Universiteit Leiden

Master Computer Science

The automatic labelling of medical entities in Dutch endoscopy reports

Name: Martin Koole
Student ID: s2383179
Date: 25/11/2021

Specialisation: Advanced Data Analytics

1st supervisor: Dr. Suzan Verberne
2nd supervisor: Prof.dr. Daniël Hommes

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

The automatic labelling of medical entities in Dutch endoscopy reports

Koole, M.
s2383179

November 25, 2021

Abstract

DEARhealth aims to improve healthcare experiences and outcomes by defining personalised Care Pathways for each individual patient. Suggestions to providers are given by a recommender, which takes into account multiple sources of information regarding the patient's health. In this paper, we focus on the extraction of information of a current unused source of information: Dutch endoscopy reports. We investigate AutoNER, a distant supervision model for which we define an extensive pipeline, incorporating the clinical database SnomedCT, to create the required dictionaries. We find that though AutoNER is able to label various terms, the extraction process is not yet reliable in its current state. Although future improvements could benefit AutoNER, we are not fully convinced that this particular approach is most effective for the extraction of medical entities from Dutch endoscopy reports.

Keywords: Information Extraction, Clinical Named Entity Recognition, Distant Supervision, AutoNER, DEARhealth, Endoscopy Reports

Acknowledgements

I would like to foremost thank my supervisor from Leiden University, Suzan Verberne, for continuous guidance, input and feedback throughout this research. Your insight and knowledge definitely helped me shape this project. A big thanks also goes to my DEAR supervisor, Daniël Hommes, who always showed genuine interest and enthusiasm, and often philosophised about how my research could be of further use for DEARhealth. I would also like to thank my weekly DEAR supervisors Caroline Ruiter and formerly Aisha Sie, for their advice and familiarising me with the DEARhealth environment. Thanks to Katinka de Korte, for introducing me to DEAR, and Vincent van Beek, for getting me into contact with the LUMC. A general shout-out goes to the DEARhealth team, for allowing me the opportunity to contribute to the development and improvement of patient healthcare.

Furthermore, I want to thank Andrea van der Meulen, for taking the time to gather the required patient data. My gratitude also goes to Feikje Hielkema and Sander Mertens, for showing me the workings of the Snowstorm API, and Cees Clemens, who one day happened to ask me if I would be interested in joining the test panel for the DEARhealth app, and starting this all.

Finally, I wish to thank my family for their never-ending support. This all would not have been possible without you.

Table of Contents

List of Figures

List of Tables

Introduction 1

DEARhealth is an organisation that addresses the treatment of chronic diseases by means of defining Care Pathways. Care Pathways "contain all medically required activities and associated support programs"[1]. In other words, these pathways indicate which specific type of care a patient requires at which point during the treatment. However, whereas ordinary treatment usually lasts until the patient has been sufficiently recovered, Care Pathways opt for continuous support for an extended, indefinite amount of time thereafter. The goal here is to prevent future complications from arising, by 'navigating' patients around health risks.

Naturally, patients differ from one another in terms of need regarding both physical and mental care. Thus, in order for a Care Pathway to achieve and maintain a patient's health, it should be defined and tailored specifically for each individual person. Doing this by means of manual labour is likely an unfeasible task: not only would this be highly time-consuming, but moreover, knowledge covering all relevant expertise is required.

DEARhealth pathways are therefore continuously evaluated by a recommender. Based on available medical background, current treatment plans and questionnaires directly taken from the patient via the corresponding mobile app, the recommender will suggest adjustments to the originally defined pathway. Whether these are implemented or not is decided by providers themselves. Over time, these non-static pathways will become more accurate and personalised for each individual patient. Figure 1 depicts a high-level relational overview of the involved parties, data, platforms and the recommender.

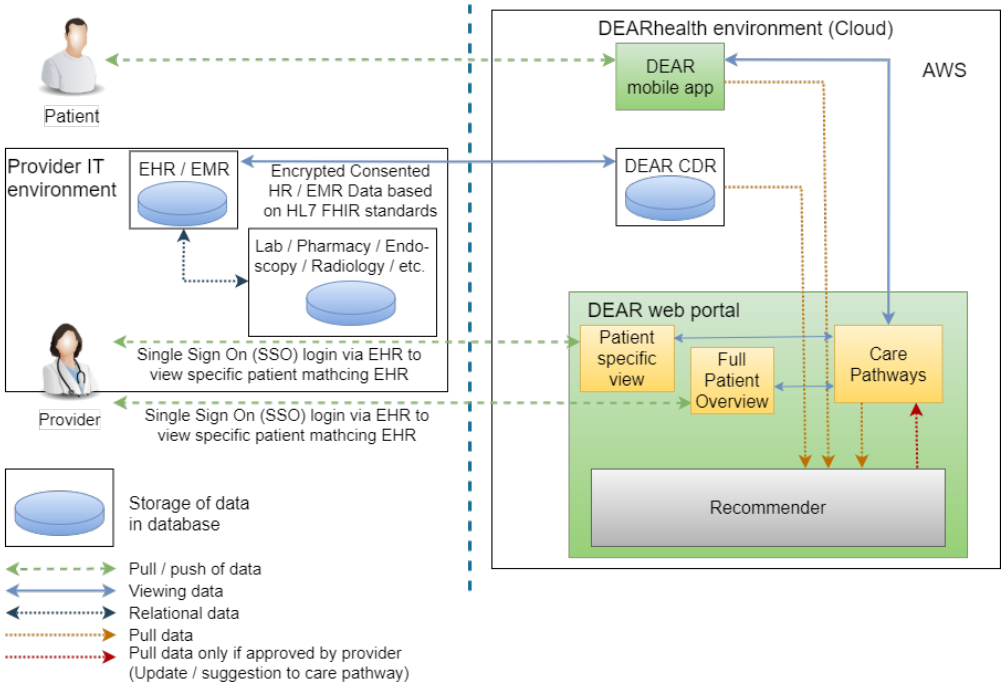


Figure 1: DEARhealth Care Pathways relational overview.

[1] https://dearhealth.com/how-it-works

A significant amount of information that could contribute to this learning and prediction process is captured within unstructured data. In the case of Inflammatory Bowel Diseases (IBD), this includes the free text fields found within endoscopy reports. Containing detailed descriptions of observations and conclusions made during an endoscopic procedure, endoscopy reports greatly reflect on the medical status of the intestines. Examples include the severity of an active illness, such as Crohn's Disease (CD) or Ulcerative Colitis (UC), the presence of fistulas, or the recovery from aforementioned or other ailments. Overall, endoscopy reports provide a lot of information regarding the development of IBD-related issues.

Currently, the recommender only allows for input filtered from texts following a small set of rules, consisting only of predefined formats. It is known that these rules do not cover a large quantity of token sets, and therefore the overall input coverage is limited. As such, DEARhealth is working towards a more robust approach. This implies the need for a newly defined method for the extraction of informative terms, as well as adjustments made to the recommender based on the more versatile resulting input.

In this research, we focus mainly on the first: information extraction from endoscopy reports, and, fitting DEARhealth's target group of Dutch patients, focus only on those written in Dutch. Furthermore, we aim to label extracted entities according to a set of categories, including at least the three main categories of *Problem* (diseases, symptoms and otherwise abnormal findings), *Treatment* (surgical procedures and medicinal substances) and *Test* (endoscopic procedures, blood tests, etc.). This way, we are able to represent terms in a structured, consistent manner, while still allowing for a greater variety of possible input compared to the aforementioned rule-based approach.

One of the main challenges however, is that we cannot rely on supervised learning methods, as the endoscopy reports are completely unlabelled. Doing so by means of manual effort would be highly time-consuming, and future addition of other data, which may include new terms, would require another iteration of labelling, lest these not appear in the training, development or test set, and would therefore not be susceptible to evaluation. Consequently, we investigate the effects of a model called AutoNER [41, 40], a knowledge-based, distantly supervised approach, which already yielded promising results on other clinical datasets. The knowledge-based aspect of AutoNER refers to its use of *dictionaries*, which contain terms including their type, providing a ground truth for the model.

In order to construct these dictionaries ourselves, we make use of a medical term database called SnomedCT[2]. An important property of SnomedCT is that it supports multiple languages, including English and Dutch. Therefore, we can use highly similar approaches for the construction of the dictionaries, regardless of lingual content. The execution, and thus reliability of our methods is thereby maintained across datasets, despite content differing on a lingual level.

Another implication resulting from the lack of any annotation is that we can hardly evaluate the performance of our methods using general evaluation metrics, such as f1-scores. Therefore, we define our methods on additional annotated clinical data. This also includes numerous reproductions to further validate the use of AutoNER. The specific use of clinical data ensures similar content to the endoscopy reports regarding medical terms, while the annotations introduce the possibility of evaluating our methods according to aforementioned evaluation metrics.

[2]https://www.snomed.org/

In summary, the main challenges of this project include: annotating a yet unlabelled dataset according to predefined categories, evaluating and improving the quality of these labels by using a distantly supervised model in the form of AutoNER, and tuning each individual step, opting for a reliable information extraction approach for in particular the Dutch endoscopy reports.

The main questions we will be focusing on are:

- What kind of information can we extract from endoscopy reports, with the general aim to improve the DEARhealth Care Pathways?
- Which methods deem the most reliable for this problem?

We make several contributions to this field of research. Foremost is applying the distant supervision model called AutoNER to Dutch biomedical texts. Second is the exploitation of Snowstorm's language independent search methods, so that our approach of constructing AutoNER's dictionaries can be applied to both English and Dutch corpora in the exact same manner. Overall, we find that the application of AutoNER is not as reliable and time efficient as the name of the model suggests.

To formulate well-supported answers to our research questions, this thesis follows a structured layout. In Chapter 2, we address relevant literature and discuss experiments done in highly similar fields of research. Next, in Chapter 3, we discuss the data we used, which includes the Dutch endoscopy reports, the clinical datasets of NCBI-Disease, BC5CDR and i2b2, the clinical term database SnomedCT and the tools we used to access the latter. We then give a thorough description of our main AutoNER pipeline setup and corresponding decision-making in Chapter 4, followed by extensive assessments, comparisons and evaluations of the obtained results in Chapter 5. In Chapter 6, we provide an additional discussion section. In Chapter 7, we draw conclusions based on the made observations and results, and formulate answers to our research questions. Finally, in Chapter 8, we propose several directions for follow-up research. From these, several are aimed at the improvement of the AutoNER model for in particular the Dutch endoscopy reports, and the remainder suggests additional tasks to help create more complete and reliable input for the DEARhealth recommender.

Background 2

Because the endoscopy reports do not yield any form of annotation, we need to investigate suitable alternatives to widely used supervised methods. As is often the case with patient data, it is due to unstructured documentation of procedures and privacy legislation rules that the topic of analysing these particular type of texts has yet to be addressed in further detail.

2.1 A brief summary on the demographics and effects of IBD

The term Inflammatory Bowel Disease generally refers to Crohn's Disease (CD) and Ulcerative Colitis (UC). Both diseases are chronic in nature, and are characterised by inflammations of the gastrointestinal tract. For UC, affected areas are continuous and uniform, found within the large intestine and rectum. For CD, affected areas tend to be more patchy, and may occur anywhere within the gastrointestinal tract.

Although IBD is a phenomenon found on a global scale, this type of anomaly is more prevalent among the Caucasian race, an observation already made in 1971 by Rogers *et al.* [37]. It was also found that IBD is more likely to occur among woman than men. To this day, this trend persists.

Additionally, a series of analyses conducted by Kappelman *et al.* [19] indicated that the number of IBD patients in the US has been increasing over time. This holds for both paediatric and adult patients, the latter for up the age of 30. For CD, it is also found that the prevalence has been increasing more for boys when compared to girls, and vice versa for adults. Although there is no such deviation between the sexes for UC, the overall prevalence has been increasing as well.

The exact cause of IBD has yet to be laid bare, but recent research suggests that, with CD in particular, this group of diseases is hereditary in nature. A series of conducted experiments by Baldassano *et al.* [4] indicates a relation between a specific gene (ATG16L1) and paediatric IBD.

Regardless, it is well known that IBD causes a wide variety of physical and psychological health problems. Especially children and adolescents are at an increased risk of developing extraintestional manifestations, which are affected regions of the body outside of the intestines. This in turn can lead to growth failure, delayed puberty and joint issues [21, 20]. IBD can also lead to intestinal scarring, strictures, and the development of ulcers and fistulas. These types of problems may regularly require surgery [11].

In turn, such persistent issues can lead to psychological disorders, affecting mental and emotional health and expression, and social behaviour [38]. Other frequently found problems in this regard include neuroticism [36, 33], perfectionism [16], and depression and anxiety [17].

Because aforementioned problems all may or not may not occur, a specific, personalised plan of treatment could be extremely beneficial. This could be a series of therapeutic measures based on affected regions of the body or resistances to particular medicine [21] (Table 2), the prescription of specific diet choices [12], and even the application of new forms of psychological therapy, which are specifically designed for those affected by IBD [46].

2.2 Big data in IBD

Over the past few decades, big data methods for IBD related data have become increasingly more important. Raw data is often generated in large amounts, and includes medical images, genomics, clinical trials, social media, electronic health records, administrative databases, e-Health applications, questionnaires - similarly to those found in the DEARhealth mobile app, and cohort studies [39].

Naturally, endoscopy reports are also a contributing factor, as is the set of images that explain and support statements made in these reports. For the latter, a very important task comes in the form of artefact detection [2].

However, regardless of format, and whether structured or unstructured, this data always has to be made suitable for big data platforms (i.e., Hadoop[3]) and analysis first, before it eventually becomes applicable to prescriptions, predictive risk models, etc. With the number of IBD affected patients consistently increasing, an important overall goal is to improve IBD care cost-effectiveness [32].

Hou *et al.* [18] for example addresses the issue of cost-effectiveness by showing the importance of distinguishing surveillance from non-surveillance colonoscopy procedures when it comes to defining a care treatment plan. Namely, without discriminating procedures, it is difficult to research surveillance practices and corresponding outcomes. This becomes especially significant when patients are more susceptible to developing life-threatening complications, such as colorectal cancer.

2.3 Natural Language Processing (NLP) techniques for clinical texts

Written clinical data is highly prone to contain names and descriptions of diseases, medicine, procedures and findings. Terms often are, like the overall content of such texts, domain specific. Because of this, performing any kind of task on clinical data comes with a variety of challenges.

Common among these, is obtaining structured outputs from unstructured clinical texts. Different types of clinical data naturally require specific NLP techniques, an issue addressed thoroughly by Kreimeyer *et al.* [23].

Recent research also focuses on the detection of elliptical coordinated compound noun phrases [6]. Sentences such as *brain and spine tumour* imply both *brain tumour* and *spine tumour*. However, most NLP systems lack the ability to detect these kind of noun phrases, which may lead to inaccurate classifications, such as labelling *brain* as *organ*, and *spine tumour* as *cancer* instead.

Capturing relevant contextual information of entities occurring in biomedical texts is often very difficult, and consequently, information passed into the deep layers of any neural model is prone to be incomplete. For a reliable biomedical NER task, this opts for the combination of widely known approaches, such as Conditional Random Fields (CRF) and Bi-directional Long Short-Term Memory (BiLSTM), including incorporated n-grams [10]. Also, due to a general lack of sufficiently annotated training data, including automatically processed syntactical information proves to be of great importance [44].

More advanced tasks such as the classification of relations among entities in clinical notes (medicines, medical terms, etc.) is another highly addressed issue. State of the art methods

[3] https://hadoop.apache.org/

are found in the form of Long Short-Term Memory (LSTM) networks [29], where also the conclusion is drawn that taking into account the word embeddings of medical domain terms enhances the performance of these networks. Using word embeddings for clinical NER tasks in general may also outperform CRF, as described by Wu *et al.* [50].

2.4 NLP techniques targeted at gastrointestinal reports

Aforementioned challenges such as the generation of structured data and entity recognition are also found in IBD related research, which naturally includes endoscopy reports. Covering such highly specific content naturally opts for newly defined methods, but little research has been done on this type of data.

The work of Zeki *et al.* [53] describes the application of EndoMineR [51], a package designed to generate quality metrics for a range of symptoms found in Barett's oesophagus. Other research by Zeki investigates the complexity of endoscopy reports using several of the programming language R's most popular readability-package scores: Flesch-Kincaid, Gunning-Fog Index and Coleman-Liau. It is found that the reports are at the 'language-level' of an early high school student, and therefore the author argues that extraction using rule-based methods is a great alternative compared to for example machine learning. Phrase removal does not have any significant impact on the scores [52].

In addition, even the construction of IBD databases is a recurring issue. Brown *et al.* [9] states the importance of doing so automatically, simply because of the sheer amount of available related data yet to be stored.

2.5 Endoscopy report quality

There are also numerous challenges regarding the quality of endoscopy reports. For general issues such as spelling errors, the Levenshtein distance can be used to determine the correct spelling of (important) terms. However, from content alone, it might be difficult to determine whether the report is about CD or UC, unless explicitly stated [9]. This is due to the displayed similarity of symptoms occurring in these diseases.

As such, studies like those by Kuipers *et al.* [24] and Bretthauer *et al.* [8] indicate the importance of standardised report systems for endoscopy reports. This includes the use of pre-defined text blocks, rather than free text fields, making the generation of the reports user-friendly, maintaining clarity and providing easy access to captured information.

Furthermore, the European Crohn's and Colitis Organisation (ECCO) suggests a consensus regarding the indication and application of endoscopic procedures, in an effort to enhance efficiency and consistency [3]. This is done by categorising the procedures according to four main topics: diagnosis and follow-up, score of endoscopic activity, small bowel endoscopy, and surveillance. Also, ECCO provides a terminology of endoscopic lesions in IBD, which includes the agreed terms for different types of mucosal damage, how these should be described, and which grading scale should be applied for each individual type [3] (Table 2.1).

Another factor that may improve the quality of endoscopy reports comes in the form of simple audit interventions. In a study on Spanish reports performed by Lisboa-Gonçalves *et al.* [27], it was found that in nearly half of these, the descriptions of observations and conclusions regarding the state of the gastrointestinal tract were incomplete, and no further support to these claims was provided. The percentage of incorrectly described lesions was

found to be just over 10%.

The lack of following a general consensus affects the quality of endoscopy reports on a global scale. Benchimol *et al.* [5] for example compared several reporting systems from multiple countries regarding the development of IBD in children in both developed in lesser developed countries. However, accurate rates are often missing and cannot be easily derived from current unstructured texts. It is argued that a more thorough investigation of these numbers might lead to a better understanding of paediatric IBD incidence rates, both environmental and genetic.

The implementation of aforementioned suggestions could enhance the quality of endoscopy reports, in turn making it easier to automatically analyse texts, train models and extract specific information.

2.6 Annotating unlabelled clinical data

The automatic annotation of unlabelled clinical data heavily depends on the presence or absence of labelled data. When labelled training and evaluation sets are present, and these labels are accurate, supervised approaches can be used to annotate additional, similar data.

One state-of-the-art supervised method for the annotation of clinical data comes in the form of BioBERT, a Bidirectional Encoder Representations from Transformers (BERT)-based [13] model pre-trained on biomedical texts. Lee *et al.* [26] show that the BioBERT outperforms numerous other methods on clinical data. This includes the i2B2 VA challenge dataset, which we will also use to evaluate our approach.

BioBERT is also suitable for transfer learning for biomedical NER [43]. Transfer learning implies the training of a model to for example detect types of cars, but use it for trucks instead, i.e., different, yet still (remotely) similar data. Our approach relies on this principle as well. Since the endoscopy reports are unlabelled, we have to resort to evaluating our methods on other clinical data in order to compute any evaluation metric scores automatically. Symeonidou *et al.* [43] show that with only a relatively small amount of annotated data, transfer learning can help in specialised information extraction tasks.

However, annotated training or other sets for (raw) patient data are extremely scarce, mainly due to privacy legislation rules. Therefore, it is often impossible to use supervised methods unless manual annotation is performed on a rather large scale. This opted for other approaches, including rule-based methods combined with deep representation [48], distant supervision methods to augment data by using only a small pool of manually annotated data [42], weakly supervised methods for the creation of vast training data [35], and unsupervised methods that use clinical term databases such as SnomedCT [30].

2.6.1 AutoNER

As mentioned in the introduction, we will primarily use a model called AutoNER to perform the information extraction task. Proposed by Shang *et al.* [41, 40], AutoNER is a knowledge-based, distant-supervision approach, allowing for NER tasks on unlabelled data. The neural model uses two dictionaries which provide a ground truth. The *core dictionary* contains entities labelled according to a set of categories. The *full dictionary* contains all terms that AutoNER should regard as candidate entities. Besides the terms captured within the *core dictionary*, the *full dictionary* additionally holds all terms that thus should be considered

relevant.

For the provided BC5CDR dataset, the *core dictionary* is the result of combining the MeSH database[4] and the CTD Chemical and Disease entity lists[5]. Here, entities are either labelled as *Chemical* or *Disease*. The *full dict* has been further extended upon by entities extracted from the texts by the AutoPhrase[6] module [40, 28].

AutoNER distinguishes itself by using a so called Tie or Break tagging scheme to further improve upon the distant supervision. Here, the focus lies on adjacent token pairs. If both tokens are of the same entity, the token span is considered a *Tie*. If either one or both of the two tokens occur only within the *full dictionary*, i.e., occurrence among the unknown high-quality phrases, the token span is labelled as *Unknown*. Tokens spans are always separated from one another by a *Break*. According to Shang *et al.* [41, 40], the use of this specific tagging scheme better exploits dictionary knowledge.

[4] https://www.nlm.nih.gov/mesh/download_mesh.html
[5] http://ctdbase.org/downloads/
[6] https://github.com/shangjingbo1226/AutoPhrase

Data 3

In this chapter, we present and elaborate on the used data. We start by analysing the Dutch endoscopy reports and the additional English clinical datasets, which include the NCBI-Disease, the i2b2 and the BC5CDR dataset. For each of the latter, we also state during which part of the research they were considered. Next, we discuss the clinical term database SnomedCT, as well as the tools we investigated to extract relevant concepts. Finally, we state which pre-trained word embeddings we used for the AutoNER model.

3.1 Endoscopy reports

The Dutch endoscopy reports we will be using are provided by the department of Gastroenterology at the Leiden University Medical Centre hospital (LUMC), totalling 1322 XML files. We anonymised these documents manually, removing all patient names, IDs and any other sensitive values that may relate to the person concerned. As mentioned, as these documents are raw patient data, they are completely unlabelled.

The majority of the texts are structured according to the 'new' format, adopted in 2017. Compared to the old layout, these reports provide more containers for storing specific data, such as descriptions for findings made during the endoscopic procedure. Therefore, more recent reports tend to have larger content by default, even if XML fields are left empty.

Furthermore, it is important to note that either format mainly consists of free text fields, with the exception of a few default choices for certain containers. Because no further check on the quality of these texts has been done, these texts are prone to spelling and other syntactical errors. After manual inspection, it becomes clear that this is indeed the case for many included documents. Some examples regarding this issue are shown in Table 1. Other examples include inconsistent use of capitalisation and words missing from sentences.

Term	Type of issue	Corrected	Translation
ontstekinh	Spelling	ontsteking	inflammation
rectunpolipeje	Spelling	rectumpoliepje	small rectal polyp
wsch	Unofficial abbreviation	waarschijnlijk	probably or likely
kwetsbaarslijmvlies	Incorrect concatenation	kwetsbaar slijmvlies	fragile mucosa
lijkt de nauw	Incorrect word usage	lijkt te nauw	seems too narrow
mn	Incomplete abbreviation	m.n. (met name)	particularly or mainly

Table 1: Examples of (reoccurring) textual issues found within the endoscopy reports.

3.2 Clinical NER datasets

Because the endoscopy reports lack any annotation, we require labelled data in order to evaluate our methods by means of a development and test set. Given perfect circumstances, the contents of such additional data will be highly similar to the endoscopy reports, and are written in the same language. However, such data cannot not be obtained easily, and we therefore sought sets which are freely distributed, are not violating any rules regarding patient privacy, and contain clinical data as to approximate the data contained in the endoscopy reports.

3.2.1 NCBI-disease dataset

The NCBI-disease corpus [14] consists of 793 fully annotated PubMed[7] abstracts. It was constructed specifically for the extraction of biomedical concepts, therefore providing a gold-standard for tasks such as clinical NER.

For our research, we will make use of the version provided by the BioBERT model, which already has all files converted to the required TSV format used by BioBERT. The annotation however only indicates entities themselves, and not any additional typing such as the group to which a disease is related, whether a term is a symptom referring to a disease, etc. Furthermore, BioBERT itself has to be altered quite extensively in order for it to allow the inclusion of typing. We therefore only use this dataset to reproduce the BioBERT results as stated in its corresponding paper by Lee *et al.* [26].

3.2.2 2010 i2b2/VA challenge

The 2010 i2b2/VA Challenge was a workshop composed of three tasks: medical concepts extraction, assertion classification, and relation classification [45]. Partners Healthcare, Beth Israel Deaconess Medical Center, and the University of Pittsburgh Medical Center were the instances providing the used anonymised patient reports. From these, 394 files are suitable for training, 477 for testing, and the remainder of 877 files was left unannotated. However, these numbers also include the files specifically annotated for the assertion classification and relation classification tasks, which we do not need. Omitting these leaves us with 170 files for training, and 256 for testing.

Instead of listing each word of the text with its corresponding entity type, the i2b2 dataset provides per text file a separate annotation file, in which annotations are stated as follows: entity, coordinates, entity type. Coordinates state the sentence number, followed by the index of the word itself - on token level. Entities can be labelled according to one of the following three categories: *Problem*, *Treatment* and *Test*. These are the categories we will also be using for the endoscopy reports.

This dataset is also used in the work of Lee *et al.* [26] to evaluate BioBERT, though again without the use of any of the aforementioned categories. This is the first benchmark dataset we will be using to define our methods on.

3.2.3 BioCreative V Chemical Disease Relation (BC5CDR) challenge

The final additional dataset we will be using is the BioCreative V Chemical Disease Relation (BC5CDR) corpus [49], which is a benchmark dataset generated for the extraction of relationships between disease and chemical entities. Consisting of 1500 human annotated PubMed articles, which include 4409 annotated chemicals, 5818 diseases and 3116 chemical-disease interactions, it covers a large variety of disease chemical interaction descriptions.

This dataset was used for evaluating the AutoNER model in its original paper, and the required dictionaries, as well as the texts, which are formatted to support the specific format AutoNER uses, are freely available. The BC5CDR dataset will function as our second benchmark dataset.

[7] https://pubmed.ncbi.nlm.nih.gov/

3.3 Snomed Clinical Terms

Snomed Clinical Terms (SnomedCT) is a collection of clinical terms, corresponding codes, synonyms, and detailed descriptions. For each concept, there is the Fully Specified Name (FSN) field, a unique string in SnomedCT, which also includes the assigned SnomedCT category between parenthesis. There is also the Preferred Term (PT), used to represent the concept within clinical applications or environments. If applicable, a (series of) synonym(s) may be provided for a particular concept.

SnomedCT also yields relationships between concepts. For example the term *brain tumour*, labelled a *disorder* in SnomedCT, has several child concepts like *Benign tumor of sella turcica* and *Chiasmal glioma*, both of which are also labelled *disorder*. The term also has two parent concepts: *Mass of intracranial structure*, which is a *finding* and *neoplasm of head*, a *disorder*. SnomedCT also holds information of the *finding site*, i.e., the place where the concept is located within the body – here, *intracranial structure*, and the *associated morphology*, which is *Neoplasm*.

From this example, we get a view of SnomedCT's multi-hierarchical structure: there are multiple child and parent concepts, as well as connections to other related concepts. Taking into account that the January 2020 release of SnomedCT's International version contains over 350.000 concepts, it is clear that the hierarchy is very complex.

Besides the high coverage of medical concepts, SnomedCT also supports multiple languages, including Dutch. Every term in SnomedCT has a direct mapping to other available lingual variants. This allows us to query SnomedCT in English – which we will do for our benchmark datasets – and use the same query to obtain the resulting concepts in Dutch, without any additional effort to translate these. Therefore, our SnomedCT search methods are language independent.

3.3.1 Accessing SnomedCT with PyMedTermino

In order to deal with SnomedCT's complexity, several tools have been defined that allow for efficient querying. We addressed two of these, the first being the PyMedTermino package, an open source Python implementation of the API as presented by Lamy *et al.* [25]. PyMedTermino makes it possible to exploit the SnomedCT hierarchy, search for concepts, retrieve the parent and child concepts thereof, etc. Although PyMedTermino does prove its ability to accurately extract concepts, there are some noticeable drawbacks.

For PyMedTermino to be used, one has to separately download and assign a release of SnomedCT database. The latter is prone to change over time (if only minimal), due to the introduction of new or removal of outdated (inactive) concepts, and other structural changes. In order for these changes to be passed into PyMedTermino, a new version of SnomedCT has to be initialised. The required initialisation also implies that whenever the user wishes to use a different language, a corresponding version of SnomedCT has to be loaded.

If search methods are already well defined, and the initialisation only needs to be done a few times, the swapping remains manageable. However, since we are continuously seeking to improve our own search method, this aspect of PyMedTermino becomes rather impracticable.

3.3.2 Accessing SnomedCT with Snowstorm

The second tool we use is the Snowstorm Terminology Server[8] API, developed by Snomed International[9]. Similar to PyMedTermino, it allows for concept search, parent and child concept retrieval, access to synonyms, etc. In addition, Snowstorm enables further exploitation of the multi-hierarchy, including a greater variety of search options, filters, and parameter settings. Moreover, there is the option of traversing the data using SnomedCT Expression Constraint Language (ECL) queries. ECL is a specifically constructed language for "defining bounded sets of clinical meanings represented by either precoordinated or postcoordinated expressions"[10]. ECL for example includes disjunctive and conjunctive operators, constraint operators like is-member-of, is-ancestor-of and is-descendant-of, and numerous other options [31]. The Snowstorm interface itself makes it clear which options are available, and which values these can take.

In order to efficiently work with Snowstorm, we made crawlers for two of its search modules: *Descriptions :* /*branch*/*descriptions* and *Concepts:* /*branch*/*concepts*. The first enables us to directly lookup concept strings, and retrieve their FSN and PT terms, as well as other (inactive) synonyms. The second allows us to exceed the maximum response limit of 10,000 terms by using the *SearchAfter* parameter. Such a high amount of resulting concepts occurs when we for example use the query: *« 64572001 |disorder (disorder)|*. This subsumes all concepts within the category *disorder* - one of SnomedCT's main categories - totalling a number of 81,973.

In contrast with PyMedTermino, changing between lingual versions of SnomedCT here only requires the alteration of a single parameter. And, since the terminology is continuously updated, the relevance of all obtained results is preserved.

In conclusion, although PyMedTermino does offer quite a large variety of operations, Snowstorm proves to be more extensive and reliable, and its ECL queries are language independent. Except for our baseline method, we will therefore continue with Snowstorm as our main tool for accessing SnomedCT.

3.4 Pre-trained word embeddings

In order for AutoNER to properly train the model, a pre-trained word embedding file is required. For the i2b2 and BC5CDR datasets, we use the embeddings resulting from the work of Pyysalo *et al.* [34]. This embedding file is automatically downloaded when initially running AutoNER. The 200 dimensional word vectors are the result from applying a Word2vec model to an extremely large quantity of Pubmed article titles, abstracts and full text documents.

For the endoscopy reports, we require a separate, Dutch embedding file - as embeddings are language dependant. Here, we use an embedding from the NLPL word embeddings repository[11]. The 100 dimensional word vectors are the result from applying a word2vec model to the Dutch CoNLL17 corpus (id 32) [15].

[8] https://snowstorm.test-nictiz.nl/swagger-ui.html
[9] https://www.snomed.org/
[10] https://confluence.ihtsdotools.org/display/slpg/snomed+ct+expression+constraint+language
[11] http://vectors.nlpl.eu/repository/

Methods 4

In this chapter, we thoroughly discuss and explain each specific step of the overall methodology. Included is the determination of upper bound scores using the state-of-the-art clinical NER model BioBERT, a naive baseline NER approach using SnomedCT, an intermediate evaluation of the thus far obtained results to better define our pipeline, and the application of AutoNER, as to perform the distant supervised NER task on the BC5CDR-Disease, i2b2 and endoscopy report datasets.

4.1 Method reproduction of a state-of-the-art supervised NER model: BioBERT

Prior to setting up a pipeline for our distant supervised model, we attempt to reproduce the results with BioBERT reported by Lee *et al.* [26] on three datasets: NCBI-Disease, BC5CDR-Disease, and i2b2. It is expected that BioBERT, given an annotated dataset, will outperform any distantly supervised approaches on the same data. However, the scores obtained with BioBERT will grant us an insight in what can be regarded as upper bound, and we will therefore strive to get as close as possible to these numbers.

In contrast to the NCBI-Disease and BC5CDR datasets, the i2b2 dataset is not freely distributed with the BioBERT model, and has to be requested via the corresponding 2010 i2b2/VA challenge website[12]. Annotations are provided in separate files, rather than being placed directly after each entity of the text. Moreover, annotations are denoted in a coordinate-like format. Here, entities are referred to merely by two indices, which correspond to the sentence and the entity's position therein, followed by the entity type.

Because of this, we first have to apply the necessary pre-processing steps to obtain the format required by BioBERT, linking each entity type to its corresponding entity using aforementioned indices. We also introduce IOB-tagging to ensure entities consisting of multiple consecutive tokens will still be regarded as a single entity by BioBERT.

Hereafter, we apply BioBERT to each of the three datasets, following the procedure as stated by Lee *et al.* [26]. It has to be noted that the evaluation of BioBERT is limited to the mere recognition of entities, disregarding any typing. The NCBI-Disease and BC5CDR-Disease datasets cover, as the names suggests, only entities of the type *Disease*. The NER task therefore does not impose the need for additional type annotation.

However, we opt to take typing into account for our distantly supervised approach. For i2b2, we therefore also investigate BioBERT's performance if IOB-annotations for all of it's three types, *Problem*, *Treatment* and *Test*, are used simultaneously, rather than separately.

[12] https://www.i2b2.org/NLP/DataSets/Main.php

	Precision	Recall	F1-score
NCBI-Disease	0.867	0.891	0.879
i2b2 (avg)	0.874	0.879	0.877
i2b2 Problem	0.692	0.697	0.694
i2b2 Treatment	0.871	0.843	0.857
i2b2 Test	0.871	0.865	0.868
BC5CDR-Disease	0.852	0.868	0.860

Table 2: Precision, recall and f1-scores of BioBERT on the NCBI-Disease, i2b2 and BC5CDR-Disease datasets. The results on the NCBI-Disease and i2b2 (avg) are reproductions, and are near equal to those stated by Lee *et al.* [26]. It has to be noted that the NER task on all three categories of the i2b2 dataset simultaneously, as stated in the i2b2 (avg) row, does not include the type labelling of these entities.

The values as presented in Table 2 will be regarded as the upper bound obtainable scores. We therefore aim to get as close as possible to these scores using distant supervised methods, but consequently also do not expect that these values will be exceeded. Surprisingly, we find that when using BioBERT on solely the *Problem* category, the model performs substantially worse compared to on all three categories simultaneously, as well as the other two categories of *Treatment* and *Test*. This may indicate that NER task on the *Problem* category of the i2b2 dataset is relatively more difficult, and we will take this observation into account when using distant supervision.

4.2 Naive dictionary-based NER using SnomedCT

Our first attempt investigates the degree of ease of mapping extracted SnomedCT terms to the three categories of i2b2 dataset, *Problem*, *Treatment* and *Test*.

For the baseline method, we use a naive dictionary-based approach for entity recognition. We determine for each term we extracted from SnomedCT where it occurs within the text, and label accordingly.

Initially, we make use of the PyMedTermino in order to traverse and exploit the SnomedCT hierarchy. As SnomedCT's eight main categories do not directly correspond with those of i2b2, e.g., there is no *Problem* category in SnomedCT, we selected terms in various ways, and map these to the required three entity types instead. We investigated three approaches we considered to be worth examining:

1. We search for *Problem*, *Treatment* and *Test* in SnomedCT, and extract all the descendants of the resulting terms, followed by their synonyms.

2. We select SnomedCT terms which hold descriptors – the term found in between parenthesis trailing the actual term – most likely to be matching the required format. For *Problem*, we select all terms within the *disorder* branch of the *clinical finding* main category. For *Treatment*, we select all terms within the *regime / therapy* branch of the *procedure* main category. Finally, for *Test*, we select all terms within the *procedure* branch of the eponymous *procedure* main category, and also add all terms found in the *observable entity* category. Similarly to the previous approach, we extract all descendants and synonyms.

3. We select terms for each of the three categories according to the given examples in the i2b2 challenge description. Using this approach, we end up with a larger variety of search terms. For *Problem*, we query *disease*, *syndrome*, *sign*, *symptom*, *mental status*, *behavioural status*, *virus*, *bacterium*, *injury*, *abnormality* and *test result*. For *Treatment*, we search for *medication*, *biological substance*, *drug delivery device*, *procedure*, *device*, *hardware* and *Treatment*. And finally, for *Test*, we query *Test*, *body fluid analysis*, *physiologic measure*, *vital sign*, *examination* and *evaluation*.

For in particular approaches one and three, we expect to find overlap among the descendants of terms, as they do not map to distinctive categories. Therefore, we may end up with terms for example labelled as both *Problem* and *Treatment*, which in turn leads to erroneous labelling.

4.3 Evaluating the baseline method: a mapping problem

When we put aforementioned approaches into practice, we find the evaluation metric scores as shown in Table 3. It is clear that our baseline methods are indeed rather naive approaches, and that there is a large possible improvement to be made when we draw a comparison to the scores obtained by BioBERT.

Approach	**Precision**	**Recall**	**F1**
1	0.252	0.039	0.064
2	0.106	0.080	0.086
3	0.155	0.141	0.145

Table 3: Precision, recall and f1-scores resulting from the three baseline approaches

The mapping of SnomedCT terms onto the three categories of the i2b2 dataset will also be the initial step for future pipelines, with in particular the dictionary construction for AutoNER. Naturally, if we are to build and continuously improve upon a certain approach, it is important to provide a solid foundation. This implies that the number of false positives (terms erroneously assigned a specific category) and false negatives (words incorrectly unlabelled by specific category) for each of the three categories, must be reduced as far as possible.

The core issue of this challenge lies in the complexity of the SnomedCT database. As already stated, terms can have a multitude of both parental and child relationships to other terms, allowing for multi-hierarchical clusters. Recalling our example from previous chapter: *brain tumour*, labelled a *disorder* in SnomedCT, has several child concepts like *Benign tumor of sella turcica* and *Chiasmal glioma*, both of which are also labelled *disorder*. The term also has two parent concepts: *Mass of intracranial structure*, which is a *finding* and *neoplasm of head*, a *disorder*. SnomedCT also holds information of the *finding site*, i.e., the place where the concept is located within the body – here, *intracranial structure*, and the *associated morphology*, which is *Neoplasm*.

We can see here that if we extract all relations and assign the same category, which in this case would be *Problem*, the term *intracranial structure* would also be assigned this label. However, *intracranial structure* does not refer to a *Problem* at all, but merely a location of the body. When this method is executed on a larger scale, we automatically end up with a lot of erroneous labels.

We also found that the same issues arise when we approach the mapping of labels from the opposite perspective. Rather than mapping SnomedCT categories onto those of i2b2, we try to map the categories of i2b2 onto those of SnomedCT. In order to extract noun phrases, we use an approach by Bowe [7] that combinations NLTK's part-of-speech tagger with the regular expressions reported in the work of Kim *et al.* [22] (Table 2). Next, we use SnomedCT as a 'filter': if querying a noun phrase in PyMedTermino does not return any results, then it most likely does not contain any clinical terms. By omitting phrases for which this holds true, we end up with a set composed of clinical entities only.

It would be ideal if all terms end up in three different categories, or an otherwise small number thereof, which would sufficiently represent *Problem*, *Treatment* and *Test*. However, we instead end up with many different (sub-)categories. One could argue that backtracking would be an option, by iterating through parental relationships until a main, or at least a high-level category is found, as to simplify the mapping problem. This however did not prove to be a reliable method either. As mentioned earlier, terms can have multiple parental relationships, and each respective ancestor can have its corresponding multitude of parental relationships, etc. The amount of manual decision-making would be an infeasible task for texts which may hold thousands of different medical entities – as is the case for i2b2. Consequently, no further improvement is made on the results as denoted in Table 3.

4.4 Knowledge-based, distantly supervised Named Entity Recognition: AutoNER

The significance of the mapping problem is reflected greatly by the evaluation metric scores resulting from all three baseline approaches, and we therefore deem the direct mapping of categories to be highly impracticable. We continue instead with AutoNER, a knowledge-based, distantly supervised approach which has proven itself a reliable alternative to supervised methods for clinical NER.

Ultimately, we want to use AutoNER for the Dutch endoscopy reports. Before we are able to do so however, we need to define a reliable method to construct the two dictionaries on which the model relies. As mentioned in Section 2.6.1, the provided dictionaries for the BC5CDR are based on the MeSH database[13] and the CTD Chemical and Disease entity lists[14], and the *full dictionary* is further expanded upon by using AutoPhrase. However, these databases are not available in Dutch. Moreover, it is not completely clear how the provided dictionaries are tailored. Because of this, we define our own pipeline to make AutoNER applicable to both Dutch and English corpora.

The content of these dictionaries is of similar format to the output of our baseline methods: lists of (candidate) entities that should be labelled according to any present category. However, this means that we again encounter the mapping problem when we incorporate the use of SnomedCT. The provided BC5CDR data already includes these dictionaries, so we do have an idea of what well-defined dictionaries should resemble. The question that arises here however is, to what extend can we actually use SnomedCT to reconstruct these dictionaries, as well as define new ones for other data such as the endoscopy reports?

In order to implement suitable pipelines, we performing the following tasks:

[13] https://www.nlm.nih.gov/mesh/download_mesh.html

[14] http://ctdbase.org/downloads/

- Validate AutoNER by reproducing the results as stated by Shang *et al.* [41] on the BC5CDR dataset using the provided dictionaries.
- Investigate how AutoNER performs if we use the entities resulting from our current baseline methods as dictionary content for the i2b2 dataset.
- Reproduce the results on the BC5CDR dataset, this time by defining our own dictionaries. Here, we disregard the *Chemical* category, and focus solely on the *Disease* entities. Compared to i2b2, where we have three categories, the NER task is thus narrowed down. This takes away a great part of the mapping complexity, and therefore allows us to better understand how we are to approach this problem.
- Return to the i2b2 dataset, and determine if we can successfully apply any new strategy that allows for dictionary construction for all three categories.
- Evaluate AutoNER on the Dutch endoscopy reports, taking into account the results and approaches on the two benchmark datasets.

The main reoccurring aspect of these steps is the construction of the dictionaries. During this research, we have investigated a lot of different approaches and search methods within SnomedCT. Eventually however, we decided upon a general method for dictionary construction, which we applied to all three datasets.

4.4.1 Extensive querying in SnomedCT

Although it is evident the mapping problem will remain present, we refine our baseline search methods in order to extract as many relevant terms as accurately as possible. However, PyMedTermino's search options are rather limited. We will therefore make use of the Snowstorm API to browse SnomedCT instead. As mentioned in Section 3.3.2, Snowstorm allows for the use of specified ECL queries, and fully supports both English and Dutch, so our queries and results will be language-independent.

In order to circumvent Snowstorm's *Descriptions : /branch/descriptions* search method's maximum number of returned results, which equals 10,000, we make use of the SearchAfter method in the *Concepts: /branch/concepts* module. We set the *limit* parameter to 1,000, and leave all other parameter settings to their default values. For all resulting main terms, we also collect the *fully specified name* (fsn) and the *preferred term* (pt).

For BC5CDR, we will only attempt to perform NER for the *Disease* category. We use the following ECL queries:
« 404684003 |Clinical finding (clinical finding)|
« 64572001 |Disorder (disorder)|
« 49755003 |Morphologic abnormality (morphologic abnormality)|

Due to an unknown reason, we barely retrieved any concepts related to types of cancer. We therefore additionally searched for *cancer* in the *Descriptions : /branch/descriptions* branch – as the number of terms for this particular search request does not exceed the corresponding maximum amount of 10,000.

For i2b2, we will define queries for all three categories, while ensuring as little overlap between results from different categories as possible. Based on the data, we find that the

Problem category is highly similar to BC5CDR's *Disease* category, and we therefore use the same ECL queries.

For *Problem*, we use:
« 404684003 |Clinical finding (clinical finding)|
« 64572001 |Disorder (disorder)|
« 49755003 |Morphologic abnormality (morphologic abnormality)|

Again, we additionally search for *cancer* in the *Descriptions : /branch/descriptions* branch to complete the entity list for this category.

For *Treatment*, we use:
« 410942007 | Drug or medicament (substance) |
« 373873005 | Pharmaceutical / biologic product (product) |
« 705208008 | Complementary therapy device (physical object) |
« 706036000 | Device for body fluid and tissue management (physical object) |
« 707727004 | Device substance (physical object) |
« 707728009 | Device system (physical object) |
« 705288005 | Ear, nose and throat device (physical object) |
« 705178002 | Physical therapy device (physical object) |
« 243120004 | Regimes and therapies (regime/therapy) |
« 277132007 | Therapeutic procedure (procedure) |
« 128303001 | Surgical removal (procedure) |

And finally for *Test*, we use:
« 86273004 | Biopsy (procedure) |
« 386053000 | Evaluation procedure (procedure) |
« 108252007 | Laboratory procedure (procedure) |
« 363788007 | Clinical history/examination observable (observable entity) |
« 103693007 | Diagnostic procedure (procedure) |

For the Dutch endoscopy reports, we use the exact same queries for the i2b2 dataset, again obtaining term lists for *Problem*, *Treatment* and *Test*. Naturally, we now retrieve the Dutch variant of each term.

After querying, we extract any synonyms for each term captured within the three lists – one list for each dataset. We have to perform this task in an additional matter because inactive concepts are not included within the results of the aforementioned queries. In order to retrieve these additional terms, we search for each term in the *Descriptions : /branch/descriptions* branch. We do have to introduce the condition that resulting terms *fsn* and *pt* fields may only be taken into account if the searched term and the retrieved *main* term are exact matches. Otherwise, we would include matches in which the searched term is a sub-string. For example, if search for *Colon cancer* without this measure, we would find *Colon cancer screening declined* (and its *pt* and *fsn* entries), which holds no value for the *Problem* category.

To expand the lists further, we add the pluralised format of each present term – if appli-

cable – using the Python *pattern*[15] package.

To finalise our initial lists of terms, we check which terms occur in the full corpus of their respective dataset (disregarding all capitalisation), and only keep those. At this point, the only difference between the *core* and *full dictionaries* are the added annotations for the former.

4.4.2 Additional sources

There is also a (large) variety of abbreviations within each corpus, many of which are relevant. SnomedCT however does not hold a sufficient amount thereof, and consequently the use of the aforementioned queries barely resulted into any. We therefore extracted all abbreviations from each corpus by using regular expressions, matching only fully capitalised tokens with a length >= 2 and =< 4.

The issue that arises however is that we cannot annotate the abbreviations the same way we did with the terms from SnomedCT, for which we relied on the mapping of categories. Instead, make use of lists of clinical abbreviations from external sources. We then annotate these manually according to the description given for each abbreviation.

For the BC5CDR and i2b2 datasets, we use a list of medical abbreviations provided by ResourcePharm[16]. For the Dutch endoscopy reports, we use the list of abbreviations contained in a Dutch IBD guideline document, called *Diagnostiek en behandeling van inflammatoire darmziekten bij kinderen* (*Diagnostics and treatment of inflammatory bowel diseases in children*) [1].

Thereafter, we check for each abbreviation we extracted from the corpus if it occurs within the annotated list. If so, we also add the abbreviation including its annotation to the *core dictionary*. Otherwise, the entity type remains unknown, and we can therefore only add the abbreviation to the *full dictionary*.

Another external source we use is the Termprofiling[17] module [47]. As mentioned in Section 2.6.1, AutoNER makes use of AutoPhrase to extract additional relevant terms for the *full dictionary*. However, AutoPhrase does not support Dutch without additionally training the model on Dutch corpora, and we therefore made use of a language-independent alternative in the form of Termprofiling. For the English datasets, we use a gamma of 0.8. For the Dutch endoscopy reports, we use a gamma of 0.5. Finally, we set the n-gram parameter to 3.

4.4.3 Manual dictionary tailoring

Although we tried to make all aforementioned steps as accurate and precise as possible, we could not avoid introducing a significant amount of false positives, i.e., terms with no relevance at all, to our dictionaries. There are also numerous terms that are labelled incorrectly for the i2b2 dataset and the endoscopy reports. This is due to a small overlap between the terms corresponding to each category, which was an inevitable result from the used queries.

To deal with these erroneous terms, we had to make manual adjustments to the dictionaries, traversing each list and removing terms accordingly.

[15] https://github.com/clips/pattern

[16] https://www.resourcepharm.com/pre-reg-pharmacist/medical-abbreviations.html

[17] https://github.com/suzanv/termprofiling

4.4.4 Development and test set construction

For the AutoNER model to train, a development and a test set are required. This at the same time allows us to gain any evaluation metric scores regarding the quality of the NER process. For the BC5CDR dataset, the development and test set are already included in the format suitable for AutoNER, while for the i2b2 dataset, we have to parse the existing sets into the correct format ourselves. For the endoscopy reports however, we constructed both sets by manually annotating a selection of documents.

The resulting development and test set consist of 9666 and 4680 annotated tokens respectively. Tokens are annotated as either *Problem*, *Treatment*, *Test* or *None*.

Results 5

In this chapter, we present and evaluate the results we obtain by various means. For all experiments regarding AutoNER, we use similar parameter settings as stated in the work by Shang *et al.* [41]: a stochastic gradient descent (SGD) update layer, a batch size of 10, a momentum of 0.9, and a dropout rate of 0.5. Only the learning rate, which we set to 0.01, differs from the 0.05 default value. We do so based on observations made from conducted experiments: while the results on the BC5CDR dataset when using the provided dictionaries remained practically unchanged, a slightly lower learning rate gave us slightly better results when our custom made dictionaries are used. Due to the scope of this project, we do not investigate further tuning on AutoNER's parameters.

Finally, as we also mentioned in previous sections, we only focus on the *Disease* category of the BC5CDR dataset. Thus, we do not attempt to define the dictionary part for the *Chemical* category.

5.1 Reproduction of AutoNER

To validate AutoNER, we use the provided dictionaries, the unannotated training set, and the annotated development and test sets. From the highly similar scores as presented by Table 4, we can conclude that AutoNER functions correctly on our machines. The negligible difference also shows that using a learning rate of 0.01 does not affect the NER process in a negative manner. Other factors that may have contributed to the slight deviation in the resulting numbers include the random initialisation of the model, and the random sampling of examples during the training phase. It is also shown that the scores of the *Disease* category are lower than those of the *Chemical* category, arguably implying that it is more difficult to correctly perform the NER task *Disease* entities.

	Precision	**Recall**	**F1-score**
Reported by Shang *et al.* (avg)	0.890	0.810	0.848
Obtained by reproduction (avg)	0.879	0.815	0.846
Chemical	0.907	0.848	0.876
Disease	0.844	0.779	0.808

Table 4: Precision, recall and f1-scores obtained on the BC5CDR dataset, using the provided dictionaries. Scores in the top two rows are the average of the two categories. We additionally provide the results for only the *Chemical* and *Disease* category obtained on our own machines.

When we investigate the predictions on token level, we can also confirm that AutoNER extracts entities that do not occur in the dictionaries. Namely, besides the 6877 dictionary entities, AutoNER labels 425 new entities (total over the development and test sets), the majority of which are valid candidate entities. Here, false positives come for example in the form of stand-alone numbers, abbreviations and names of body parts.

5.2 Initial AutoNER results on the i2b2 dataset

Now that we have confirmed that the original AutoNER results are reproducible, we apply the model to the i2b2 dataset. Our *core* and *full dictionary* consist of the entities extracted by the third method as described in Section 4.2.

It is important to note that at this point, we have made alterations to the development and test set. Possessive adjectives (his, her, etc.), demonstrative pronouns (these, that, etc.) and articles (a, an and the) do not yield any specific importance, though they are often included in entities of the i2b2 dataset. We also do not include these kind of tokens in entities of the endoscopy reports. Therefore, if an entity starts with any of these tokens, then that part of the entity is not taken into account, i.e., labelled *None*.

	Precision	**Recall**	**F1-score**
Problem	0.445	0.155	0.229
Treatment	0.177	0.012	0.022
Test	0.000	0.000	0.000
Average	0.207	0.055	0.084

Table 5: Precision, recall and f1-scores obtained on the i2b2 dataset, using the entity list obtained by our baseline method as dictionary content.

From the results presented in Table 5, it becomes evident that using our baseline method for dictionary building is far from optimal. Beside the low scores for *Problem*, and *Treatment* even less so, we actually find a precision, recall and f1-score equal to 0.0 for the *Test* category. This rather exceptional finding can be explained by the manner in which AutoNER evaluates the results it achieves on the development set, and how the model trains itself accordingly. Namely, if AutoNER does not find any improvement for a certain number of iterations, the learning rate drops. In this case, the model has most likely become stuck in a local optimum, as the possible improvements made on the *Test* category do not outweigh the simultaneous loss on the other two categories.

From the aforementioned baseline approach, however, it was already clear that the resulting entity lists, and in turn the dictionaries we have now used, are of rather low quality regarding both false positives and negatives. This particularly holds for the *Test* category.

Based on these observations, we can conclude that the dictionaries require a lot of improvement before the NER process will become as reliable as it is when using the provided dictionaries.

5.3 AutoNER with custom dictionaries on the BC5CDR dataset

Through thorough experimentation, we make use of a dictionary resulting from querying SnomedCT by means of the Snowstorm API, the addition of synonyms and plurals, abbreviations, terms extracted by the Termprofiling module, and manual pruning to remove as many false positives as possible. Our final *core* and *full dictionaries* hold 2512 annotated terms, and 2679 unannotated terms, respectively. The difference in numbers is caused by the additional unlabelled abbreviations and terms resulting from Termprofiling.

Table 6 shows the results of the NER process on the *Disease* category when using our custom-tailored dictionaries. For comparison, we also state the scores of the *Disease* category

obtained when using the provided dictionaries.

	Precision	**Recall**	**F1-score**
Disease (provided dictionaries)	0.844	0.776	0.808
Disease (custom dictionaries)	0.742	0.513	0.606

Table 6: Precision, recall and f1-scores obtained on the BC5CDR dataset, using the custom-tailored dictionaries.

Compared to the use of the provided dictionaries, we experience a significant drop in the f1-score when applying our custom dictionaries. This decrease is primarily caused by a substantially lower recall. The latter is most likely due to a smaller number of high frequency entities contained in the dictionary: whereas the custom *core dictionary* yields 2512 annotated *Disease* entities, the provided dictionary contains 'just' 1288 annotated entities. On the other hand, the custom *full dictionary* contains 2679 terms, whereas the original *full dictionary* contains 3761 terms.

We have also observed that the NER process converges much more quickly, requiring only 15 epochs compared to the 49 before.

The number of entities that AutoNER extracts which aren't in either dictionary has now increased from 425 to 1131. The majority of these entities are, as we also observed during the result reproduction, valid, with the exception of an unknown, though relatively small percentage of false positives.

Regardless, the quality, and quantity even more likely so, of the custom dictionaries do not seem to meet that of the provided dictionaries. Besides the use of different clinical databases from which the terms are extracted, there is another important difference found in the construction of in particular the *full dictionary*. Namely, the provided *full dictionary* contains additional terms extracted from the corpus by the AutoPhrase module. A small experiment, which involves the removal of these specific terms from the provided dictionary, leads to an f1-score of around 0.72, a decrease of approximately 0.08.

5.4 Final AutoNER results on the i2b2 dataset

As shown in Section 5.2, our preliminary results on the i2b2 dataset are far from sufficient. Now that we have defined a more extensive pipeline based on experiments conducted on the BC5CDR dataset, we return to the i2b2 dataset, and apply this pipeline to construct the corresponding dictionaries. We obtain 1518 entities for *Problem*, 556 entities for *Treatment*, and 286 for *Test*, totalling a number of 2360 annotated entities for the *core dictionary*. The corresponding *full dictionary* contains 3299 entities.

Table 7 denotes the obtained precision, recall and f1-scores for the NER process on the i2b2 dataset.

	Precision	**Recall**	**F1-score**
Problem	0.457	0.342	0.391
Treatment	0.515	0.234	0.321
Test	0.402	0.148	0.217
Average	0.458	0.241	0.310
Initial average	0.207	0.055	0.084

Table 7: Precision, recall and f1-scores obtained on the i2b2 dataset, using the custom-tailored dictionaries. Values in the *Initial average* row are the results we obtained by using our baseline method for dictionary building, as shown in Table 5.

Although a noticeable improvement is found compared to earlier results on the i2b2 dataset, particularly for the *Treatment* and *Test* categories, the overall scores indicate that the made predictions still aren't very accurate. Also, when we compare the f1-score obtained on the BC5CDR *Disease* category with the custom dictionaries, we find that the averaged f1-score on the i2b2 dataset is lower by almost 30%.

When we look at the confusion matrix, as shown in Figure 2, we can see that the number of incorrect predictions that occur among the three main categories is relatively low. Rather, most errors are the due to AutoNER missing out on entities that should be labelled according to any of these categories, as displayed by the rightmost column. Values found within the latter also further reflect the low recall scores.

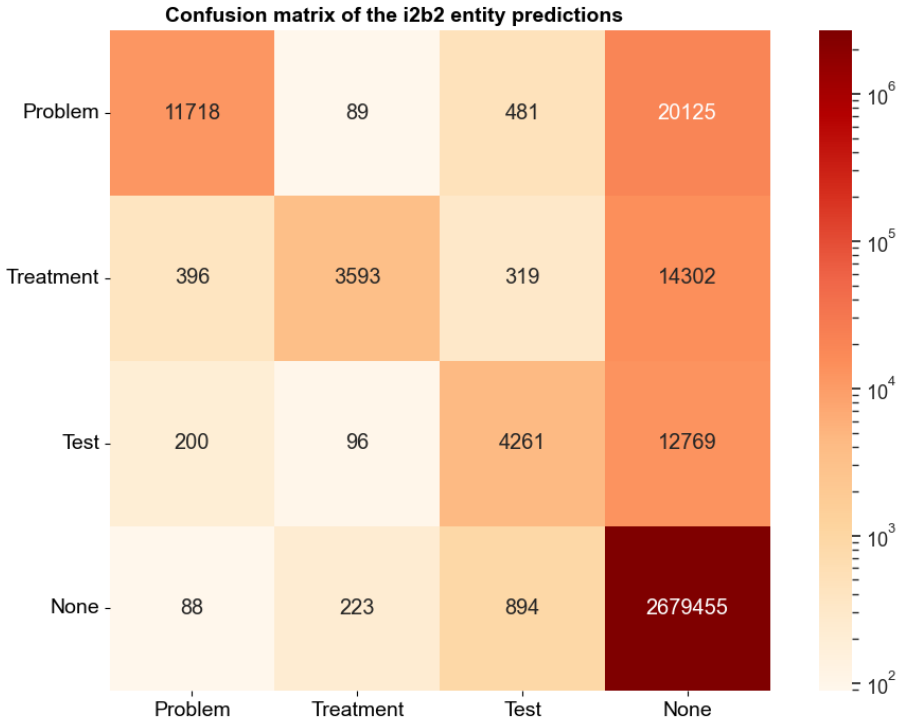


Figure 2: Confusion matrix for the three main categories of the i2b2 dataset, and the *None* category. Predictions are presented on the X-axis, while the truth values are on the Y-axis. Note that the heatmap colouring is done according to a logarithmic scale.

5.5 AutoNER with custom dictionaries on the Dutch endoscopy reports

For the endoscopy reports, the corresponding *core* and *full dictionaries* yield 348 annotated and 524 unannotated entities respectively. We immediately find that brand names rarely occur in our dictionaries. This is because such terms are already scarcely appearing within SnomedCT. The few that are present here, are often concatenated with dosage prescriptions. These kind of entities were always filtered out when we check whether or not they appear within the corpus, simply because of their format.

We also find that the Termprofiling module extracts several new valuable candidate entities from the corpus, which are included within the *full dictionary*.

Finally, taking into account all aforementioned results and observations, we apply our pipeline to the Dutch endoscopy reports. Table 8, shows the obtained precision, recall and f1-scores for the three main categories of *Problem*, *Treatment* and *Test*.

	Precision	Recall	F1-score
Problem	0.800	0.333	0.471
Treatment	0.818	0.188	0.305
Test	0.867	0.186	0.306
Average	0.828	0.236	0.361

Table 8: Precision, recall and f1-scores obtained on the Dutch endoscopy reports, using the custom-tailored dictionaries.

Although the precision score for all three categories is remarkably high, the recall scores are, similarly to those on the i2b2 dataset, low. Consequently, this leads to quite low overall f1-scores.

Term	Entity type
endoscopie	Test
midazolam	Treatment
bisacodyl	Treatment
poliep	Problem
anastomose	Problem
stenose	Problem
poliepen	Problem
aften	Problem
IBD	Problem
colitis	Problem
littekens	Problem

(a) The eleven most frequent correctly predicted entities.

Term	Entity type
Sedatie	Treatment
Colonoscopie incl biopsie	Test
Rectaal toucher	Test
scopie	Test
Picoprep	Treatment
M. Crohn	Problem
Dormicum	Treatment
ulcera	Problem
Poliep	Problem
Fentanyl	Treatment
Crohn	Problem

(b) The eleven most frequent incorrectly predicted entities.

Table 9: The eleven most frequent correctly and incorrectly predicted entities of the endoscopy reports, ordered by descending frequency.

We find that despite the relatively small development and test sets, AutoNER extracts a number of terms that are not in the dictionaries. From these fifteen additional terms, three were deemed to be irrelevant. The remaining terms were labelled correctly in nearly all cases.

Arguably the most prominent factor contributing to the low recall scores isn't the complete absence of certain concepts from the dictionary, but rather the format in which they occur within the texts themselves. Namely, if we look at the texts, we see numerous concepts that can be referred to in a variety of ways. One of these terms is *ziekte van Crohn* (*Crohn's disease*), which in the corpus may also appear as *M. Crohn*, *M.C.*, *MCrohn*, and *Crohn*. Even though *ziekte van Crohn* appears within our *core dictionary*, none of its variations do, and none are recognised by AutoNER as *Problem* entities. Moreover, aforementioned variations are far more frequent than the *core* entity, resulting in numerous false negatives. This is also shown in Table 9, where both *M. Crohn* and *Crohn* appear among the entities that are mislabelled most frequently.

There are 38 tokens that do not occur in the used embedding. This also holds for nearly all of the abbreviations if we distinguish the fully capitalised tokens as they appear in the

texts, from the lower-cased formats that may occur in the embeddings. We initially assumed that the low recall scores are caused by terms not occurring in the embeddings we use. In this case, AutoNER resorts to a 'default' embedding array, one that does not hold any related information to the actual entity. However, when we look at the token-level predictions made on both the development and test set, this assumption does not seem to uphold itself in all cases. For example, both the terms *IBD* and *MTX* do not appear in the embedding, and both are contained within the *core dictionary*, labelled as *Problem* and *Treatment* respectively. However, whereas *IBD* is in all cases but one predicted correctly, *MTX* is never regarded as a *Treatment* by AutoNER.

Furthermore, there are terms that do occur in both the embeddings and the *core dictionary*, but are still not assigned any of the three categories by AutoNER. Examples in this case are *Coloscopie*, a *Problem*, and *Fentanyl*, a *Treatment*. We also find that AutoNER does not adapt well in cases of capitalised terms. Here, an example is that of *poliep* and *Poliep*, which is also shown in Table 9. Despite *poliep* occurring in both dictionaries and the embeddings, *Poliep* is never labelled correctly. The definition of this term is not affected by capitalisation, and should only be excluded from the *Problem* category based on context, e.g., no *Poliep* is found. This however does not apply, and therefore this term should have been labelled by its respective category. In general, it was surprising to see that the correlation between the prediction made on a particular entity, its presence or absence in the embeddings and the dictionaries, was not as consistent as we expected it to be.

Based on the confusion matrix as shown in Figure 3, we are able to conclude that the number of erroneous predictions is almost solely due to entities not being labelled by any of the three actual categories, and that incorrect predictions between entities of these three categories are actually never made. Consequently however, this introduces a new oddity: based on the precision scores for the categories, we expected to see more, if any false positives portrayed in the confusion matrix. This is particularly the case for the *Test* category, for which the matrix only holds false negatives, i.e., entities that should be labelled as *Test* are missed out on; labelled *None*. However, for *Test*, AutoNER returns a precision score of 0.867, meaning that not all entities AutoNER has labelled as *Test* should be assigned this label.

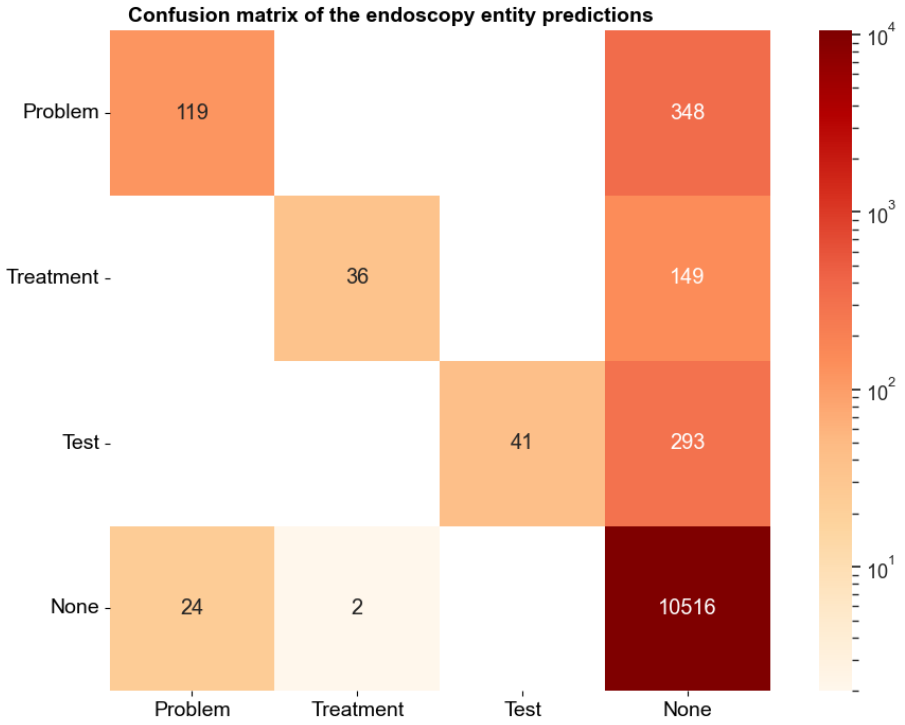
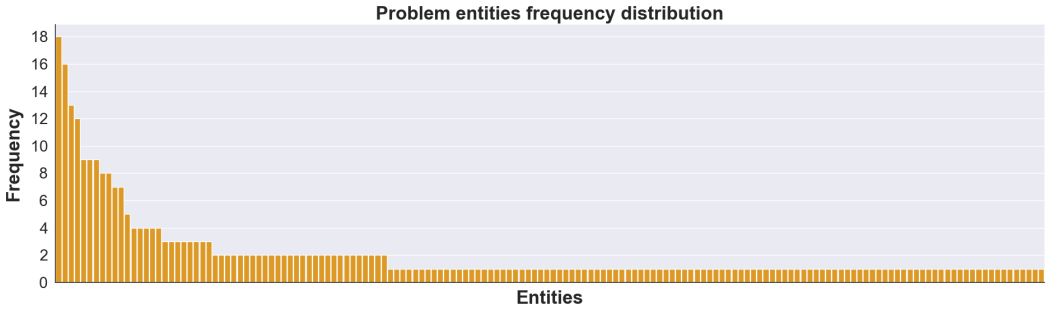


Figure 3: Confusion matrix for the three main categories of the endoscopy report dataset, and the *None* category. Predictions are presented on the X-axis, while the truth values are on the Y-axis. Note that the heatmap colouring is done according to a logarithmic scale.

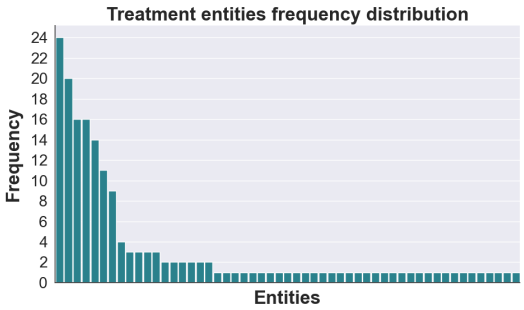
Manual evaluation of the predictions confirms that the confusion matrix is completely accurate, and that there are indeed no false positives among the predicted *Test* entities. Despite further investigation, we were unable to determine which exact part of the extraction process introduces this inconsistency. We reason that during training, the model indeed achieves the aforementioned scores. Once the training is finished and the model is applied to the development and test set, it happens to not make any predictions that lead to false positives. Rerunning the experiments with exactly the same settings however did not lead to any false positives.

A more general factor that likely affected the NER process as well, is the size of the development and test set. Compared to those of the other two datasets we used, the number of lines is approximately 25 times less.

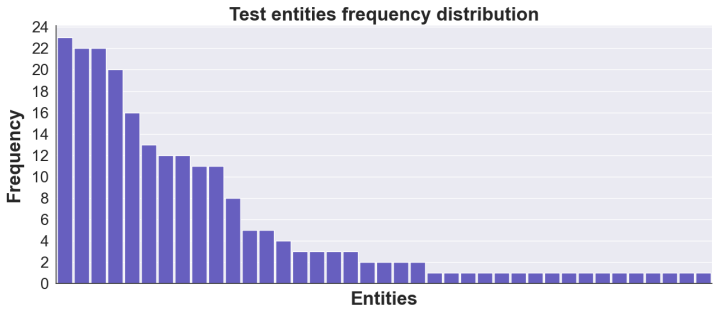
The frequency distribution of entities for all three categories is also skewed, as is shown in Figure 4. Although not necessarily long-tail distributions, it is still fairly obvious that if AutoNER fails to correctly label any entities that are on the far left side of the spectrum, the obtained scores are lowered significantly.



(a) Frequency distribution of the 158 unique entities labelled as *Problem*.



(b) Frequency distribution of the 53 unique entities labelled as *Treatment*.



(c) Frequency distribution of the 39 unique entities labelled as *Test*.

Figure 4: The frequency distributions of entities labelled as *Problem*, *Treatment* and *Test* contained in the development and test set of the Dutch endoscopy reports dataset.

The size of the development and test set paired with the uneven distribution of entity frequencies therein, makes it more difficult to draw conclusions that are completely reliable. It could therefore be the case that when these issues are addressed, AutoNER may perform better without making any alterations to the dictionaries or other parameters. On the other hand, taking into account the results on the i2b2 dataset, we've seen that larger development and test sets do not automatically imply that AutoNER will perform (significantly) better.

Discussion 6

We started with several result reproductions from related research on other clinical datasets, with in particular the i2b2 and the BC5CDR dataset. These scores were used to determine which results we could expect as upper bound. We then applied AutoNER to these datasets with dictionaries we defined using a clinical term database in the form of SnomedCT. Finally, taking into consideration all observations and conclusions made on these benchmark datasets, we applied AutoNER with again specifically tailored dictionaries to the Dutch endoscopy reports.

Already in early stages of this research, it became clear that the mapping of SnomedCT terms onto the categories of other datasets was very challenging, and that more thorough approaches were required in order to improve upon the naive baseline f1-score of 0.145 for the i2b2 dataset. One of the tools we opted to change was the PyMedTermino package. Although it allowed for a decent amount of options to exploit SnomedCT's multi-hierarchical design, it did not provide us with the necessary options to extract all the terms we required.

We then defined a pipeline for the application of AutoNER. The main challenge here was the construction of the dictionaries, a process also affected by the persisting mapping issue. In order to make the problem slightly less challenging, we first focused solely on the *Disease* category of the BC5CDR dataset, thereby narrowing down the number of categories to one. This eventually led to a pipeline for reliable dictionary construction, taking into account the differences between the *core* and *full dictionaries*. The extraction of terms and their respective synonyms from SnomedCT is now done using Snowstorm. Consequently, the resulting terms are considerably more accurate because of the ECL queries. For all these terms, we obtained synonyms and added plurals.

After filtering the list of terms based on the contents of the full corpus, we again extended upon it by adding annotated abbreviations for the *core dictionary* and unannotated abbreviations and terms resulting from the Termprofiling module for the *full dictionary*. Finally, we executed a series of manual modifications to the dictionaries in order to remove as many false positives as possible.

By applying our pipeline to the BC5CDR dataset, we managed to obtain an f1-score of 0.606 for the *Disease* category. This was 0.202 lower compared to f1-score when the provided dictionaries are used. The i2b2 dataset proved to be more difficult, and even after thorough dictionary tailoring, we achieved f1-scores of 0.391 for *Problem*, 0.321 for *Treatment*, and 0.217 for *Test*. Compared to our supervised approach in the form of BioBERT, the gap is significantly larger: 0.567 averaged over the three categories.

For both datasets, we saw that the f1-scores were always affected greatly by low recall values. We argued that, despite the quality of the dictionaries, they might still lacking quantity-wise. Consequently, AutoNER was unable to extract a sufficient amount of new terms in order to achieve higher f1-scores.

On the endoscopy reports, AutoNER performed slightly better than it did on the i2b2 dataset, obtaining an f1-score of 0.471 for *Problem*, 0.305 for *Treatment*, and 0.306 for *Test*. AutoNER's ability to extract new terms that do not occur in the dictionaries was not sufficient enough, leaving the majority of terms unrecognised.

For domain specific dictionary construction, as was the case for the endoscopy reports,

we conclude that even though SnomedCT does contain all clinical concepts, we still miss out on a large amount of corresponding variants. For example, we extracted *Ziekte van Crohn* (*Crohn's disease*), but there are various ways for this term to occur in the actual reports: *M.C.*, *M.Crohn*, etc., which are naturally not covered by SnomedCT. This in turn leaves our dictionaries not fully representing the collection of relevant entities that occur in the corpus, impairing AutoNER's learning process.

Based on all aforementioned observations and conclusions, we have to address the fact that AutoNER may not be as practically applicable as we initially assumed. First and foremost is that defining and applying a suitable pipeline for dictionary construction is extremely time-consuming, as the dictionaries should contain as many relevant terms as possible, but at the same time yield relatively no false positives. In particular for the latter, a lot of manual labour is required, as basically all initial dictionary terms have to be filtered based on relevance and whether the assigned label is correct or not.

Second, AutoNER still requires a development and test set in order for the model to train, and consequently to actually allow for the extraction of new relevant entities. Then, there is the issue of AutoNER occasionally not recognising entities that are present in the dictionaries. Although terms should or should not be labelled based on context, this is undesired behaviour, especially if the dictionaries are tailored to hold specific labelled terms, which still may not end up being extracted.

Conclusions 7

In this paper, we have addressed the automatic extraction of medical entities from Dutch endoscopy reports using a distant supervised method called AutoNER. Our methodology included a series of reproductions, a baseline approach and the application of AutoNER to the BC5CDR-Disease, i2b2 and the Dutch endoscopy reports datasets. Taking all made observations into account, we will now address our research questions. The first research question was:

- What kind of information can we extract from endoscopy reports, with the general aim to improve the DEARhealth Care Pathways?

We have seen that we are to able extract entities from the Dutch endoscopy reports using the knowledge-based, distantly supervised model called AutoNER. We are able to recognise terms implying the presence of a disease or other medical issue (*Problem*), terms that refer to the administration or use of medicine, surgeries and therapies (*Treatment*) and clinical examinations and assessment procedures (*Test*). Using this information, we can prepare the foundation of relevant medical history for each patient in a structured manner, which in turn could be used for the recommender.

Our second research question was:

- Which methods deem the most reliable for this problem?

We have to conclude that the pipeline we defined for AutoNER requires additional tailoring and improvement before this approach could be considered completely reliable. Though we are able to extract entities belonging to either of the aforementioned categories, a relatively large amount of entities is still missed out on.

In general, AutoNER can make for a decent initial round of Named Entity Recognition, but considering the amount of time required in order to produce both dictionaries, a development and a test set, we are not convinced that AutoNER is a reliable and efficient alternative to fully supervised models, like BioBERT. One could arguably spend an equal amount of time producing sufficiently large enough annotated training, development and test sets, thus allowing for the training of supervised models.

It isn't exactly clear as to where the trade-off point lies between additional (manual) tailoring of the dictionaries, and the reliability of automatic entity extraction by AutoNER with these dictionaries.

Future Work and recommendations 8

There are several aspects of our work that could prove interesting for follow-up research.

The first is to address the rather low recall scores we obtained by using AutoNER and our custom dictionaries on the Dutch endoscopy reports. We already argued that we missed out on a lot of concept variants, despite the concept itself being covered by SnomedCT and consequently appearing within our dictionaries. A suggestion could be that the entity lists are extended upon by those who have written the reports, or other experts in this domain specific field who are highly familiar with the contents of endoscopy reports. This should include as many contributors as possible, as each may have his or her own writing style and semantic preferences. However, it has to be noted that even though this might be beneficial for AutoNER, the approach will become more akin an unsupervised dictionary-based approach, rather than a distant supervised one (which AutoNER is suggested to be).

Another option that could be taken into consideration is the training of AutoPhrase on a large amount of Dutch (clinical) corpora. Although we have seen that Termprofiling did extract some additional terms for the *full dictionary*, it could be that AutoPhrase is actually able to detect the aforementioned variants. This would thus also further automate the dictionary tailoring process.

Furthermore, the relatively small development and test set may also negatively affect the training process. A rather straightforward addition to our research may therefore be the expansion of the development and test sets, and determine whether some immediate improvement is found. Involvement of domain related experts could also prove to be worthwhile here. Although we were able to construct reliable annotated development and test sets ourselves, an expert might still insist on some slight alterations. These could be the addition or removal of entities, the inclusion or exclusion of certain words from entities, etc. When including multiple experts, we suggest annotation is done by using a tool such as Doccano[18] in order to reach consensus.

If one continues to experiment with AutoNER, there exists the option of hyper-parameter optimisation through for example grid search. However, taking into account the issues we encountered, dictionary and set modifications would most likely prove to be a more important factor when it comes to achieving more accurate predictions. Naturally, once this is more thoroughly addressed, hyper-parameter optimisation might lead to even more precise results.

Regarding the DEARhealth recommender, the implementation of follow-up tasks such as Named Entity Linking (NEL) and Relation Extraction (RE) would result into additional, highly relevant knowledge in a structured format. Currently, we have addressed the mere extraction of entities. If for example a report states that *formerly observed scarring has lead to an intestinal blockage*, we will see in our results that both the terms *scarring* and *intestinal blockage* are labelled as a *Problem*. However, we cannot yet derive that the scarring actually caused the intestinal blockage. Similarly, we are not yet able to link dosages to any *Treatment* entity, or determine whether a *Test*, such as an endoscopy, caused the patient any pain because of an inflammation in the terminal ileum.

Finally, in combination with the aforementioned follow-up tasks, the presence of discontinuous entities should be addressed. An example here is: *abnormal vascular patterns and*

[18] https://github.com/doccano/doccano

haustra markings. Although *abnormal vascular patterns* may be recognised as a *Problem* entity, we will not yet find *abnormal haustra markings* among the results. Finding a reliable method to extract these kind of discontinuous entities will undoubtedly lead to more complete and accurate knowledge representation.

References

[1] Diagnostiek en behandeling van inflammatoire darmziekten bij kinderen. pages 271–276. Van Zuiden Communications B.V, 2008. ISBN 978-90-8523-179-0. URL http://www.kindergeneeskunde-mca.nl/images/stories/medische_protocollen/ibd_rl_cbo_08.pdf.

[2] S. Ali, F. Zhou, C. Daul, B. Braden, A. Bailey, S. Realdon, J. East, W. G., V. Loschenov, E. Grisan, W. Blondel, and J. Rittscher. Endoscopy artifact detection (ead 2019) challenge dataset. *CoRR*, 05 2019.

[3] V. Annese, M. Daperno, M. Rutter, A. Amiot, P. Bossuyt, J. East, M. Ferrante, M. Götz, K. Katsanos, R. Kiesslich, I. Ordas, A. Repici, B. Rosa, S. Sebastian, T. Kucharzik, and R. Eliakim. European evidence based consensus for endoscopy in inflammatory bowel disease. *Journal of Crohn's and Colitis*, 7:982–1018, 12 2013. doi: 10.1016/j.crohns.2013.09.016.

[4] R. Baldassano, J. Bradfield, D. Monos, C. Kim, J. Glessner, T. Casalunovo, E. Frackelton, F. Otieno, S. Kanterakis, J. L. Shaner, R. M. Smith, A. W. Eckert, L. J. Robinson, C. C. Onyiah, D. Abrams, R. Chiavacci, R. Skraban, M. Devoto, S. Grant, and H. Hakonarson. Association of the t300a non-synonymous variant of the atg16l1 gene with susceptibility to paediatric crohn's disease. *Gut*, 56:1171 – 1173, 2007.

[5] E. A. Benchimol, K. J. Fortinsky, P. Gozdyra, M. van den Heuvel, J. van Limbergen, and A. Griffiths. Epidemiology of pediatric inflammatory bowel disease: A systematic review of international trends. *Inflammatory Bowel Disease*, 17(1):423–429, 2010. doi: 10.1002/ibd.21349.

[6] C. Blake and T. Rindflesch. Leveraging syntax to better capture the semantics of elliptical coordinated compound noun phrases. *Journal of Biomedical Informatics*, 72, 07 2017. doi: 10.1016/j.jbi.2017.07.001.

[7] A. Bowe. Au naturale - an introduction to nltk, 03 2011. URL https://alexbowe.com/au-naturale/.

[8] M. Bretthauer, L. Aabakken, E. Dekker, M. Kaminski, T. Rösch, R. Hultcrantz, S. Suchanek, R. Jover, E. Kuipers, R. Bisschops, C. Spada, R. Valori, D. Domagk, C. Rees, and M. Rutter. Reporting systems in gastrointestinal endoscopy: Requirements and standards facilitating quality improvement: European society of gastrointestinal endoscopy position statement. *United European Gastroenterology Journal*, 4:172–176, 04 2016. doi: 10.1177/2050640616629079.

[9] J. Brown and S. Zeki. Oth-07 identification of ibd cohorts from linked endoscopy and histology reports using natural language processing. *Gut*, 68(Suppl 2):A224–A224, 2019. ISSN 0017-5749. doi: 10.1136/gutjnl-2019-BSGAbstracts.426. URL https://gut.bmj.com/content/68/Suppl_2/A224.1.

[10] H. Cho and H. Lee. Biomedical named entity recognition using deep neural networks with contextual information. *BMC Bioinformatics*, 20, 12 2019. doi: 10.1186/s12859-019-3321-4.

[11] R. Cima and B. Wolff. Reoperative crohn's surgery: Tricks of the trade. *Clinics in colon and rectal surgery*, 20:336–43, 11 2007. doi: 10.1055/s-2007-991034.

[12] G. Cope. Overview of dietary choices for ulcerative colitis and crohn's disease. *Gastrointestinal Nursing*, 13:35–41, 02 2015. doi: 10.12968/gasn.2015.13.1.35.

[13] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL `http://arxiv.org/abs/1810.04805`.

[14] R. Dogan, R. Leaman, and Z. lu. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47, 01 2014. doi: 10.1016/j.jbi.2013.12.006.

[15] M. Fares, A. Kutuzov, S. Oepen, and E. Velldal. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In J. Tiedemann and N. Tahmasebi, editors, *Proceedings of the 21st Nordic Conference on Computational Linguistics, NODALIDA 2017, Gothenburg, Sweden, May 22-24, 2017*, pages 271–276. Association for Computational Linguistics, 2017. ISBN 978-91-7685-601-7. URL `https://aclanthology.info/papers/W17-0237/w17-0237`.

[16] G. Flett, C. Baricza, A. Gupta, P. Hewitt, and N. Endler. Perfectionism, psychosocial impact and coping with irritable bowel disease: A study of patients with crohn's disease and ulcerative colitis. *Journal of health psychology*, 16:561–71, 02 2011. doi: 10.1177/1359105310383601.

[17] L. Graff, J. Walker, and C. Bernstein. It's not just about the gut: Managing depression and anxiety in inflammatory bowel disease. *Practical Gastroenterology*, 34:11–25, 07 2010.

[18] J. Hou, M. Chang, T. Nguyen, J. Kramer, P. Richardson, L. Sansgiry, S.and D'Avolio, and H. El-Serag. Automated identification of surveillance colonoscopy in inflammatory bowel disease using natural language processing. *Digestive diseases and sciences*, 58, 10 2012. doi: 10.1007/s10620-012-2433-8.

[19] M. Kappelman, K. Moore, J. Allen, and S. Cook. Recent trends in the prevalence of crohn's disease and ulcerative colitis in a commercially insured us population. *Digestive diseases and sciences*, 58, 08 2012. doi: 10.1007/s10620-012-2371-5.

[20] J. Kelsen and R. N. Baldassano. Inflammatory bowel disease: the difference between children and adults. *Inflammatory Bowel Disease*, 15(9):1438–1447, 2008. doi: 10.1002/ibd.20560.

[21] S. C. Kim and G. D. Ferry. Inflammatory bowel diseases in pediatric and adolescent patients: Clinical, therapeutic, and psychosocial considerations. *Gastroenterology*, 126 (6):1550–1560, 2004. doi: 10.1053/j.gastro.2004.03.022.

[22] S. N. Kim, T. Baldwin, and M.-Y. Kan. Evaluating n-gram based evaluation metrics for automatic keyphrase extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 572–580, Beijing, China, 08 2010. Coling 2010 Organizing Committee. URL https://aclanthology.org/C10-1065.

[23] K. Kreimeyer, M. Foster, A. Pandey, N. Arya, G. Halford, S. Jones, R. Forshee, M. Walderhaug, and T. Botsis. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *Journal of Biomedical Informatics*, 73, 07 2017. doi: 10.1016/j.jbi.2017.07.012.

[24] E. Kuipers, G. Henegouwen, P. Fockens, and R. Ouwendijk. Computerisation of endoscopy reports using standard reports and text blocks. *The Netherlands journal of medicine*, 64:78–83, 04 2006.

[25] J.-B. Lamy, A. Venot, and C. Duclos. Pymedtermino: An open-source generic api for advanced terminology services. *Studies in health technology and informatics*, 210:924–8, 05 2015. doi: 10.3233/978-1-61499-512-8-924.

[26] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics (Oxford, England)*, 36, 09 2019. doi: 10.1093/bioinformatics/btz682.

[27] P. Lisboa-Gonçalves, D. Libanio, J. Marques-Antunes, M. Dinis-Ribeiro, and P. Pimentel-Nunes. Quality of reporting in upper gastrointestinal endoscopy: Effect of a simple audit intervention. *GE - Portuguese Journal of Gastroenterology*, 26, 04 2018. doi: 10.1159/000487145.

[28] J. Liu, J. Shang, C. Wang, X. Ren, and J. Han. Mining quality phrases from massive text corpora. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, page 1729–1744, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450327589. doi: 10.1145/2723372.2751523. URL https://doi.org/10.1145/2723372.2751523.

[29] Y. Luo. Recurrent neural networks for classifying relations in clinical notes. *Journal of Biomedical Informatics*, 72, 07 2017. doi: 10.1016/j.jbi.2017.07.006.

[30] I. Martinez Soriano and J. Castro. Dner clinical (named entity recognition) from free clinical text to snomed-ct concept. *WSEAS Transactions on Computers*, 16:83–91, 01 2017.

[31] B. Martínez-Salvador, M. Marcos, A. Mañas, J. Maldonado, and M. Robles. Using snomed ct expression constraints to bridge the gap between clinical decision-support and electronic health records. 08 2016. doi: 10.3233/978-1-61499-678-1-504.

[32] P. Olivera Sendra, S. Danese, N. Jay, G. Natoli, and L. Peyrin-Biroulet. Big data in ibd: a look into the future. *Nature Reviews Gastroenterology & Hepatology*, 01 2019. doi: 10.1038/s41575-019-0102-5.

[33] J. Prasko, D. Jelenova, and V. Mihal. Psychological aspects and psychotherapy of inflammatory bowel disease and irritable bowel syndrome in children. *Biomedical papers of*

the Medical Faculty of the University Palacký, Olomouc, Czechoslovakia, 154:307–14, 12 2010. doi: 10.5507/bp.2010.046.

[34] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, and S. Ananiadou. Distributional semantics resources for biomedical text processing. *Proceedings of Languages in Biology and Medicine*, 01 2013.

[35] A. Ratner, S. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré. Snorkel: rapid training data creation with weak supervision. *The VLDB Journal*, 29, 05 2020. doi: 10.1007/s00778-019-00552-1.

[36] D. Robertson, J. Ray, I. Diamond, and J. G. Edwards. Personality profile and affective state of patients with inflammatory bowel disease. *Gut*, 30:623–6, 06 1989. doi: 10.1136/gut.30.5.623.

[37] B. H. G. Rogers, L. M. Clark, and J. B. Kirsner. The epidemiologic and demographic characteristics of inflammatory, bowel disease: An analysis of a computerized file of 1400 patients. *Journal of Chronic Diseases*, 24(12):743–773, 1971. ISSN 0021-9681. doi: 10.1016/0021-9681(71)90087-7.

[38] M. Sajadinejad, K. Mobarakeh, H. Molavi, M. Kalantari, and P. Adibi. Psychological issues in inflammatory bowel disease: An overview. *Gastroenterology research and practice*, 2012:106502, 06 2012. doi: 10.1155/2012/106502.

[39] N. S. Seyed Tabib, M. Madgwick, P. Sudhakar, B. Verstockt, T. Korcsmaros, and S. Vermeire. Big data in ibd: big progress for clinical practice. *Gut*, 69(8):1520–1532, 2020. doi: 10.1136/gutjnl-2019-320065.

[40] J. Shang, J. Liu, M. Jiang, X. Ren, C. R. Voss, and J. Han. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[41] J. Shang, L. Liu, X. Ren, X. Gu, T. Ren, and J. Han. Learning named entity tagger using domain-specific dictionary. In *EMNLP*, 2018.

[42] P. Su, G. Li, C. Wu, and K. Vijay-Shanker. Using distant supervision to augment manually annotated data for relation extraction. 05 2019. doi: 10.1101/626226.

[43] A. Symeonidou, V. Sazonau, and P. Groth. Transfer learning for biomedical named entity recognition with biobert. In *SEMANTICS Posters&Demos*, 2019.

[44] Y. TIAN, W. Shen, Y. Song, F. Xia, M. He, and K. Li. Improving biomedical named entity recognition with syntactic information. 04 2020. doi: 10.21203/rs.3.rs-21994/v1.

[45] O. Uzuner, B. South, S. Shen, and S. DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18:552–6, 06 2011. doi: 10.1136/amiajnl-2011-000203.

[46] G. van den Brink, L. Stapersma, H. El Marroun, J. Henrichs, E. Szigethy, E. Utens, and J. Escher. Effectiveness of disease-specific cognitive–behavioural therapy on depression, anxiety, quality of life and the clinical course of disease in adolescents with inflammatory bowel disease: study protocol of a multicentre randomised controlled trial (happy-ibd). *BMJ Open Gastroenterology*, 3:e000071, 03 2016. doi: 10.1136/bmjgast-2015-000071.

[47] S. Verberne, M. Sappelli, D. Hiemstra, and W. Kraaij. Evaluation and analysis of term scoring methods for term extraction. *Information Retrieval*, 2016. doi: 10.1007/s10791-016-9286-2.

[48] Y. Wang, S. Sohn, S. Liu, F. Shen, L. Wang, E. Atkinson, S. Amin, and H. Liu. A clinical text classification paradigm using weak supervision and deep representation. *BMC Medical Informatics and Decision Making*, 19, 2019.

[49] C.-H. Wei, Y. Peng, R. Leaman, A. P. Davis, C. Mattingly, J. Li, T. Wiegers, and Z. lu. Overview of the biocreative v chemical disease relation (cdr) task. pages 154–166, 09 2015.

[50] Y. Wu, J. Xu, M. Jiang, Y. Zhang, and H. Xu. A study of neural word embeddings for named entity recognition in clinical text. *AMIA Annual Symposium Proceedings*, 2015: 1326–1333, 11 2015.

[51] S. Zeki. Endominer for the extraction of endoscopic and associated pathology data from medical reports. *Journal of Open Source Software*, 3:701, 04 2018. doi: 10.21105/joss.00701.

[52] S. Zeki. Ptu-144 removal of normal and negative phrases from endoscopic semi-structured text for accurate automated endoscopic audit. *Gut*, 67(Suppl 1):A267–A269, 2018. ISSN 0017-5749. doi: 10.1136/gutjnl-2018-BSGAbstracts.522. URL `https://gut.bmj.com/content/67/Suppl_1/A267.2`.

[53] S. Zeki, R. Hackett, J. Dunn, A. Bancil, S. Preston, J. Chin-Aleong, J. Brown, and S. McDonald. Ptu-105 automated, algorithm based extraction of barrett's surveillance metrics from natural language text is reliable. *Gut*, 68(Suppl 2):A243–A243, 2019. ISSN 0017-5749. doi: 10.1136/gutjnl-2019-BSGAbstracts.464. URL `https://gut.bmj.com/content/68/Suppl_2/A243.1`.