Clinical Named Entity Recognition in Swedish Patient Records

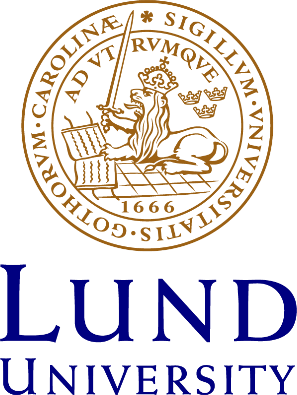
using ICD-10-SE

**Keywords**: named entity recognition, clinical text mining, corpora development

Klinisk Entitets-identifiering

i svenska patientjournaler användande ICD-10-SE

**Nyckelords**: naturlig språkbehandling, klinisk text mining, korpusutveckling

**Author**: Carl Ollvik Aasa

**Email**: [tna14col@student.lu.se](mailto:tna14col@student.lu.se)

**Supervisors:**

*Sonja Aits* (main supervisor)

*Cell Death, Lysosomes and Artificial Intelligence Group*, Department of Experimental Medical Science, Faculty of Medicine, Lund University, Lund, Sweden

*Johanna Berg* (co-supervisor)

Malmö University Hospital and Department of Global Public Health, Karolinska Institute

**Examinator: Pär Halje**

# Table of contents

Table of contents 2

Abstract 3

Populärvetenskaplig sammanfattning 4

1 Introduction 6

1.1 Purpose and Research Questions 6

2 Methods 7

2.1 Computational Environment 7

2.2 Corpus and Annotation 7

2.2.1 SpaCy and Swedish transformer model 8

2.2.2 NegEx 8

2.3 Evaluation 8

2.4 Citizen Science 9

3 Ethical Considerations 9

4 Results 10

4.1 Creation of an annotated gold standard corpus with fictional Swedish patient records 10

4.2 Creation of Natural Language Processing Model 10

4.3 Citizen Science 10

5 Discussion 11

5.1 Pretrained ML model 11

5.2 Negation Recognition 11

5.3 Symptoms and Findings Recognition 12

5.4 Limitations 12

5.5 Relation to Other Work 14

6 Conclusions 14

Declaration of own contribution 15

Acknowledgements 16

References 17

Appendix 19

# Abstract

**Background**:

**Objective**:

**Methods**:

**Results**:

**Conclusions:**

# Populärvetenskaplig sammanfattning

# INTRODUCTION

Records of the medical care given to patients, called *health records* or *medical charts* along with other terms, contain a wealth of information. In them are patients’ subjective symptom descriptions, findings yielded from medical examination and tests done by medical personnel, diagnoses and diagnostic reasoning, treatments, drug administration, adverse effects and more. Their primary purpose is as a tool in the daily work of medical personnel caring for a patient, but the information they contain can also have secondary utility in improving care for other patients, be used for prioritization and structuring for health care systems and policies, as well as for various medical research and epidemiological studies.(1)(2)(3) Despite this potential, electronic health records (EHRs) are still only beginning to be used for such purposes and in Sweden are still rarely being used. This is partly due to the lack of tools to process the data. Some data in EHRs consists of structured data, such as laboratory test results, but much of the rich and nuanced information about symptoms, findings and diagnostic reasoning consists of unstructured free-text narratives of patient care (e.g., admission notes, history and background of the patient - anamnesis -, results of examinations - status-and discharge letters - *epikris* in Swedish).(1)

The automatic processing of such free-text: *natural language processing* (NLP), defined as “any computer-based algorithm that handles, augments, and transforms natural language so that it can be represented for computation”(2), is rapidly improving with utilization and refinement of machine learning (ML) algorithms. Medical language and the language of EHRs differs from other text, such as news articles or encyclopedias, containing many non-standardized abbreviations and jargon that requires knowledge of the field as well as misspellings and syntactically incomplete sentences for brevity, e.g. not mentioning the patient as the subject.(4) In addition, the data contain sensitive information, such as addresses, occupation, and id numbers, which requires solutions and policies to protect patient integrity. (18) Despite these challenges NLP has been used for multiple healthcare tasks such as care and cost analysis, risk factor identification, pseudo-anonymization and information extraction from EHRs making NLP being the most widely used tool for “big data” analysis in healthcare.(5) (6) (2) A major part of such tasks is named entity recognition (NER) which is the automatic detection and assigning of parts of text to predefined concept/categories, such as locations, measurements, diseases, symptoms and findings.(1)

Most NLP tools are developed for English and there is a great need to develop tools for other languages and to explore the possibility of using techniques and systems across languages. In addition to obvious differences in words an grammatical structure, Swedish poses special challenges, as compounding of words occurs frequently and the language is more inflective than English.(7)(8) Some adoptions of tools developed for German and English to other languages have been examined to varied degree. A study by Skeppshult et. al in 2014 looked at the entities Disorder*,* Finding*,* Pharmaceutical DrugandBody Structure in Swedish clinical text and concluded that “*English clinical entity recognition approaches were also suitable for Swedish”.* (9)(8)

The major bottleneck for developing new models is generating the data to train them. ML models require large sets of annotated text to be used as reference and the traditional solution in medicine is for clinicians to create this by hand. Doing this manual annotation is time-consuming and in a field such as medicine requires specialist knowledge.(1) This has clear limitations in scalability. In contrast to ML based NLP, terminology- and rule-based NLP, where rules for matching patterns, phrases and syntactic information are created, can be implemented and tested with small data sets and are especially useful when dealing with a finite number of examples, such as a set of symptoms. (10). Such NLP systems could be used to assist human annotators in the production of gold standard corpora and for generating computationally annotated silver standard corpora with which ML models later could be trained.

Terminology based systems for recognizing disorders, drug mentions and more have been created. A especially interesting subset use ontology based vocabularies, such as SNOWMED CT (11) and MeSH terminology has been used for NER in Swedish discharge summaries.(12) Initiatives such as open the Open Biological and Biomedical Foundry (obofoundry.org) adopted Symptom Ontology are furthering such ontologies by collaboration and open source.(13) These systems strength are that they come at a low development cost since no vocabulary has to be created and they can be deployed rapidly. It is therefore of interest to evaluate the utility of such systems using different ontological systems.

Such systems are of course limited. In 2013 Forbush et al manually annotated 750 clinical texts for subjective symptom expressions and found that 31% of symptom expression weren’t coded or could be mapped to standard terminology using International Classification of Diseases-Ninth edition CM (ICD-9-CM.(14) This serves to illustrate the complexity of symptom description and difficulties in creating such ontological systems. For the purpose of pre-annotation and development of silver standard corpora these systems could however still provide benefits.

Most previous NLP studies, both for Swedish and other languages have focused on the identification and extraction of disease information in clinical text, along with pharmaceutical information and detection of patient identifiers for anonymization.(2) (9) (15) However a great deal of the rich information in EHRs are symptoms, that is the subjective indications of disease experienced by patients and signs which are objective findings that can be made by examination or measurement by healthcare personnel. A 2019th systematic review of NLP in healthcare by Koleck et al (2) found that symptom information was presented as a primary outcome in only half of studies and about a third of studies using symptom as just a means of classifying disease. The sheer volume of data available on symptoms and findings through EHRs offers a major opportunity for improving symptom assessment and management, patient quality of life, disease categorization and developing NLP for symptoms and signs is an important step in improving healthcare.

The tenth edition of ICD adds major changes to the previous 9th edition, from which ICD-9-CM is based, adding many new codes. ICD-10 is available in the six official languages of the World Health Organization (WHO) and in 36 additional languages including Swedish (ICD-10-SE), which is used by *Socialsytelsen* and the Swedish health care system.(16)(17) It contains ca. 32,000 different diagnosis codes for diseases, symptoms, and findings adding about twice as many categories as previous editions which makes evaluating the utility ICD-10-SE codes for automatic labeling of symptoms and findings of interest. In 2012 Skeppsteadt et al evaluated a multi-step preprocessing with several terminology-based systems with the primary aim of matching of disease mentions in clinical notes from a Swedish emergency unit. They also used chapter 18 of ICD, *“Symtom, sjukdomstecken och onormala kliniska fynd och laboratoriefynd som ej klassificeras på annan plats”* for assigning the class *finding,* with which they found a precision of 0,57 and recall of 0,3 with no preprocessing of the codes. .. Maybe just drop this as it was found after? And explain? A bit of a weird situation as this isn’t really the main point of the project, just a test, but it’s the only thing we have data on

## OBJECTIVE

The purpose of the present study is to develop and evaluate a vocabulary based NLP model using ICD-10-SE codes for named entity recognition of symptoms and signs in free-text from health records; posing the research question: *Can symptoms and findings be automatically prelabelled to assist manual annotation in Swedish electronic health records using Swedish ICD10 codes?* In particular we aim manually annotate symptoms and findings, along with negations in a corpus of fictional patient charts from Swedish emergency rooms and compare to the automatic labeling done computationally matching ICD-10-SE codes from the R00-R99 categories applying and comparing preprocessing.,

# Methods

## Creation and

Ten which mirrors common clinical practice. independently

(19)Annotated entity categories where A symptom being a subjective experience while findings are objective evidence of disease and a finding, while a negation is any linguistic construction which inverts or denies the default meaning of a word or statement that it affects. Further definitions and clarifications in collaboration within accordance with SNOWMED CT and ICD descriptions and definitionsand which were incorporated in the annotation guidelines**.** , see appendix ?.?. Rules guiding annotation decisions were produced from a base set of definitions and principles prior to annotation and refined based on cases encountered in the annotation process. documented along and resolved.Complicated cases were discussed in the project team consisting of members trained in medicine and natural language processing to come to a consensus. Annotations were then redone by *APH1* and a medical student in accordance with the final guidelines before evaluation

### Evaluation & Gold Standard Corpora

The two manual annotation sets of the corpus was evaluated for inter-annotator agreement (IAA) by calculating the Precision, Recall, F-score, and Cohen’s kappa between the two annotators along with a confusion matrix for the entity classes. From the two annotation sets a gold standard was produced by the chief annotator.

## Named Entity Recognition

### SpaCy and Swedish transformer model

The NLP model was built using the spaCy[[1]](#footnote-2) library [version: 3.0.6 ]. A pretrained transformer-based model for Swedish [sv\_pipeline-0.0.0][[2]](#footnote-3) from The National Library of Sweden was used as base to add tokenization (i.e. splitting of texts into individual terms), dependency parsing (i.e. syntactic relationships), part-of-speech tagging (i.e. syntactic identities) and NER for time modifiers, organizations and metric terms for Swedish text. (20) The model was equipped with dictionaries of ICD codes and negations for NER and symptoms, signs, and negations, see the following sections for details. For these named entities all matching was done to the normalized lowercase version of texts and dictionaries.

### ICD

ICD-10-SE codes where collected from Socialstyrelsen (socialstyrelsen.se/globalassets/sharepoint-dokument/dokument-webb/klassifikationer-och-koder/icd-10-se-2021-text.zip, accessed 2021-04-03). Codes of categories R00-R99, “*Symtom, sjukdomstecken och onormala kliniska fynd och laboratoriefynd som ej klassificeras på annan plats* “ were extracted.

Automatic pre-processing of the ICD-10 code list was made to remove clarifying text within parentheses . Results of naïve application of raw codes was compared to processed code.

As ICD-10-SE does not provide distinctions between finding and symptoms for most categories no distinction of symptoms and findings is made in the NLP model or analysis.

### Negations

For NER of negations a Swedish version of NegEx, a rule-and dictionary-based algorithm for negation recognition, was added to the pipeline (7). The previously published list of negation terms was expanded to adapt it for use in patient journals with assistance from two project members with medical expertise (*APH1* and a senior biomedical researcher). This was added as a dictionary for the entity class NEG.

### Evaluation

The performance of the NLP pipeline for NER of negations, symptoms and signs was evaluated against the gold standard corpus. No weights were assigned to any specific document. The following metrics were calculated: precision: the percentage of named entities found by the NLP pipeline that are correct; recall: the percentage of named entities present in the corpus that are found by the system; F1 score: harmonic mean of precision and recall. corpus Span overlap. Cohens kappa

### Error Analysis

Error analysis was performed with a. False Positives (FP) are spans that the model incorrectly labeled as either symptom, sign, or negation. False negatives are …

## Computational Environment

Computations were performed using python 3.8.5. and bash scripting on WSL2 running ubuntu 20.04. INCEpTION was run locally on windows 10 and as a docker container in WSL2 for Windows 10 using MySQL as database backend.

## Ethical Considerations

In this stage of development, the immediate ethical issues are minimal, as all the chart information used is fictional. However, the intention is to deploy these techniques and models on real patient data (ethics permission for this has already been granted). It is important to consider proper data management to protect highly sensitive patient information. NLP models can easily be scaled to great proportions given the right hardware and therefore the consequences are large. Reliable de-identification is needed for many applications of these models, as models could be used for profiling, e.g. for sexual orientation, or by insurance companies for life. It’s important to only use the necessary data, such as only the data under a specific header, e.g. “anamnesis” and to carefully consider who can access raw results or processed data.

Another issue concerns the bias of the model. Training on real world data will reflect real word inequalities, such as different care for different patient groups or the assignment of lower risk to a patient group which is merely underdiagnosed because of bias. Defining inclusion/exclusion-criteria for data collection is of major importance to avoid these problems. This is of course a problem in the creation of gold-standard corpora as well were annotators bias and idiosyncrasies will affect the labelling. Standards for comparing the labels of different annotators, as well a thorough and transparent guidelines must be implemented and presented.

In summary the NLP has the potential to improve healthcare, but if patient integrity is not properly protected in law and by the implemented system, the potential of the system becomes the problem. It’s therefore important to continuously consider and develop safe protocols, methods, and software and that this is done in an open and transparent way.

# Results

## Corpus Creation and Annotation

For the corpus of ten fictional Swedish emergency room patient records was created. Annotation guidelines, see appendix were developed and the corpus was annotated for symptoms, signs, and negations in accordance.

### Inter-Annotator Agreement

The interannotator agreement was… se table

The f1-score was x for

Text

Description automatically generated

Higher scores were given when symptoms and findings

Text

Description automatically generated

Most common error was

Chart, treemap chart

Description automatically generated

### Entity characteristics

In the gold standard corpus x annotations were present. The number of instances of each entity class is shown in,

Would be interesting with entity distribution across headings – but we don’t really have time now

## Named Entity Recognition







NER performance

Precision recall F1

### Error Analysis

Confusion matrix, symp False positives are typically diseases…

# Discussion

In this project, I created an NLP pipeline for clinical text mining pipeline using spaCy. This enabled the use of a transformer-based general ML model from The National Library of Sweden in combination with a dictionary- and rule-based approach with relatively little configuration effort.

In the project we chose to try and build a fully functioning system rather than making only one small part with a higher degree of maturity. This follows the widely used principle of agile software development (https://agilemanifesto.org/principles.html) where you start with a minimal system and then have repeated cycles of continuous improvement in collaboration with potential end users.

The study has enlightened many improvement opportunities and pitfalls which will be discussed. Furthermore, some results require further investigation and troubleshooting although some possible explanations have been identified.

## Named Entity Recognition

### 

### Negation Recognition

The implemented negations recognition has high precision and recall and can be deployed for pre-annotation in its current state. Some performance improvements can be made by integrating the pseudo-negation dictionary containing phrases that appear as negations but are not such as “kan ej uteslutas”. Another improvement could be made by preprocessing of compound negations in the format of “o” joined with another word, such as “ostadig” [[3]](#footnote-12) into a negation and root word. This also helps to reduce granularity as the root entities can be compared instead of being two separate instances.

Standardization of model performance comparisons could help

## 

## Limitations

In this project, the evaluation of the performance of the pipeline was made qualitatively from visual comparison o to our gold standard corpus of 10 fabricated charts, which was annotated by a single annotator. These are important limitation of this that need to be improved in subsequent phases of the project.

Firstly, quantification off NLP-model performance should be implemented. This should compare the corresponding entities in the annotated corpus to the entity labels made by the NLP model. Alignment of tokens must be done and verified as so that the correct comparison can be made. Extraction of false and true positive, respective negative entity matches for each corresponding token should be done from which calculations of precision, recall and F1 score should be performed, see section 6.4. Moreover, the number of characters and tokens in every document should be quantified and the distribution of entities. I have already implemented and tested scripts for performing this task but there are currently difficulties exporting the annotated corpus from INCEpTION because of a bug in the export module of the program. Fixing this would be the first step to address the limitations of the project and the team behind INCEpTION has been contacted. Using these scripts, the annotated corpus was exported from INCEpTION in CoNLL-2003 format. After aligning the tokenization of these texts using spaCy alignment module, these true annotations were compared to the annotations produced by the NLP using the scripts as well as SpaCy evaluation module.

Secondly, evaluation of the gold-standard annotations of the corpus should be employed and annotations potentially improved. As a simple step intra-annotator agreement could be evaluated, which can give an indication of the difficulty of the corpus and the quality of the annotator. The use of several different annotators would be the next step. The corpus now only reflects the individual interpretation of one annotator, both making it error prone and susceptible to bias. Analysis and metric thresholds for different aspects of inter-/intra annotator agreement (IAA) should be defined as inclusion and exclusion criteria of annotations to be included in a corpus. It is also of great importance to consider the recruitment of annotators for an annotation task and state recruitment criteria. (23). Creating guidelines presents many challenges and a previously implemented guideline structures and creation methods should be reviewed qualitatively and quantitatively to determine best practices and implement workflows for structured guideline creation principles.

Thirdly, while a small corpus annotated is at least in part suitable for evaluation of the NLP pipeline, a much larger corpus is required for training new ML model components. The citizen science platform can also be used to collect such a larger annotated corpus that reflects a large variety of writing styles. Emergency medicine was chosen as chart form because all doctors work there as part of their education regardless of their later specialization and because emergency medicine covers a very large variety of patients. Each record is to be written by a single doctor who also labels symptoms and findings. Records could afterwards be annotated by a second annotator after importing the data to INCEpTION. To save some work for both human annotators pre-annotation with the improved dictionary based NER module could be implemented. If the use of a second annotator is not feasible due to the corpus size, the dictionary-based system could also be used to help identify poorly written and annotated records, which are expected to occur due to misunderstandings of guidelines and junk input.

Using fictional patient records limits the applicability to real documents. Models trained on fictional records should be thoroughly tested for applicability before application in real world scenarios. The advantage is open sharing without privacy concerns. Thus, using the citizen science platform, a variety of annotators could be annotating the corpus. This reduces the bias for the decisions of a specific annotator so that models are not overfitted to match specific persons.

### Misspellings sound be discussued

## Relation to Other Work

This study expands upon previous work for NLP for medical texts in Swedish done by the Clinical Text Mining Group of Stockholm University[[4]](#footnote-13). Contact with the group has been initialized. Their work has been done on the Stockholm EPR (Electronic Patient Record) Corpus which contains data from over 512 clinical units from Karolinska University Hospital encompassing the years 2006‒2014 and over two million patients that have been deidentified. However, the corpus cannot be shared outside of their local system due to ethical concerns. This would be the strength of our approach to create a corpus that could be freely used and expanded. Work by the group has ben done for automatic deidentification of patients which could be interesting for future project(21) (24).

There are other frameworks being developed than spaCy for the creation of NLP pipelines that meet most of the same qualifications. In particular Flair, which is developed the Humboldt University of Berlin could be worth comparing to spaCy them.(25)

# Conclusions and Future Work

The strategy of using terminologies for pre-annotation holds promise as it’s very low cost and principles could be generalized to include and process several terminologies. The risk is including false positives, lowering precision. In the function of pre-annotation is more important that the system has high precision than recall as it’s not meant to find all entities but to aid the annotator. If suggestions are incorrect it will bias the annotator and make automated suggestion and accepting workflows less practical.

This study functions as a base for the development of a powerful NLP pipeline and human-annotated gold corpora for Swedish clinical texts. these .A

Some pipeline components have sufficient maturity for immediate use whereas others require modification, for which more extensive data is needed. The NLP pipeline created can be used to pre-annotate large corpora and shorten manual annotation processes for creation of training data for fine-tuning machine learning models.

detailed description of the clinical population from which

symptom information was extracted and analyzed, open sharing of

user-developed symptom-related NLP algorithms or pipelines and

vocabularies, and the establishment of formal reporting standards

for investigations using NLP methodologies.

Declaration of own contribution

I started out by learning to use the INCEpTION tool which was chosen by my supervisor but had not been used in the group. I set up the annotation procedure in the group and developed a simple annotation pipeline. I also set up prototype annotation guidelines in conjunction with a simple tracker for logging of annotation problems and developing the guidelines. I made preliminary annotations and was also part of the discussions in which encountered annotation issues were resolved. I then expanded the guidelines in close collaboration with a team member who did the final annotations.

I performed almost all programming for this project. First, I set up the spaCy pipeline. This was done in a stepwise pattern of exploration with a small data sample using a Jupyter notebook running in Google Colaboratory. Programming languages I used were python3, bash and, for version control, git. After collecting the Swedish language models, I set up a stable version of a spaCy pipeline able to run both locally and on Google Colab. I then tested various modules and pipeline configurations.

I went on to pre-process the corpus and dictionaries and implemented interfaces to spaCy commands. For this, I created python scripts to process the ICD-10-SE code descriptions. I identified Swedish NegEx as a suitable project tool and then expanded its dictionary together with a collaborator from the project group and added it to the NLP pipeline. I also set up the scripts for evaluating the performance of the pipeline and for the visualization of the results.

Acknowledgements

I would like to thank my supervisors, project members and collaborators for all their help with project design, data creation, technical support, scientific discussions, and other input: Rafsan Ahmed, Salma Kazemi Rashed, Anton Werin, Anders Björkelund, Martin Gerdin Wärnberg, Johanna Berg, Sonja Aits. I also wanted to thank the AIR Lund project for making available their infrastructure for processing sensitive patient data even though it was not used in the end

References

1. Dalianis H. Clinical Text Mining: Secondary Use of Electronic Patient Records. 1st ed. 2018. Cham: Springer International Publishing : Imprint: Springer; 2018. 1 p.

2. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inform Assoc JAMIA. 2019 Apr 1;26(4):364–79.

3. Mehta N, Pandit A. Concurrence of big data analytics and healthcare: A systematic review. Int J Med Inf. 2018 Jun;114:57–65.

4. Allvin H, Carlsson E, Dalianis H, Danielsson-Ojala R, Daudaravičius V, Hassel M, et al. Characteristics of Finnish and Swedish intensive care nursing narratives: a comparative analysis to support the development of clinical language technologies. J Biomed Semant. 2011;2 Suppl 3:S1.

5. Kieuvongngam V, Tan B, Niu Y. Automatic Text Summarization of COVID-19 Medical Research Articles using BERT and GPT-2. ArXiv200601997 Cs [Internet]. 2020 Jun 2 [cited 2021 Jun 3]; Available from: http://arxiv.org/abs/2006.01997

6. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020 Feb 15;36(4):1234–40.

7. Skeppstedt M. Negation detection in Swedish clinical text: An adaption of NegEx to Swedish. J Biomed Semant. 2011 Jul 14;2(3):S3.

8. Skeppstedt M, Kvist M, Nilsson GH, Dalianis H. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. J Biomed Inform. 2014 Jun 1;49:148–58.

9. Almgren S, Pavlov S, Mogren O. Named Entity Recognition in Swedish Health Records with Character-Based Deep Bidirectional LSTMs. In: Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016) [Internet]. Osaka, Japan: The COLING 2016 Organizing Committee; 2016 [cited 2021 Aug 3]. p. 30–9. Available from: https://aclanthology.org/W16-5104

10. Cook HV, Jensen LJ. A Guide to Dictionary-Based Text Mining. Methods Mol Biol Clifton NJ. 2019;1939:73–89.

11. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010 Sep 1;17(5):507–13.

12. Kokkinakis D, Thurin A. Identification of Entity References in Hospital Discharge Letters. In: NODALIDA. 2007.

13. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007 Nov;25(11):1251–5.

14. Forbush TB, Gundlapalli AV, Palmer MN, Shen S, South BR, Divita G, et al. “Sitting on pins and needles”: characterization of symptom descriptions in clinical notes". AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci. 2013;2013:67–71.

15. Dalianis H. Pseudonymisation of Swedish Electronic Patient Records Using a Rule-Based Approach. 2019 Sep 30 [cited 2021 Jun 7]; Available from: https://ep.liu.se/en/conference-article.aspx?series=ecp&issue=166&Article\_No=3

16. Klassifikationen ICD-10 [Internet]. Socialstyrelsen. [cited 2021 Jun 1]. Available from: https://www.socialstyrelsen.se/utveckla-verksamhet/e-halsa/klassificering-och-koder/icd-10/

17. Classification of Diseases (ICD) [Internet]. [cited 2021 Jun 1]. Available from: https://www.who.int/standards/classifications/classification-of-diseases

18. Cameron S, Turtle-Song I. Learning to Write Case Notes Using the SOAP Format. J Couns Dev. 2002;80(3):286–92.

19. Klie J-C, Bugert M, Boullosa B, de Castilho RE, Gurevych I. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In: Proceedings of the 27th international conference on computational linguistics: System demonstrations [Internet]. Association for Computational Linguistics; 2018. p. 5–9. Available from: http://tubiblio.ulb.tu-darmstadt.de/106270/

20. SpaCy – ny svensk modell för storskalig textanalys [Internet]. [cited 2021 Jun 3]. Available from: https://www.kb.se/samverkan-och-utveckling/nytt-fran-kb/nyheter-samverkan-och-utveckling/2020-11-23-spacy---ny-svensk-modell-for-storskalig-textanalys.html

21. Berg H, Dalianis H. Augmenting a De-identification System for Swedish Clinical Text Using Open Resources and Deep Learning. In: Linköping Electronic Conference Proceedings [Internet]. 2019 [cited 2021 Jun 3]. p. 8–15. Available from: https://ep.liu.se/en/conference-article.aspx?series=ecp&issue=166&Article\_No=2

22. Engel Thomas C, Bjødstrup Jensen P, Werge T, Brunak S. Negation scope and spelling variation for text-mining of Danish electronic patient records. In: Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi) [Internet]. Gothenburg, Sweden: Association for Computational Linguistics; 2014 [cited 2021 Jun 3]. p. 64–8. Available from: http://aclweb.org/anthology/W14-1109

23. Wissler L, Almashraee M, Monett D, Paschke A. The Gold Standard in Corpus Annotation. In 2014.

24. Dalianis H, Boström H. Releasing a Swedish Clinical Corpus after Removing all Words – De-identification Experiments with Conditional Random Fields and Random Forests. In 2012 [cited 2021 Jun 3]. p. 45–8. Available from: http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-221475

25. Akbik A, Bergmann T, Blythe D, Rasul K, Schweter S, Vollgraf R. FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations) [Internet]. Minneapolis, Minnesota: Association for Computational Linguistics; 2019 [cited 2021 Jun 3]. p. 54–9. Available from: https://www.aclweb.org/anthology/N19-4010

# Appendix

|  |  |
| --- | --- |
| **class** | **definition** |
| symptom | Subjective sign of disease apparent to the patient |
| finding | Objective observation made by a physician and/or the result of a medical examination of the patient |
| negation | Term which inverts or denies the default meaning of a word or statement |

\* Developed with a resident in Emergency Medicine at Skåne University Hospital in Malmö

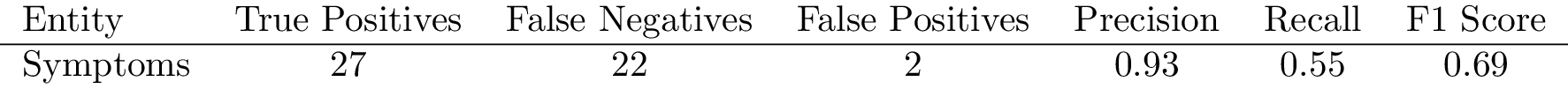
\* Definitions of symptoms and findings are based SNOMED CT Editorial Guide <https://confluence.ihtsdotools.org/display/DOCEG>

**Table 1**: Definition of the Named Entity Classes for Clinical Texts Annotated in the Gold-Standard Corpus

|  |  |
| --- | --- |
| XPOS tagger (accuracy): | 97.96 |
| UPOS tagger (accuracy): | 98.40 |
| Parser (UAS/LAS): | 93.51/90.74 |
| Sentencer (F score): | 94.73 |
| NER (F score): | 90.06 |

\* All scores were quotes from the group that provided the model and not verified by me

Table 2: Performance scores of transformer model [sv-pipeline-0.0.0] provided by the Swedish National Library https://github.com/Kungbib/swedish-spacy



One of the 10 fabricated charts from the gold standard corpus with annotations made by the NLP pipeline and evaluated manually. Visualization of detected entities came from the displacy module of spacy and evaluations were added manually. The number of true and false positive and false negative predictions was extracted and used to calculate the precision recall and F1 score.

Table 3: Preliminary pipeline performance for named entity recognition of symptom and findings

**Figure. 1** Visualization and evaluation of the annotations made by the NLP pipeline

One of the 10 fabricated charts from the gold standard corpus with annotations made by the NLP pipeline and evaluated manually. Visualization of detected entities (grey boxes) comes from the displacy module of spacy and evaluations were added manually. Boxes are false-negative NER labels, red crosses are false-positive labels and stars are correct labels. Letter combinations indicate NER class: SYM: symptom/finding, NEG: negation, TME: time-related expression, MSR: measurement



1. spaCy is an open-source library for advanced Natural Language Processing (NLP) in Python (https://spacy.io/) [↑](#footnote-ref-2)
2. <https://github.com/Kungbib/swedish-spacy> [↑](#footnote-ref-3)
3. Resolves to “unsteady” [↑](#footnote-ref-12)
4. https://dsv.su.se/en/research/research-areas/health/clintextgroup-1.113084 [↑](#footnote-ref-13)