

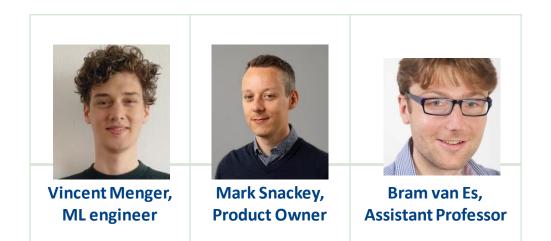
## 🖺 clinical + 🖛 nl + 🗎 NLP = 抹 clinlp

# **clinlp**: a collaborative platform for advancing Dutch clinical NLP



Vincent Menger, Mark Snackey & Bram van Es 3rd Dutch Clinical NLP Workshop June 24, 2024

### About us...



- Work on AI, ML, NLP @ UMCU
  - Team analytics
  - UPOD
- About myself
  - Technical background (CS/AI), ~4yrs PhD @ UU/UMCU, ~2yrs industry experience, now ~3yrs DIT Analytics



## Introduction

| If I work on                    | I typically use    |  |  |
|---------------------------------|--------------------|--|--|
| Processing tabular data         | pandas             |  |  |
| Classical machine learning      | learn              |  |  |
| Neural Networks                 | TensorFlow OyTorch |  |  |
| Pretrained models, transformers | Hugging Face       |  |  |
| Natural Language Processing     | spaCy              |  |  |
| NLP with Dutch clinical text    | clinlp             |  |  |



## What do these libraries have in common?



- 1. Provide useful tools
- 2. Standardized framework
- 3. Production-ready quality
- 4. Open source collaboration



```
Mot
                                                                                            Ē
           tokenizer = AutoTokenizer\
                        .from_pretrained("UMCU/MedRoBERTa.nl_NegationDetection")
           model = AutoModelForTokenClassification\
     Ca
                       .from pretrained("UMCU/MedRoBERTa.nl NegationDetection")
     ps'
           some_text = "De patient was niet aanspreekbaar en hij zag er grauw uit. \
           Hij heeft de inspanningstest echter goed doorstaan."
           inputs = tokenizer(some_text, return_tensors='pt')
 Geo
           output = model.forward(inputs)
UMCU/Mee
           probas = torch.nn.functional.softmax(output.logits[0]).detach().numpy()
Token Class
              associate with tokens
           #
 Setup
           input_tokens = tokenizer.convert_ids_to_tokens(inputs['input_ids'][0])
 User Gui
           target_map = {0: 'B-Negated', 1:'B-NotNegated',2:'I-Negated',3:'I-NotNegated'}

    Develope

           results = [{'token': input_tokens[idx],
 For Algorit
                             'proba_negated': proba_arr[0]+proba_arr[2],
                             'proba_not_negated': proba_arr[1]+proba_arr[3]
                             3
                            for idx,proba_arr in enumerate(probas)]
                                  NL.txt
                               E
                                  NL1.txt
```

5

**MC Utrecht** 

### **Motivation**



● Java + ♀ 0 + ☆ 1 + ⊙ 0 + № 0 - Updated on Mar 22, 2021

● Java ・ ♀ 0 ・ ☆ 1 ・ ⊙ 0 ・ ∏ 0 ・ Updated on Jan 11, 2018

Summaries

ContextDutch Private

biosemantics.

Electronic supplementary mate

Additional file 1: Regular copre Temporality module. (DOCX 41 ContextD-2 Research article | <u>Connaccess</u> | Published: 29 Newmber 2014 **ContextD:** an algorithm to identify contextual properties of medical terms in a Dutch clinical corpus Zuber Atal<sup>10</sup>, Jeward Pan, Ning Jam, Mrian C.M. Statember, Marin J. Schwerie 4 Jan A.Kers

BMC Bioinformatics 15, Article number: 373 (2014) Cite this article

```
• Case: detecting negation for psychosis symptoms
```



import spacy
import clinlp

```
nlp = spacy.blank("clinlp")
```

# Rule based
nlp.add\_pipe("clinlp\_context\_algorithm")

# Transformer based
nlp.add\_pipe("clinlp\_negation\_transformer")





### Provide useful NLP tools



### **1. Provide useful NLP tools**

- 2. Standardized framework
- 3. Production-ready quality
- 4. Open source collaboration



## Provide useful NLP tools

- Incudes various components that perform generic NLP tasks
- Tailored to Dutch clinical text
- Work out of the box, but customizable
- Different methods: rule-based, classical ML, transformers, etc.

### Tokenizing

| Prednisolon gebruik 20 mg/dag      | Prednisolon gebruik 20 mg / dag   |
|------------------------------------|---|
| Mn 's nachts wel last              | Mn 's nachts wel last   |
| NIBP[mmHg]: 124/85(98)             | NIBP         [         mmHg         ]         :         124         /         85         (         98 |
| Zie status [PERSOON-1] [DATUM-3]   | Zie status [PERSOON-1] [DATUM-3]  |
| disharmonisch profiel VIQ>PIQ      | disharmonisch profiel VIQ > PIQ   |
| Contracties;-                      | Contracties ; -   |
| VG/ o.a. alcoholische levercirrose | VG / o.a. alcoholische levercirrose   |



## Provide useful NLP tools: components (1)

### Sentence boundary detection

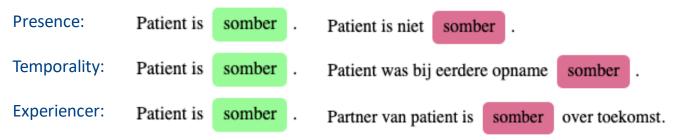
| Betreft: dhr. [NAAM].  |
|--|
| Geachte collega,   |
|  |
| Bovengenoemde patient bezocht onze polikliniek jl. februari.    Er werd de |
| volgende medicatie z.n. voorgeschreven:                                    |
| II- Prednisolon 20 mg/dagli  |
| II- Pantoprazol 1 ddll   |
|  |
| Hoogachtend, [NAAM]  |

### **Entity recognition**

Hypotensie bij prematuur **Exact matching** geboren baby. ٠ Hypotensie Normalized matching bij prematuur geboren baby. • Hypotensie bij prematur geboren baby. **Fuzzy matching** ٠ Bloeddruk sterk verlaagd bij prematuur geboren baby. **Proximity matching** ٠ Pseudo matching Hypotensie bij baby met prematuur ademhalingspatroon. ٠

## Provide useful NLP tools: components (2)

### **Entity qualification (rule-based, transformer-based)**



- Currently: information extraction components
- Plans for: NER+L, prodigy compatibility, active learning, summarization, speech to text, ... anything!



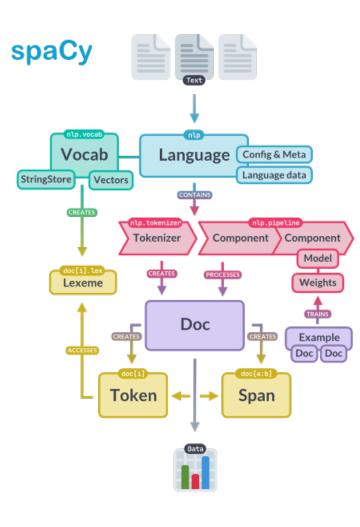
## Standardized framework



Provide useful NLP tools
 Standardized framework
 Production-ready quality
 Open source collaboration



## Standardized framework



- Organized in spaCy framework
  - Easy implementation
  - Consistency
  - Good integration
  - Mix and match components in pipeline
  - Wrap pre-trained models

#### @dataclass(frozen=True)

#### class Qualifier:

A qualifier for an entity.

A qualifier is a piece of information that provides additional context to an entity. For example, a ``Presence`` qualifier with a value of ``Present`` or ``Absent``. A qualifier has a fixed value.

name: str = field(compare=True)
"""The name of the qualifier."""

value: str = field(compare=True)
"""The value of the qualifier."""

is\_default: bool = field(compare=True)
"""Whether the value is the default value."""

priority: int = field(default=0, compare=False)
"""The priority of the qualifier."""

prob: Optional[float] = field(default=None, compare=False)
"""The probability of the qualifier."""



## Standardized framework: full IE pipeline

```
ιQ
import spacy
from clinlp.ie import Term
nlp = spacy.blank("clinlp")
# Normalization
nlp.add_pipe("clinlp_normalizer")
# Sentences
nlp.add_pipe("clinlp_sentencizer")
# Entities
concepts = {
    "prematuriteit": [
        "preterm", "<p3", "prematuriteit", "partus praematurus"
    ],
    "hypotensie": [
        "hypotensie", Term("bd verlaagd", proximity=1)
   ],
    "veneus_infarct": [
        "veneus infarct", Term("VI", attr="TEXT")
    1
}
entity_matcher = nlp.add_pipe("clinlp_rule_based_entity_matcher", config={"attr": "NORM", "fuz
entity_matcher.load_concepts(concepts)
# Qualifiers
nlp.add_pipe("clinlp_context_algorithm", config={"phrase_matcher_attr": "NORM"})
```



## **Production-ready quality**



Provide useful NLP tools
 Standardized framework
 Production-ready quality
 Open source collaboration



## **Production-ready quality**

### Follows industry best practices, such as:

| class QualifierDetector(Pipe):   |
|--|
| """Abstract pipeline component for detecting qualifiers in clinical text.""" |
|  |
| definit(   |
|  |
| @property  |
| @abstractmethod  |
| def qualifier_classes(self) -> dict[str, QualifierClass]:…                   |
|  |
| @staticmethod  |
| def add qualifier to ent(entity: Span, new qualifier: Qualifier) -> None:    |
|  |
| <pre>def _initialize_ent_qualifiers(self, entity: Span) -&gt; None:</pre>    |
|  |
| @abstractmethod  |
| <pre>def detect qualifiers(self, doc: Doc) -&gt; None:</pre>                 |
|  |
| defcall(self, doc: Doc) -> Doc:  |
|  |

### Proper system design with relevant abstractions, etc.

| ⊘ 634/634              | 73.6s 신 |
|------------------------|---------|
| $\sim \oslash$ clinlp  |         |
| $\sim$ $\oslash$ tests |         |
| > 🕗 integration        |         |
| > 🕗 regression         |         |
| $\sim \oslash$ unit    |         |
| > ⊘ ie                 |         |
| > 🕗 metrics            |         |
| > 🕗 test_language.py   |         |
| > 🕗 test_normalizer.py |         |
| >                      |         |
| ) ⊘ test_util.py       |         |
|                        |         |

### Tests with good coverage

| ٠ | >  | ruff  | for   | nat | &&   | ruff   | check |
|---|----|-------|-------|-----|------|--------|-------|
|   | 41 | .fil€ | es le | eft | unc  | :hange | ed    |
|   | Al | l ch€ | ecks  | pas | sseo | 11 -   |       |

### Clean and consistent codebase

| () 2 weeks ago 852f345    | 📓 zip 📓 tar.gz   | 🗋 Notes |
|---------------------------|------------------|---------|
|                           |                  |         |
| V0.7.0 m                  |                  |         |
| () on May 16 -O- ac4a54c  | 🖁 zip 📓 tar.gz [ | Notes   |
| v0.6.6 🚥                  |                  |         |
| 🕒 on Apr 24 - 🗢 262d56c   | 🗿 zip 👔 tar.gz [ | Notes   |
| v0.6.5 🚥                  |                  |         |
| ( ) on Apr 24 - ↔ 009bf34 | 🕅 zip 🕼 tar.gz [ | Notes   |

🕓 on Feb 13 🗢 Øb19a52 👔 zip 👔 tar.gz 🌓 Notes

### **Frequent releases**

@ .#

| 266 workflow runs  |                         | Event + | Status +           | Branch + | Actor - |
|--|-------------------------|---------|--------------------|----------|---------|
| Merge pull request #101 from umcu/add-documentation-pages<br>format-lint #121: Commit Clar1332 pushed by vmenger | main                    |         | ⊟ 20 ho<br>⊘ 33s   | urs ago  |         |
| Merge pull request #101 from umcu/add-documentation-pages<br>tests #121: Connit Ots1332 pushed by vmanper        | main                    |         | 台 20 ho<br>⑦ 5m 3d |          |         |
| Add documentation pages tests #120: Pall request #10! synchronize by vmenger                                     | add-documentation-pages |         | 런 20 ho<br>🕑 4m 43 |          |         |
| Add documentation pages format-lint #120: Pull request #101 synchronize by vmenger                               | add-documentation-pages |         | 20 ho              | urs ago  |         |

### CI/CD, automated workflows



### docs: clinlp.readthedocs.io



## **Open source collaboration**



Provide useful NLP tools
 Standardized framework
 Production-ready quality
 Open source collaboration



## **Open source collaboration**

- Committed to maintaining clinlp as a free and open source package
- The more people join, the more it works for everyone
- Benefits: reproducibility, comparing results, fast prototyping, robust pipelines, availability, sharing work, easier collaboration
- If you have some code/project lying around, please contact us!



## We need your help!



### How to join?

If this is relevant to you  $\rightarrow$  use it If you want features or found bugs  $\rightarrow$  give feedback If you have a contribution  $\rightarrow$  **contact us**, or fork, change and PR

### **Contact** GitHub issue / <u>v.j.menger-2@umcutrecht.nl</u> / <u>bes3@umcutrecht.nl</u>) / afterwards ©

Thanks to: Bram, Daniel, Karin, Malin, Mark, Martijn, Richard, Rosanne, Ruben, Sander, Saskia, Saskia, Teus, Tuur, Vincent, Zimbo

