181', '185', 'nose', 'http://purl.
 ↪obolibrary.org/obo/UBERON_0000004'], ['200', '206', 'muscle', 'http://purl.
 ↪obolibrary.org/obo/UBERON_0005090'], ['214', '222', 'headache', 'http://purl.
 ↪obolibrary.org/obo/HP_0002315'], ['224', '232', 'coughing', 'http://purl.obolibrary.
 ↪org/obo/HP_0012735'], ['246', '251', 'tired', 'http://purl.obolibrary.org/obo/HP_
 ↪0012378'], ['288', '294', 'oxygen', 'http://purl.obolibrary.org/obo/CHEBI_15379'], [
 ↪'295', '304', 'transport', 'http://purl.obolibrary.org/obo/GO_0006810'], ['335',
 ↪'340', 'fever', 'http://purl.obolibrary.org/obo/HP_0001945'], ['342', '351',
 ↪'tachypnea', 'http://purl.obolibrary.org/obo/HP_0002789'], ['415', '419', 'cell',
 ↪'http://purl.obolibrary.org/obo/CL_0000000'], ['175', '185', 'runny nose', 'http://
 ↪purl.obolibrary.org/obo/HP_0031417'], ['187', '198', 'sore throat', 'http://purl.
 ↪obolibrary.org/obo/HP_0033050'], ['288', '304', 'oxygen transport', 'http://purl.
 ↪obolibrary.org/obo/GO_0015671'], ['410', '419', 'stem cell', 'http://purl.
```

(continues on next page)

(continued from previous page)

```
>>> entities = merpy.get_entities(document, "bireme_decs_por2020")
>>> print(entities)
[['0', '9', 'Influenza', 'https://decs.bvsalud.org/ths/?filter=ths_regid&q=D007251'],␣
↪['78', '87', 'influenza', 'https://decs.bvsalud.org/ths/?filter=ths_regid&q=D007251
↪'], ['378', '383', 'febre', 'https://decs.bvsalud.org/ths/?filter=ths_regid&
↪q=D005334'], ['385', '395', 'taquipneia', 'https://decs.bvsalud.org/ths/?filter=ths_
↪regid&q=D059246'], ['410', '414', 'stem', 'https://decs.bvsalud.org/ths/?filter=ths_
↪regid&q=D017348']]

>>> merpy.create_lexicon(["gene1", "gene2", "gene3"], "genelist")
wrote genelist lexicon
>>> merpy.process_lexicon("genelist")
================
gene1
gene2
gene3
================
================
================
================
>>> merpy.delete_lexicon("genelist")
deleted genelist lexicon

>>> merpy.download_lexicon("https://github.com/lasigeBioTM/MER/raw/biocreative2017/
↪data/ChEBI.txt", "chebi")
wrote chebi lexicon
>>> merpy.process_lexicon("chebi")
================
lannate
1.2.di.o.oleoyl.3.o..beta.d.galactopyranosyl..sn.glycerol
manganese.2..
n.butyraldoxime
brocillin
beta.d.glc..1..4...l.alpha.d.hep...1..3...l.alpha.d.hep
.clo3....
tetrasulfanide
bis.perfluorobutylethene
3..gmp
================
.5z.8z.14z..11.12.epoxyicosa.5.8.14.trienoic acid
presqualene diphosphate
tuberculosinol diphosphate
1.2.3.4.butanetetralyl tetranitrate
soyasaponin bb
trilithium citrate
.s..2..o.chlorophenyl..2..methylamino.cyclohexanone hydrochloride
beta.l.ara4n.lipid a.2..
7.chloroindole.3.acetic acid
.e..3..trifluoromethyl.cinnamic acid
================
dopamine dimethyl ether
acetyl.2...5...phosphoribosyl..3..dephospho.coenzyme a serine residue
3.methylbut.3.enyl diphosphate trianion
disodium 3.3..azobis .6.hydroxybenzoate.
potassium mercuric iodide
nickel.ii. sulfate .1.1.
```

(continues on next page)

(continued from previous page)

```
fentanyl dihydrogen citrate
.gal.1 .glcnac.1 .man.1
sodium nitroprusside dihydrate
cholic acid taurine conjugate
================
.s..4.amino.5.oxopentanoic acid
glutathione disulfide
hydroquinone benzyl
6.hydroxyriboflavin 5...trihydrogen
dopamine dimethyl
3.methylbut.3.enyl diphosphate
disodium 3.3..azobis
potassium mercuric
fentanyl dihydrogen
cholic acid
================
```

CHAPTER 4

Semantic Similarities

```
$ curl -L -O http://labs.rd.ciencias.ulisboa.pt/dishin/chebi202302.db.gz
$ gunzip -N chebi202302.db.gz
```

```
>>> import merpy
>>> merpy.process_lexicon("lexicon")
>>> document = "α-maltose and nicotinic acid was found, but not nicotinic acid D-
↪ribonucleotide"
>>> entities = merpy.get_entities(document, "lexicon")
>>> merpy.get_similarities(entities, 'chebi.db')
[['0', '9', 'α-maltose', 'http://purl.obolibrary.org/obo/CHEBI_18167', 0.
↪026437365432380268], ['14', '28', 'nicotinic acid', 'http://purl.obolibrary.org/obo/
↪CHEBI_15940', 0.07969957014235445], ['48', '62', 'nicotinic acid', 'http://purl.
↪obolibrary.org/obo/CHEBI_15940', 0.07969957014235445], ['48', '79', 'nicotinic acid␣
↪D-ribonucleotide', 'http://purl.obolibrary.org/obo/CHEBI_15763', 0.
↪07969957014235445]]
```

CHAPTER 5

API

CHAPTER 6

Changelog

1.1.0 - Run get_entities with multiprocessing

1.1.1 - Add get_similarities function

CHAPTER 7

Reference

More information about MER can be found in:

- MER: a Shell Script and Annotation Server for Minimal Named Entity Recognition and Linking, F. Couto and A. Lamurias, Journal of Cheminformatics, 10:58, 2018 [https://doi.org/10.1186/s13321-018-0312-9]
- MER: a Minimal Named-Entity Recognition Tagger and Annotation Server, F. Couto, L. Campos, and A. Lamurias, in BioCreative V.5 Challenge Evaluation, 2017 [https://www.researchgate.net/publication/316545534_MER_a_Minimal_Named-Entity_Recognition_Tagger_and_Annotation_Server]

7.1 Indices and tables

- genindex
- modindex
- search