UNIVERSITY
OF OSLO

Master's thesis

Negation Resolution for Norwegian Medical Text

Annotation, modeling and domain portability

Marie Emerentze Fleisje

Informatics: Language Technology
60 ECTS study points

Department of Informatics
Faculty of Mathematics and Natural Sciences

Spring 2023

Marie Emerentze Fleisje

Negation Resolution for Norwegian Medical Text

Annotation, modeling and domain portability

Supervisor:
Lilja Øvrelid

Abstract

Detecting negation and resolving its scope is an essential task in NLP and a priority area in the clinical subfield of NLP. For larger languages such as English, there has been much research in the development of datasets and models for negation resolution, including efforts targeting the medical and clinical domains. Neural methods have come to dominate negation modeling in recent years, but simpler, rule-based approaches are still popular in medical applications. For Norwegian, the availability of resources for negation resolution has until recently been quite sparse.

In this thesis, we train an end-to-end negation resolution system utilizing a negation dataset of Norwegian review articles and a simple neural approach inspired by previous work. Using standardized evaluation metrics, the models achieve good results on in-domain test data. Furthermore, we evaluate the applicability of the existing dataset and its guidelines to future projects. Our review shows that better specification of the guidelines is desirable and reveals inconsistencies and annotation errors in the dataset.

Building on a previously released dataset, we present $\text{NorMed}_{\text{neg}}$, a publicly available Norwegian negation dataset of biomedical journal articles annotated according to an adjusted version of the mentioned guidelines. The transfer of our models to the medical domain represented by $\text{NorMed}_{\text{neg}}$ leads to poor performance, but we find that this can be compensated for by further training on parts of $\text{NorMed}_{\text{neg}}$. A positive effect is observed even with small amounts of training data.

Considering the focus on negated symptoms and findings in clinical NLP, we provide our thoughts on the use of models trained according to the existing annotation scheme in a clinical setting. We conclude that adjustments are necessary if the goal is to identify the specific clinical entities described as absent.

Acknowledgements

I would like to express my gratitude to my main supervisor Lilja Øvrelid for excellent guidance and invaluable feedback over the past two years.

Many thanks to Pål Brekke for productive discussions during the initial phases of this project and for being accessible and willing to answer technical and medical-related questions.

I am also grateful to Thor Stenbæk and DIPS, who played an important role in initiating the project.

Special thanks to Petter Mæhlum for his annotation effort and help with the computation of inter-annotator agreement, and for much good advice along the way.

Lastly, I would like to mention my family and friends. I really appreciate the encouragement and moral support they provided me throughout the sometimes stressful process of writing this master's thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Negation is a well-known linguistic phenomenon, traditionally viewed as an operator that changes the truth value of a proposition (Morante and Sporleder, 2012). It is frequently used in spoken and written language and intuitively understood by the language-processing parts of the human brain. For example, to English speakers it will be obvious that the two sentences below are substantially different with respect to their meaning:

(1.1) (a) He will pass the exam.
(b) He will not pass the exam.

According to Morante and Sporleder (2012), negation is an important contributor to the so-called *extra-propositional* aspects of meaning. Parallel to the example of Prabhakaran et al. (2010), we can illustrate the proposition in (1.1) as PASS(HE, THE EXAM). Both (a) and (b) contain this proposition, but to express the meaning of (b), we must add something extra-propositional, i.e. the negation operator.

In natural language processing (NLP), one is interested in detecting the presence of negation in text, and in determining which parts of the text are affected by it. This task is usually straightforward for humans, but research has shown that even state-of-the-art models in NLP struggle with handling negation correctly (Hosseini et al., 2021). Improving the processing of negation will be beneficial to many tasks, such as machine translation (Hossain et al., 2020), sentiment analysis (Wiegand et al., 2010) and textual entailment recognition (Helwe et al., 2022), and bring NLP closer to the ultimate goal of understanding text (Morante and Sporleder, 2012).

Negation also plays an important role in the medical and clinical field, where the objective is the identification of symptoms and findings as present or absent (Dalianis, 2018). Thus, many efforts have been directed toward this domain specifically. For the Norwegian language, the number of contributions is very limited, and there is a need for more research. In general, there has been a lack of Norwegian resources for negation resolution until the release of $\text{NoReC}_{\text{neg}}$ (Mæhlum et al., 2021), a dataset built from review articles.

The task of negation modeling has been approached in numerous ways by the NLP community. In recent years, neural methods have gained popularity, and some of the most successful systems have used a sequence labeling approach (Khandelwal and Britto, 2020; Khandelwal and Sawant, 2020). As of today, there are no publications on neural sequence labeling models for negation in Norwegian, and thus, we plan to experiment with such an approach in this thesis, using $\text{NoReC}_{\text{neg}}$ for training and evaluation.

The $\text{NoReC}_{\text{neg}}$ dataset is based on an annotation scheme created by adapting the guidelines of previous annotation efforts to Norwegian (Mæhlum et al., 2021). This scheme has not yet been subject to a thorough evaluation. With future annotation efforts and modeling projects in mind, it is important to ensure the quality of these guidelines as well as the consistency and correctness of the annotations, and this will be one of our focus areas.

The portability of the guidelines and models trained on the dataset into other *domains* has not been assessed either. Given the importance of negation resolution in medicine, it is of interest to evaluate the transferability of these resources from review texts to medical text. In this thesis, we will therefore annotate a dataset of biomedical articles and use it to explore the performance of models based on $\text{NoReC}_{\text{neg}}$ in the medical field. Furthermore, we will briefly consider the applicability of the annotation guidelines in a medical setting.

A change of domain is likely to cause a drop in model performance due to differences in writing style and vocabulary. Therefore, we aim to investigate the effects on performance seen when allowing the models to learn using data from the new target domain as well. Through such experiments, we can also get an indication as to the size of the effect in relation to the amount of in-domain training data.

1.1 Research questions

Based on the discussion above, we formulate the following research questions, which we aim to investigate as part of this thesis:

RQ1: Can we achieve state-of-the-art results for negation resolution in Norwegian with a neural sequence labeling system?

RQ2: How applicable are the $\text{NoReC}_{\text{neg}}$ resources to new projects and new domains?

RQ3: Can a negation resolution system fine-tuned on review articles be ported into the medical domain without a loss of performance?

RQ4: How are the results in the medical domain affected by further fine-tuning of the aforementioned system on medical text?

1.2 Outline

The contents of this thesis are structured as follows:

Chapter 2 introduces the role of negation in NLP. We provide an overview of differences in annotation schemes in existing datasets and put the most emphasis on the annotation scheme of the dataset we use to train our models in chapter 3, $\text{NoReC}_{\text{neg}}$. Negation modeling approaches in earlier work are summarized. Furthermore, the field of clinical NLP is presented, including an overview of work targeting negation in clinical text, especially for Norwegian.

Chapter 3 details our approach to building a system for detecting negation and resolving its scope in Norwegian text. We describe our experimental setup and evaluation methods. Modeling results are presented, and a detailed error analysis is provided.

Chapter 4 reviews the annotation guidelines and practice in $\text{NoReC}_{\text{neg}}$. We identify cases of unclear guidelines and how these cases are treated by the annotators. Also, we point to cases of clear violations of the guidelines.

Chapter 5 describes the process of annotating a dataset of biomedical journal articles according to the annotation scheme of $\text{NoReC}_{\text{neg}}$ and our additional assumptions. Statistics of the annotated dataset, which we refer to as $\text{NorMed}_{\text{neg}}$, are provided and discussed.

Chapter 6 presents the results of applying models from chapter 3 to the annotated dataset from chapter 5, including an error analysis. It also contains an attempt to view our models and annotation scheme in a medical or clinical context, accompanied by preliminary results of a further fine-tuning of the models on parts of $\text{NorMed}_{\text{neg}}$.

Chapter 7 summarizes the key findings and contributions of this thesis, and provides recommendations for future work.

Chapter 2

Background

The purpose of this chapter is to introduce the topics of *negation in natural language processing* and *clinical natural language processing* and thus establish the academic context of this thesis. In 2.1, we provide an overview of negation annotation and available datasets. Furthermore, a summary of relevant modeling approaches is given. The part on clinical NLP, section 2.2, discusses characteristics of clinical text and applications of NLP in a clinical context. Our two main themes are brought together in section 2.3, where the topic of negation in clinical NLP is further detailed, focusing on Norwegian. Finally, section 2.4 connects the theoretical background to the subsequent chapters of this thesis.

2.1 Negation in NLP

As mentioned in chapter 1, recognizing negation correctly is important in many NLP tasks. A prerequisite for this development is labeled data. We therefore present common approaches to marking negation in text and provide a brief overview of resources annotated with respect to negation. Subsequently, we explore the development in the field of negation modeling, including recent advances.

2.1.1 Negation datasets

When working with negation in NLP, it is essential to have access to annotated data. Years of research in the field has resulted in numerous negation datasets. Jiménez-Zafra et al. (2020) provide an overview of these datasets, their similarities and differences. Among the criteria they use for comparison are *negation components* and *negation types*.

They identify four negation components, each related to a specific task: (i) negation cue detection, (ii) negation scope identification, (iii) negated event recognition and (iv) negation focus detection. A *cue* is the negation marker, and its *scope* is the span of text whose meaning is affected by the presence of the cue. A negated *event* denotes the event (in most cases a verb, a noun or an adjective) that the cue directly negates. *Focus* refers to the part of the scope on which the negation cue acts most strongly.

As this thesis will focus on cue detection and scope identification, these two concepts are illustrated with examples, see (2.1) - (2.5), which will be discussed shortly. The sentences are obtained from the Norwegian $\text{NoReC}_{\text{neg}}$ dataset (Mæhlum et al., 2021). Negation cues are in boldface and their scopes inside brackets.

Jiménez-Zafra et al. (2020) compare the datasets according to which of the mentioned components have been annotated. Their findings show that none of the four components (cue, scope, negated event, focus) are present in all datasets. Cue and scope are most commonly annotated. Some datasets contain event annotations, either alone or in addition to both cue and scope. All of the four datasets annotating focus also annotate cue, and some of these also mark negation scope.

Furthermore, they describe three negation types: *syntactic*, *lexical* and *morphological* (or *affixal*) negation. In the syntactic case, the negated meaning is conveyed by a syntactically independent unit. An instance of this can be seen in (2.1), where the cue is the sentence adverb *ikke* 'not'. In Norwegian text, syntactic negation also typically occurs with cues such as *aldri* 'never', as can be seen in (2.2), *ingen* 'no' and *uten* 'without'. Lexical negation is present when an aspect of a word's meaning is negative, as with *fraværende* 'absent' in (2.3). In morphological negation, the source of negation is a morpheme or an affix, i.e. a unit below the word level. (2.4) is an example of morphological negation with the suffix *-fri* '-free/-less'. (2.5) contains an example with the prefix *u-* 'un-'. Another common affixal negation marker in Norwegian is the suffix *-løs* '-less'.

(2.1) *Vi har kjempet så hardt for dette , [jeg har]* ***ikke*** *[ord]*
We have fought so hard for this , I have not words
'We have fought so hard for this, I have no words'

(2.2) *[Eva Mendes er]* ***aldri*** *[bedre enn middelmådig i sine filmer] .*
Eva Mendes is never better than mediocre in her movies .
'Eva Mendes is never better than mediocre in her movies.'

(2.3) *... men [magien er] fullstendig* ***fraværende*** *.*
... but magic.the is completely absent .
'... but there is no magic to it at all.'

(2.4) *Filmen ... er helt [feil]****fri*** *, etter hva jeg kan se .*
Movie.the ... is wholly flaw.free , after what I can see .
'From what I can see, the movie is completely flawless.'

(2.5) *Boka er* ***u****[vanlig] fint illustrert .*
Book.the is unusually nicely illustrated .
'The book is unusually nicely illustrated.'

All datasets described by Jiménez-Zafra et al. (2020) where information about negation types is available include, either completely or partially,

syntactic negation. Three contain only lexical negation in addition to this, while the remaining datasets are annotated with morphological negation as well. Some of the datasets cover the negation types only partially.

Another dimension of comparison is which domain the documents of a dataset belong to. Vocabulary is domain-specific, and thus the domain of a dataset will limit its area of use (Jiménez-Zafra et al., 2020). Their review clearly shows that the biomedical and clinical domain dominates. This can probably be explained by the importance of identifying the presence or absence of negation in medical records, which will be discussed further in 2.3. There are also several datasets built from review articles.

Out of the datasets discussed by Jiménez-Zafra et al. (2020), most are in English. There are also several Spanish datasets and a few in other languages, but there were none available in Norwegian at this point in time.

NoReC$_{neg}$ – a negation dataset for Norwegian

The publication of NoReC$_{neg}$ (Mæhlum et al., 2021) marks a step forward for research in NLP and negation in Norway, as this is the first annotated negation dataset for the Norwegian language. NoReC$_{neg}$ consists of 11,346 sentences, where about 20 % are subject to negation. The sentences are distributed across 414 documents, all of these being professional review articles from various domains. These are a subset of the larger Norwegian Review Corpus (NoReC), created by Velldal et al. (2018).

According to the criteria used by Jiménez-Zafra et al. (2020), NoReC$_{neg}$ annotates cue and scope, but neither negated event nor focus. All the negation types (syntactic, lexical and morphological) are annotated. The main inspiration for the NoReC$_{neg}$ annotation guidelines are the SFU Corpus, a review dataset with an English (Konstantinova et al., 2012) and a Spanish part (Jiménez-Zafra et al., 2018), and the ConanDoyle-neg corpus (Morante and Daelemans, 2012), which consists of stories written by Arthur Conan Doyle.

Unlike SFU, ConanDoyle-neg and NoReC$_{neg}$ include morphological negation. Another point of difference is with respect to the inclusion of subjects in the span of the scope. Here, NoReC$_{neg}$ follows the standard of ConanDoyle-neg, which does consider subjects as part of the scope. Due to this, discontinuous scopes occur frequently in NoReC$_{neg}$, as subjects will usually precede a common negation cue like *ikke* 'not', while the rest of the negated proposition follows after it. In neither of the three datasets is the negation cue defined as part of the scope, in contrast to the BioScope corpus (Vincze et al., 2008), upon which both the SFU Corpus and ConanDoyle-neg build. Worth mentioning is also that NoReC$_{neg}$ annotates multi-word cues but omits prepositions in such constructs, in contrast to the ConanDoyle-neg corpus. Another difference between these two is that NoReC$_{neg}$ also annotates negation in non-factual sentences. The complete annotation guidelines can be found on GitHub.[1]

[1] https://github.com/ltgoslo/norec_neg/blob/main/annotation_guidelines/guidelines_neg.md (Per May 13, 2023, these were last updated on Jun 1, 2021.)

The inter-annotator agreement for cues measured by Mæhlum et al. (2021) is very high (F_1 0.995, κ 0.841). As for the scope annotation task, there is more variation between the annotators. These scores are lower, but depend on the perspective; the token-level overlap is naturally higher (F_1 0.912, κ 0.803) than the score counting exactly matching scopes (F_1 0.632, κ 0.34).

2.1.2 Negation modeling

Morante and Blanco (2021) summarize the advances made with respect to negation processing in recent years. Based on their paper, we present a brief overview of notable efforts in the field. As they emphasize, the work on negation in NLP has to a large extent focused on scope resolution rather than tasks like focus identification, which is outside the scope of this thesis and will not be discussed further. We also want to make it clear that we refer to the task of identifying negation cues as *cue detection*. For the task of determining the correct negation scopes, we use the terms *scope resolution* and *scope identification* interchangeably.

In the early days, different rule-based approaches dominated. Among the most prominent ones is the NegEx algorithm (Chapman et al., 2001), which has proven to work well for clinical text. It has been subject to further development and attempts at improvement also more recently (Elazhary, 2017; Harkema et al., 2009; Mehrabi et al., 2015).

The publication of the BioScope corpus (Vincze et al., 2008), with annotations of negation scope, made the task of negation processing more open to machine learning approaches. Morante and Daelemans (2009) created a system combining several supervised machine learning classifiers that outperformed earlier negation scope identification systems. Other machine learning approaches have involved Conditional Random Fields (CRFs) (Agarwal and Yu, 2010; Reitan et al., 2015) and shallow semantic parsing (Li et al., 2010). There have also been scope resolution experiments using SVMs to rank syntactic constituents (Read et al., 2012) and combinations of shallow and deep methods (Velldal et al., 2012).

Starting around 2016, the field of negation scope identification seems to shift in the direction of deep neural networks. Qian et al. (2016) used convolutional neural networks (CNNs), while Ren et al. (2018) used recursive neural networks. Fancellu et al. (2016) tried both a feed-forward neural network (FFNN) and a bidirectional LSTM (Bi-LSTM), and Lazib et al. (2018) used a hybrid of a CNN and a Bi-LSTM.

Among those not covered by Morante and Blanco (2021) is the paper by Kurtz et al. (2020). The authors approached cue detection, scope resolution and negated event recognition as a graph-based problem extending an existing dependency graph parser (Dozat and Manning, 2018). For training and testing, they used two different versions of the ConanDoyle-neg corpus (Morante and Daelemans, 2012). Two sets of experiments were conducted, one where gold cues were utilized in the prediction of negation scopes and negated events, and another where cues were predicted as well. With the former method, they outperformed comparable systems

with respect to the strict FN (Full Negation F_1) score, which in this case requires predicted scopes and events to match the gold standard exactly. For the latter method, the previous state-of-the-art (Read et al., 2012) was outperformed with respect to token-level F_1-score for scopes, and the previously mentioned FN score. In terms of token-level scope F_1-score their system is also better than a rule-based system created by Packard et al. (2014).

Mæhlum et al. (2021) further explored the results of Kurtz et al. (2020) and tested several variations of dependency graph architectures on the $\text{NoReC}_{\text{neg}}$ dataset. Their approach predicts negation cues and scopes only, and the dataset and results they present serve as a benchmark for negation resolution in Norwegian.

Khandelwal and Sawant (2020) introduced one of the most influential models of recent date, which was the result of a transfer learning approach with BERT[2] (Devlin et al., 2019). Their model, NegBERT, uses BERT in two rounds, first for cue detection and then for scope resolution. For both tasks, a classification layer is added on top of BERT, and it is fine-tuned on the training set. During training, the cue detection system is fed sentences where tokens are annotated as affixal cues, normal cues, part of multiword cues or not a cue. The scope identification part of the system is trained on sentences where the cue is annotated and each token is labeled either 0 or 1 according to whether or not it is part of the cue's scope.

The NegBERT (Khandelwal and Sawant, 2020) authors trained several instances of their architecture using training sets from various corpora belonging to different domains: SFU (Konstantinova et al., 2012), Sherlock (Morante and Blanco, 2012), which appears to be another name for ConanDoyle-neg (Morante and Daelemans, 2012), and BioScope (Vincze et al., 2008). They tested each model on test sets from all these corpora to look at the ability of cross-domain generalization. For scope resolution, their results outperformed the previous state-of-the-art token-level F_1-score by a large margin (Sherlock: F_1 92.36, BioScope Abstracts: F_1 95.68, BioScope Full: 91.24 F_1, SFU: 90.95 F_1). In most datasets, however, NegBERT performed substantially poorer with respect to cue detection than the current state of the art. The authors explain this by the need for more training data in a model of this size and complexity. In terms of domain portability, they conclude that the results are not bad, but better results are desirable.

Negation modeling has also been addressed with multitask learning, by jointly training cue detection and scope resolution models for negation and speculation, respectively (Khandelwal and Britto, 2020). Here, BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2019) were used as language models. The models were trained on combinations of BioScope Full Papers, BioScope Abstracts and SFU. For the scope resolution task, the models trained on both negation and speculation outperformed the models trained on the negation task alone, as well as the

[2]We will not detail the architecture of BERT in this thesis and refer to Devlin et al. (2019) for an elaboration on this.

previous state-of-the-art results on all datasets used for testing (BioScope Full: 97.40, BioScope Abstracts: 97.06, SFU: 93.19 F_1).

Note the distinction between models that predict both cues and scopes, and models predicting only the scopes, as pointed out by Mæhlum et al. (2021). Predicting both must be considered a more difficult task than using the true cues as the basis for scope prediction. This distinction makes a comparison of results challenging. Among the works we have discussed, Kurtz et al. (2020) report results for both methods, and Mæhlum et al. (2021) use predicted cues. Fancellu et al. (2016) and Qian et al. (2016) follow the second method, predicting only the scopes. In the cases of Khandelwal and Sawant (2020) and Khandelwal and Britto (2020), cues and scopes are predicted independently. Their reported scores for scope identification seem to be based on gold cues.

2.2 Clinical NLP

As defined by Névéol et al. (2018), clinical NLP is a subfield of NLP, characterized by NLP being applied to clinical text or with a clinical outcome as the objective. As a central part of this field, they mention the application of such techniques to texts in Electronic Health Records (EHRs). Other focus areas are developing resources relevant to clinical NLP systems, research in biomedical information retrieval and the analysis of patient-produced text aiming at clinical purposes.

2.2.1 Clinical text

Clinical text is a fundamental building block for clinical NLP. The following is based on the textbook *Clinical Text Mining* by Hercules Dalianis (Dalianis, 2018), which describes the important characteristics of clinical text.

The term is used for the text contained in health record systems and can also be referred to as *electronic patient record text*. Medical professionals such as doctors and nurses are both the authors and the intended readers of these documents. Because they work in a highly specialized domain, patient records tend to contain a large number of foreign words and terminology that is unknown to the average person without any medical expertise. Many of these terms originate in Latin and Greek, but the spelling is often to some degree influenced by the morphology of the target language.

Among the domain-specific words are also various abbreviations and acronyms, a specialized type of abbreviation combining certain letters from the individual words of a phrase. The use of abbreviations is interesting because they may give rise to ambiguity. Studies on Swedish clinical text have found abbreviation rates ranging from 1 % to 14 % (Allvin et al., 2011; Isenius, 2012; Isenius et al., 2012; Nizamuddin and Dalianis, 2014; Olsson, 2011; Skeppstedt et al., 2012). Other studies have found that around 1/3 of abbreviations are ambiguous in English (Liu et al., 2001) and Swedish (Lövestam et al., 2014) clinical text.

Another important contextual factor is the stressful working environment of many health professionals. Time pressure results in a high frequency of spelling mistakes in patient records, with percentages ranging from 1.1 % (Grigonyte et al., 2014) to 7.6 % (Nizamuddin and Dalianis, 2014) in Swedish studies. One will also find that the syntactic structure of the text differs from other text types. Short and incomplete sentences are common. This might be in terms of lacking subjects and lacking helping verbs ('to be'). Time pressure might be one reason for this, however it is reasonable to believe it is also due to a desire to facilitate efficient reading.

Clinicians use health records to document symptoms and findings present in the patient. In addition, they describe which symptoms are absent, and negative clinical findings. Both are important in providing a basis for clinical decision-making, which will hopefully lead to a correct diagnosis. For this reason, negated expressions occur frequently in clinical text. In a study on Swedish clinical text concerning assessment of patients, Dalianis and Skeppstedt (2010) found that 13.5 % of the texts consisted of negated expressions or sentences.

While some expressions are negated, others are associated with a level of uncertainty. This may range from 'probably negative' to 'probably positive' and can be used to express the level of certainty of a diagnosis etc.

Taking all these characteristics of clinical text into account, it seems probable that clinical text differs greatly from review texts such as those found in $\text{NoReC}_{\text{neg}}$ (Mæhlum et al., 2021). One would expect the vocabulary of review texts to be influenced by words specific to the domain of the review object. It is also to be expected that review texts contain mostly complete sentences and show more linguistic creativity compared to clinical text. The frequencies of abbreviations and spelling errors will likely be lower in texts that have undergone a process of quality assurance before publication.

2.2.2 Access to clinical data

In Norway, accessibility and storage of personal data is regulated by the General Data Protection Regulation (GDPR)[3]. Helseforskningsloven[4] *(eng: The Health Research Act)* addresses research involving personal health data. In order to use such data for research purposes, voluntary, specific and informed consent from the research object is required (Helseforskningsloven, 2008, § 13). If the data are anonymous, consent is not necessary (Helseforskningsloven, 2008, § 20). However, the anonymization of unstructured data is difficult and an active field of research. For text to be considered anonymous, one must remove all information that may lead to the identification of individuals. It is not sufficient to eliminate direct identifiers, as combinations of so-called quasi-identifiers also may lead to identification (Lison et al., 2021). One can easily imagine this as a difficult task in health

[3] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679
[4] https://lovdata.no/dokument/NL/lov/2008-06-20-44

records, which contain information about age, sex, place of residence, family and potentially rare diagnoses.

In addition, all health research projects must be approved in advance by a regional committee for medical and health research ethics (Helseforskningsloven, 2008, § 9). These committees have the power to decide whether the requested processing of personal health data is medically and ethically sound (Helseforskningsloven, 2008, § 34).

2.2.3 Research areas and applications of clinical NLP

Despite challenges with access to clinical data, various clinical datasets and other clinical NLP resources do exist. Much of the research addresses the English language. However, Névéol et al. (2018) show that among clinical NLP publications on PubMed, one also finds work targeting languages such as French, German, Chinese, Japanese, Spanish, Dutch and Swedish. As the development of lexicons, annotated corpora and other resources is a prerequisite for other clinical NLP tasks, it is no surprise that their overview of publications shows that this is a priority.

Furthermore, there are several publications on de-identification of clinical text. Improving these methods could make it easier to get access to clinical data for research in the future. De-identification of the structured data of patient records is quite straight-forward; the problem is the presence of personal information such as names and addresses in free text (Dalianis, 2018).

The field of clinical NLP includes many tasks. The overview provided by Névéol et al. (2018) regarding publications in other languages than English include tasks related to information extraction. Here, examples are extraction of medical entities such as symptoms and findings (Named Entity Recognition), events such as adverse drug events and relations between entities. Translation of clinical text and various text classification tasks are also represented. A practical example of a classification task is the automatic assignment of ICD-10 diagnosis codes to patient records (Dalianis, 2018). Furthermore, there is detection of negation and speculation, which as discussed is frequent in clinical text, and the analysis of temporal expressions. Dalianis (2018) also emphasize text summarization and simplification of patient records as important tasks.

2.3 Negation in clinical NLP

In clinical NLP, one is interested in automatically processing patient records to obtain information on whether the symptoms, findings or diagnoses mentioned are present or not. Here, negation resolution is used to identify the entities that are absent. Note that we use the term 'negation resolution' as a reference to the general task of detecting negation in text and determining what is negated.

Many of the approaches for recognizing negation in clinical NLP are simple rule-based models or extensions of such, and these have proven

to work quite well on clinical text (Chapman et al., 2001; Elazhary, 2017; Harkema et al., 2009; Mehrabi et al., 2015). As we have seen, however, the general performance of negation resolution in NLP has improved with the introduction of neural machine learning methods. We therefore believe that applying such methods to clinical and medical text has a potential as well. From 2.1.2 we remember that Khandelwal and Sawant (2020) and Khandelwal and Britto (2020) showed promising results on the parts of the BioScope (Vincze et al., 2008) corpus consisting of biomedical text.

2.3.1 Negation in clinical NLP for Norwegian

In a global perspective, Norwegian is a small language with approximately 5.5 million speakers. Thus, the size and number of Norwegian NLP communities cannot be compared to those working with world languages such as English. Furthermore, the strict regulations of personal health data in Norway represents a challenge for the development of the clinical subfield.

As a consequence, there has been limited research in clinical NLP for Norwegian, including the task of negation resolution. Based on the similarities between the two languages, Budrionis et al. (2018) made an attempt to port the Swedish NegEx algorithm (Skeppstedt, 2011) to Norwegian. Due to the lack of a final gold standard, they were not able to properly evaluate their results.

Sadhukhan (2021) continued the work on a Norwegian NegEx in their master's thesis. Their system was tested on a corpus of biomedical articles. While the Swedish version of the algorithm has shown good results with F-scores around 0.8 (Tanushi et al., 2013), their NegEx system for Norwegian performed poorly, achieving an F-score of only 0.55.

Sadhukhan (2021) use a list of Norwegian medical terms, as NegEx is dependent on a predefined list of entities for which to decide whether they are negated or not. Their system also utilizes a list of 80 Norwegian negation triggers. For each pre-identified medical entity in an input sentence, the algorithm uses regular expressions to detect whether the entity is inside the scope of a negation trigger, where *scope* is defined as a distance of maximally 6 tokens.

An interesting result of their efforts is the *Norwegian GastroSurgery Biomedical Negation Corpus*[5], which is based on a dataset of medical articles from the domain of gastrointestinal surgery originally collected by Budrionis et al. (2018). The corpus consists of 2,330 sentences. With a total of 48 negated terms, negation is quite infrequent.

For Swedish, there have been experiments with other rule-based negation resolution algorithms in addition to NegEx, all providing F-scores close to 0.8 (Tanushi et al., 2013). To the best of our knowledge, there are no publications on clinical negation resolution with *neural* methods or other machine learning approaches for neither of the Scandinavian languages.

[5] https://github.com/DebaratiSJ/NegEx-on-Norwegian-biomedical-text/blob/main/Gold%20standard%20biomedical%20corpus/Norwegian%20GastroSurgery%20Biomedical%20Negation%20Corpus.txt

2.4 Conclusion

As previously mentioned, the scope of this thesis will be negation resolution in text from the medical domain. For our models, we will use the $\text{NoReC}_{\text{neg}}$ negation dataset (Mæhlum et al., 2021) as the basis for training. We want the focus of this work to be the medical perspective on negation and not on negation modeling in general. Hence, we will train relatively simple models based on sequence labeling in contrast to graph-based architectures. This, however, involves state-of-the-art technology such as BERT (Devlin et al., 2019).

To evaluate the performance of our models in a medical context, we need annotated data from the medical domain. For this reason, we will reannotate the Norwegian GastroSurgery Biomedical Negation Corpus (Sadhukhan, 2021) using the annotation scheme of $\text{NoReC}_{\text{neg}}$.

We are aware that biomedical journal articles and patient records represent different domains and that good performance of negation models in the former does not necessarily translate to the latter. Among the assumed differences, we find a high frequency of spelling errors, abbreviations, and incomplete sentences in clinical text (Dalianis, 2018), which is probably much less common in carefully edited, published articles. Still, we have reason to believe that clinical text and biomedical articles resemble each other more than clinical text and review articles in terms of vocabulary, the use of medical terminology, and thematically.

The annotation of biomedical text according to the $\text{NoReC}_{\text{neg}}$ standard will allow us to examine the portability of models trained on review articles into the medical domain and identify possible challenges in this process. Our hope is that these results, both in terms of the annotated dataset and the models, can be of value for the identification of negation in Norwegian clinical text as well, given the similarities between the clinical and biomedical domains.

Chapter 3

A sequence labeling system for negation resolution

In this chapter, we train models for identifying negation in Norwegian text. This is done in two steps; first, negation cues are detected, and then negation scopes are resolved for the detected cues. We approach both tasks as sequence labeling problems. Initially, we test different variations of the scope resolution model. The best configuration is used as a basis for another round of experiments, where the goal is to investigate the effect of different transformer-based language models on the performance of the negation resolution system. To evaluate our systems, we use the metrics from the *SEM 2012 Shared Task (Morante and Blanco, 2012). In both stages of experimenting, we perform a quantitative and qualitative error analysis of the best system and report our findings.

3.1 General experimental setup

We approach the task of negation resolution as a sequence labeling problem. This method is widely used for various tasks in the field of NLP, such as named entity recognition (Devlin et al., 2019; Lample et al., 2016) and sentiment analysis (Li et al., 2019; Luo et al., 2020). Although previous work has shown that the use of a graph-based architecture can be beneficial in negation resolution (Kurtz et al., 2020), sequence labeling still represents an approach that is both simple and has produced convincing results for other languages. One such example is the system known as NegBERT (Khandelwal and Sawant, 2020), which has strongly inspired our approach. As previously described in 2.1.2, it is composed of two separate models for cue and scope identification, each consisting of BERT (Devlin et al., 2019) with a classification layer on top (Khandelwal and Sawant, 2020).

In our modeling experiments, we use the same bipartite model structure in order to create a system that can predict negation scopes without being fed the correct negation cues as input. A schematic overview of the system is found in Figure 3.1 on the following page. All our models use Norwegian or multilingual transformer embeddings and a final linear classification layer.

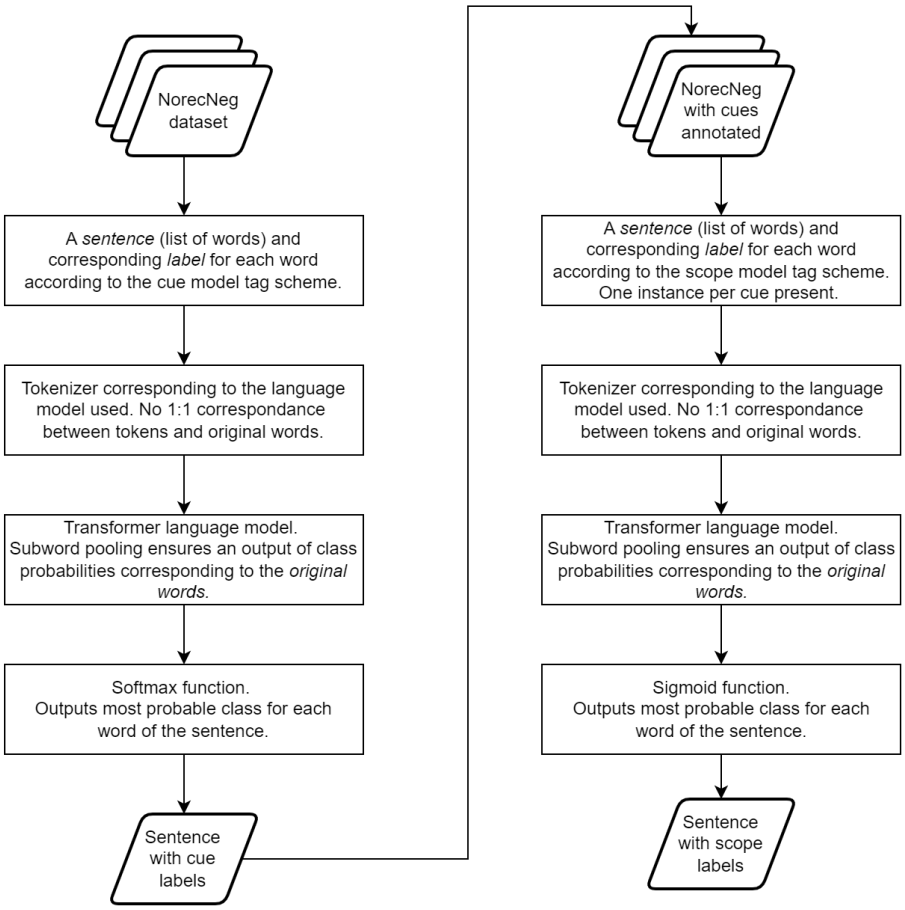

Figure 3.1: An overview of the data flow, inputs and outputs of the system.

Table 3.1 on the next page lists the training and model settings that are common to both cue and scope model throughout all our experiments. The differences between the models used for cue detection and scope resolution will be discussed in subsections 3.1.1 and 3.1.2. Unless otherwise stated, the reader can assume them to be identical in terms of all other architectural aspects, hyperparameter values and other settings.

In section 3.2, we experiment with a selection of small variations to the scope resolution model. In section 3.3, we take the best system from the initial experiments and investigate the effect of the language model used, as well as the effect of the batch size used in training.

All our models are trained on the Saga supercomputer provided by the Norwegian Research Infrastructure Services (NRIS).[1] For information on the requirements, we refer to our GitHub repository.[2]

[1] https://documentation.sigma2.no/index.html
[2] https://github.com/marieef/master-thesis_code

Model settings	
initial learning rate	3e-5
drop last	True
optimizer	AdamW
scheduler	cosine schedule with warm-up
warm-up steps	200
freeze (BERT)	False (requires_grad = True)

Table 3.1: Values of hyperparameters and other settings common to all models.

Norwegian	Direct translation	Cue label	Scope label
Boka	Book.the	3	0
er	is	3	0
uvanlig	unusually	**0**	**1**
fint	nicely	3	0
illustrert	illustrated	3	0
.	.	3	0

Table 3.2: An example of how a sentence containing an affixal cue will be tagged according to our tag scheme.

3.1.1 Cue detection model

We follow the descriptions of Khandelwal and Sawant (2020), using four classes:

0: AFFIX (affixal cue)
1: NORMAL (regular cue)
2: MULTI (part of a multi-word cue)
3: NOT_CUE (not part of a cue)

In addition, there is a fifth label used to pad sequences of varying length, but these are ignored during training. The other labels are weighted equally.

In the training phase, the model is fed a sentence as a list of words, and the corresponding list of gold labels according to the labeling scheme above. We feed each sentence in the training data into the model once per epoch, training the model to find all negation cues in the sentence in one step.

As mentioned, we treat cue detection as a multiclass problem. Therefore, we use categorical cross entropy loss as the loss function (Khandelwal and Sawant, 2020).

3.1.2 Scope resolution model

Our approach to resolve negation scopes follows the descriptions of the NegBERT paper (Khandelwal and Sawant, 2020) as well. The aim of the

Norwegian	Direct translation	Cue label	Scope label
en	a	3	**1**
svært	very	3	0
uvanlig	unusual	**0**	**1**
blanding	mix	3	**1**
av	of	3	0
jazz	jazz	3	0
og	and	3	0
hinduistisk	hinduistic	3	0
inspirert	inspired	3	0
åndelig	spiritual	3	0
musikk	music	3	0
.	.	3	0

Table 3.3: An example of how a part of a sentence containing an affixal cue will be tagged according to our tag scheme.

Norwegian	Direct translation	Cue label	Scope label
Jeg	I	3	**1**
vet	do.know	3	**1**
ikke	not	**1**	0
hva	what	3	**1**
som	that	3	**1**
skjedde	happened	3	**1**
,	,	3	0
jeg	I	3	0
bare	just	3	0
skjøt	shot	3	0
.	.	3	0

Table 3.4: An example of how a sentence containing a 'normal' cue will be tagged according to our tag scheme.

Norwegian	Direct translation	Cue label	Scope label
Verken	Neither	**2**	0
letthet	lightness	3	**1**
,	,	3	**1**
humor	humor	3	**1**
eller	nor	**2**	0
engang	even	3	0
avsky	disgust	3	**1**
.	.	3	0

Table 3.5: An example of how a sentence containing a multi-word cue will be tagged according to our tag scheme.

scope resolution model is to correctly predict, for each word in an input sentence, whether it is part of the scope of one specific negation cue in the sentence.

To do this, the model needs to know which word(s) constitute the cue whose scope is to be resolved. In order to signal this to the model, we use special tokens of the form 'token[X]', where X is replaced with the type of cue (0, 1 or 2 according to the scheme above). This can be done in two ways; either by replacing the cue word in the sentence by the special token, or by inserting the special token directly in front of the cue word in the input sentence (Khandelwal and Sawant, 2020). We experiment with both methods.

Sentences can contain more than one negation. In these cases, the sentence is fed to the model once for each negation cue.

Because scope resolution at the token level is a binary task, two labels are used:

0 (outside scope)
1 (inside scope)

Also for this task, a third label is used for padding. The predictions for these labels do not contribute to the loss. Due to the binary nature of the task, we use binary cross entropy loss as the loss function.

3.1.3 The negation resolution system as a whole

The cue and the scope model are trained independently on the predefined training portion of the $\text{NoReC}_{\text{neg}}$ dataset (Mæhlum et al., 2021) and tested on the corresponding development test set.[3]

Like NegBERT (Khandelwal and Sawant, 2020), our models output one label per word in both cue detection and scope resolution. Tables 3.2 to 3.5 on pages 17–18 contain examples that illustrate the tag scheme.

Limitations of the system

Affixal cues The dataset we use contains annotations of *affixal* cues, i.e. subtoken cues. Thus, in these cases, we make a simplifying assumption when we train our models to label the *whole* word as a cue. Because the remaining part of a word containing a negation affix is contained within its scope, our scope resolution models are trained to label the whole word as part of the scope. In this way, one gets an overlap between cue and scope in the cases of affixal negation. An illustration of this can be seen in Tables 3.2 to 3.3 on pages 17–18. According to the annotation guidelines[4] of $\text{NoReC}_{\text{neg}}$ (Mæhlum et al., 2021), avoiding overlap between a cue and its scope might be an advantage for modeling. Our approach violates this principle.

To get predictions on sub-word level, one possibility is to postprocess the raw model predictions. This can be done by matching the words

[3] https://github.com/ltgoslo/norec_neg/tree/main/data

[4] We refer to the latest version per May 13, 2023 (from Jun 1, 2021): https://github.com/ltgoslo/norec_neg/blob/main/annotation_guidelines/guidelines_neg.md

predicted as affixal cues against known affixal negation cues. In this way, a fair evaluation against the original gold standard is possible. We will address this limitation in 3.1.4.

Words containing multiple cues Since our model assigns one label to each word, it assumes that every word contains at most one cue. In the hypothetical case of multiple cues inside the same word, only one of them will be extracted to create a training example for cue detection. The only example we could think of are possible words such as *ikke-tankeløs* 'not-thoughtless'. Based on our experience with Norwegian text, such words are not commonly used, and we have checked that there are no such cases in the dataset.

Multi-word cues Multi-word cues pose a challenge for our models. For each cue present in a given sentence, the scope resolution model needs a separate copy of the sentence with this specific cue annotated. This applies both to training and inference.

The problem does not apply to cases such as the one shown in Table 3.5 on page 18, but arises when the model encounters an input sentence containing more than one multi-word negation cue. In the training phase, the model uses the annotated training set to derive which of the words tagged as parts of multi-word cues belong together. However, when applying the model to test data, this information is not available. As our simple, sequential model does not carry the syntactic information one would need to resolve this issue, we choose to treat all predicted MULTI labels of a sentence as if they belong to the same cue.

An alternative would be to use a rule-based approach, i.e. to add heuristics utilizing knowledge about multi-word negation cue patterns in Norwegian. We consider this to be outside the scope of our simple modeling experiments.

Choice of language model For our initial experiments, the choice of language model is fixed. We are fully aware that this is not necessarily the optimal choice of language model for our system. The performance of the system might benefit from choosing a larger language model trained on more data. We shall return to this in 3.3.

3.1.4 Evaluation

Throughout all our experiments, evaluation scores are reported as the average value of 5 runs, corresponding to a set of 5 random seeds. As for the evaluation metrics, we choose to adhere to the standard set by the *SEM Shared Task 2012 (Morante and Blanco, 2012) on resolving the scope and focus of negation. We report the following three scores for all our models:

Cue-level F_1-score (CUE) For a cue predicted by the system to be counted as a true positive, there has to be an exact match with the gold cue.

Scope Token F_1-score (ST) This metric operates on the token level. It calculates the total number of tokens included in each scope, and if a scope token produced by the system exactly matches a gold scope token, this counts as a true positive. For a scope token to be counted as correct, there must at least be a token of overlap between the gold cue and the predicted cue.

Full Negation F_1-score (FN) For a full negation to be counted as a true positive, both the cue and the scope prediction is required to match the gold standard exactly. Note that in the shared task of *SEM 2012 (Morante and Blanco, 2012), there is an additional requirement for the predicted negated event to be correct. Since event prediction is outside the scope of this thesis, we disregard this condition.

Adjusting the gold annotations to span whole words

As previously explained, our models assign one label to each word of an input sentence in both the cue detection and the scope resolution phase. The $\text{NoReC}_{\text{neg}}$ gold standard we compare our results to contains annotations for affixal negation cues spanning across *parts* of words. Our models would never be able to directly identify the correct span of the negation cue and scope in a case such as the one shown in example (3.1), where the gold cue is marked in bold and the gold scope within square brackets.

(3.1) *Boka* *er* ***u****[vanlig]* *fint* *illustrert* .
Book.the is unusually nicely illustrated .
'The book is unusually nicely illustrated.'

(3.2) *Boka* *er* *[**uvanlig**]* *fint* *illustrert* .
Book.the is unusually nicely illustrated .
'The book is unusually nicely illustrated.'

Our first approach to handle this challenge is to evaluate against a simplified gold standard. Here, all annotations spanning only parts of a word are replaced by the word as a whole. In this adjusted gold standard, the annotations for the sentence in example (3.1) would look as in example (3.2). This represents the same type of gold standard as the one used in model training. The reader should note that this is a simplification of the original gold standard, and these results will therefore not be directly comparable to those reported in Mæhlum et al. (2021). Furthermore, note that the results during our initial development are incomparable to theirs as we operate on the development test set, while they report results on the held-out test set only. Although simpler, we found it more suitable to evaluate how well the models learn to do what they have actually been trained for. Even so, we will report the results on the original gold standard as well, as it could be of interest to quantify the difference in performance on the adjusted and original gold standard. The script used to convert the

original gold standard to a version spanning whole words can be found in our GitHub repository.[5]

Affix matching in evaluation against the original gold standard

In spite of the fact that our models operate on word units, we would like to be able to fairly evaluate the results for the original gold standard with its subtoken annotations as well. In order to do this, we run the initial predictions of our models through a rule-based module. Here, we use regular expressions to recognize patterns of Norwegian negation affixes, inspired by descriptions in earlier work (Kurtz et al., 2020; Lapponi et al., 2012). The patterns we use are obtained from code contained in the $\text{NoReC}_{\text{neg}}$ repo.[6] We match these patterns against those predicted cues that were labeled as affixal cues by our model. If a match is found, we replace the original cue prediction with the identified affix. If the model initially predicted this word to be part of the scope as well, we replace this scope prediction with the rest of the word (i.e. the word without the part that was identified as a negation affix.)

Since this enables a fair comparison to the original gold standard, we will consider this our main evaluation method. Nevertheless, as mentioned in 3.1.4, we will also include the evaluation of the raw predictions for both the original and the adjusted gold standard. When reporting the results of our experiments, we will comment on differences in metric scores between the three methods if considered relevant.

3.2 Initial experiments

In this section, we describe our first set of experiments. We apply a few variations to the scope resolution model and present the results on the development test set. The errors made by the best system are quantified and analyzed qualitatively.

3.2.1 Model configuration

In this part, unless otherwise stated, the described configurations apply to both the cue and the scope model. Tables 3.1 on page 17 and 3.6 on page 24 list the model settings used in this round of experiments.

Regarding the hyperparameter values chosen, these were partly taken from Khandelwal and Sawant (2020) and the original BERT paper (Devlin et al., 2019) and partly chosen on the basis of being reasonable choices for our task. The learning rate is the same as used in NegBERT. It is within the range of learning rates recommended when fine-tuning BERT (Devlin et al., 2019). We combine it with a warmup phase and a cosine learning rate

[5] https://github.com/marieef/master-thesis_code/blob/main/format_conversion/simplify_gold_standard_sem.py

[6] Our script for extraction of affixes: https://github.com/marieef/master-thesis_code/blob/main/format_conversion/extract_affixes.py

scheduler, as this has been used for other Norwegian sequence labeling tasks.[7] As optimizer, we use AdamW (Loshchilov and Hutter, 2019) as opposed to NegBERT (Khandelwal and Sawant, 2020), where Adam is used. Batch size is set to 32, again according to the recommendations of Devlin et al. (2019). We set number of training epochs to 5 based on the initial observation that the development loss seems not to decrease more after this.

Language model

As reported in Table 3.6 on the next page, we use NorBERT-2 (Kutuzov et al., 2021) as the language model.[8] This version of the NorBERT model is not described in the paper by Kutuzov et al. (2021), but the details of the model are given on the website of The Nordic Language Processing Laboratory (NLPL): NorBERT-2 is developed by the NorLM initiative and trained on approximately 15 billion word tokens from the Norwegian Colossal Corpus (NCC) (Kummervold et al., 2022) and the Norwegian part of the C4 web-crawled corpus (Xue et al., 2021); the training data covers both variations of the Norwegian written language, Bokmål and Nynorsk, and the vocabulary of the model consists of 50,000 WordPiece (Wu et al., 2016) tokens (Nordic Language Processing Laboratory, 2023).

Variations

For our initial experiments, we tested four variations of the scope model, resulting in four slightly different systems. In the following, we elaborate on the aspects of variation:

Cue annotation This parameter refers to how we tell the scope resolution model which cue it is to resolve the scope for. The two strategies used are 'replace' and 'augment' (Khandelwal and Sawant, 2020). In the former, we *replace* a cue word by a special token in the input sentence, as described in subsection 3.1.2. In the *augment* method, we concatenate the special token with the cue word.[9] As an example, *ikke* 'not' becomes *token[1]ikke*.

Training data All models are trained on the predefined training portion of the $\text{NoReC}_{\text{neg}}$ (Mæhlum et al., 2021) dataset. Here, we distinguish between scope models trained on *all* its sentences, and scope models only trained on the sentences actually containing negation.

[7] Generally, we were inspired by code for a related task provided for the students attending the course IN5550 – Neural Methods in Natural Language Processing at the University of Oslo in spring 2022. We adopted the warmup settings, learning rate scheduler and optimizer from this code. This was used as a basis instead of the original NegBERT code because it represents an understandable and coherent code base. All the code for our experiments is available on GitHub: https://github.com/marieef/master-thesis_code.

[8] The model was accessed through Hugging Face: https://huggingface.co/ltg/norbert2

[9] Khandelwal and Sawant (2020) add a space between the special token and the cue, whereas we do not.

Model settings	
batch size	32
epochs	5
language model	ltg/norbert2

Table 3.6: Settings specific to initial experiments.

Model	Cue annotation	Training data
S1	Replace	Neg. sentences
S2	Replace	All sentences
S3	Augment	Neg. sentences
S4	Augment	All sentences

Table 3.7: Specifications of the different scope resolution models tested in the first round of experiments. 'Cue annotation' refers to how cues are input to the scope model: replaced by a special token ('Replace') or through a combination of a special token and the cue itself ('Augment'). 'Training data' is either all sentences in the training set, or only those containing negation.

An overview of the models is provided in Table 3.7. In S1, the scope resolution model uses only sentences actually containing negation for training, and cues are annotated using the *replace* method. The *replace* method is used for S2 as well, but this one is trained on all sentences. S3 uses the *augment* method for cue annotation and is trained on sentences with negation, exclusively. S4 combines the *augment* method with all sentences as training data.

	Original		
	CUE	ST	FN
S1	79.12 (0.27)	80.33 (0.79)	55.07 (0.48)
S2	79.12 (0.27)	80.60 (0.28)	55.00 (1.50)
S3	79.12 (0.27)	81.10 (0.60)	56.37 (0.43)
S4	79.12 (0.27)	**81.61** (0.43)	**57.63** (1.02)

Table 3.8: Results of our four initial architectures when evaluated against the original gold standard of the development test set. The metrics from the 2012 *SEM shared task are used. We report the average across 5 runs.

	Adjusted			Original+RE		
	CUE	ST	FN	CUE	ST	FN
S1	91.71 (0.28)	83.35 (0.77)	61.84 (0.64)	91.71 (0.28)	83.35 (0.77)	61.84 (0.64)
S2	91.71 (0.28)	83.64 (0.30)	61.41 (1.80)	91.71 (0.28)	83.64 (0.30)	61.41 (1.80)
S3	91.71 (0.28)	84.10 (0.58)	62.76 (0.93)	91.71 (0.28)	84.10 (0.58)	62.76 (0.93)
S4	91.71 (0.28)	84.63 (0.47)	63.72 (0.94)	91.71 (0.28)	**84.63** (0.47)	**63.72** (0.94)

Table 3.9: Results of our four initial architectures when evaluated against the adjusted gold standard, and against the original gold standard after affix extraction. The metrics from the 2012 *SEM shared task are used. We use the development test set and report the average across 5 runs.

3.2.2 Results

Tables 3.8 on the preceding page and 3.9 present the results for our different models, with each score averaged over 5 runs. The evaluation metrics are obtained from the *SEM Shared Task 2012 (Morante and Sporleder, 2012) and described in 3.1.4. Table 3.8 contains the results of evaluating the raw predictions against the original gold standard. In Table 3.9, the raw predictions have been evaluated against a gold standard simplified to span whole words (*Adjusted*). Additionally, it contains the results of evaluating the predictions against the original gold standard upon extraction of negation affixes from the word-level predictions (*Original+RE*).

Note that the metric values for *Adjusted* and *Original+RE*, respectively, are identical. As there is no drop in performance in the latter case, we conclude that the affix-matching patterns are able to correctly recognize all negation affixes present in the words labeled as affixal cues by our models. There might of course exist cases not covered in the development test set for which our patterns do not work.

As expected, all metric scores in Table 3.8 on the preceding page are substantially lower compared to the scores in Table 3.9. This is due to the unfair evaluation of word-level predictions against the original gold standard with its subtoken annotations of affixal cues.

Our models S1 through S4, as defined in Table 3.7 on the preceding page, represent all possible combinations of cue annotation method and scope model training data. We observe that S4, where cues are annotated by concatenation with a special cue token and all sentences are included in scope model training, generally performs best. Both systems using the 'augment' method for cue annotation outperform their 'replace' counterparts. This is in accordance with our expectations. It is not surprising that information about the actual word predicted as a cue, and not simply the cue *type*, is valuable to the model. There is also a tendency that the scope models trained on *all* sentences outperform those trained only on sentences containing negation cues. The scope resolution model seems to benefit from exposure to non-negated examples as well.

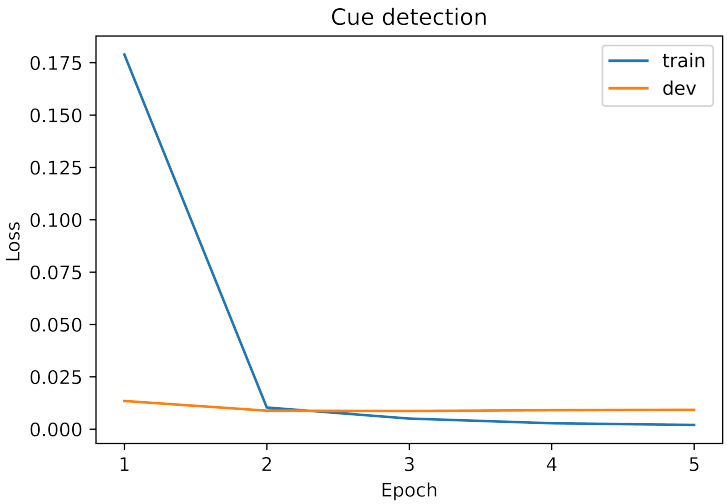

Figure 3.2: Loss as a function of number of epochs for the cue detection model used in S4.

Loss curves

We plot the loss curves of the best system in Figures 3.2 and 3.3 on the facing page. We observe that the development loss actually starts increasing from epoch 3 for the scope resolution model. For the cue detection model, as shown in 3.2, it remains almost flat after epoch 2.

From this, we infer that training for a fixed number of 5 epochs might not be optimal. Thus, we choose to introduce early stopping in both the cue and the scope model. One possibility would have been to use development loss as the stopping metric. However, we choose to adhere to NegBERT (Khandelwal and Sawant, 2020) and use their implementations of cue and scope F_1. These metrics are closely related to our final evaluation metrics (Morante and Blanco, 2012).

We set the upper limit on epochs to 20, as it has been suggested to use this number of epochs when fine-tuning BERT (Mosbach et al., 2021). Like NegBERT (Khandelwal and Sawant, 2020), we use 6 epochs as patience and save the models at the point of the highest F_1 score.

The results from the evaluation on the development test set are presented in Table 3.10 on the next page. As before, raw predictions are evaluated against both the original and the adjusted, word-level gold standard, and predictions matched against affixes are evaluated against the original gold standard. The introduction of early stopping leads to an overall improvement of the results regardless of the evaluation method. There is a small decrease in FN (0.10), but CUE and ST are increased by 0.36 and 0.37, respectively, compared to the previous results in Table 3.9 on the preceding page.

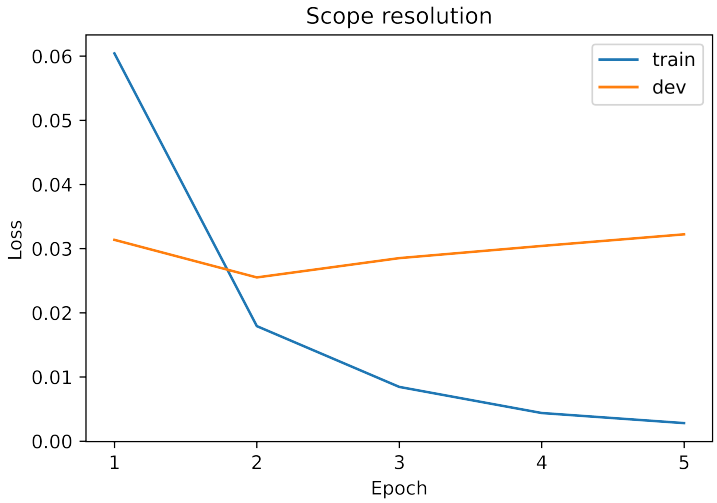

Figure 3.3: Loss as a function of number of epochs for the scope resolution model used in S4.

	CUE	ST	FN
Original	79.45 (0.31)	82.03 (0.27)	57.42 (1.27)
Adjusted	92.07 (0.10)	85.00 (0.23)	63.62 (2.14)
Original+RE	92.07 (0.10)	85.00 (0.23)	63.62 (2.14)

Table 3.10: Evaluation results of early stopping applied to cue and scope model, using F_1 as the stopping metric. Raw predictions are evaluated against both original and adjusted, word-level gold standard. Affix-matched predictions are evaluated against the original gold standard. The metrics from the *SEM Shared Task 2012 are used, and results are averaged across 5 runs.

3.2.3 Error analysis

We perform a quantitative and qualitative error analysis on our best system, i.e. after the introduction of early stopping, and use the development test set for this purpose.

For the analysis, we choose to utilize the predictions that have been run through the regular expression module to split the words predicted as affixal cues into a cue part (the negation affix) and a scope part (the rest of the word). The numbers reported in the following are averaged across five instances of the model trained with different random seeds.

Cue errors

The development test set contains 301 sentences with at least one negation. In total, there are 342 negation cues. On average, our model predicts 312 (312.2) of these correctly, leaving 30 (29.8, 8.7 %) false negatives. In addition, we find 26 (25.6) false positives among the predicted cues (7.58 % of the total number of predicted cues). Our system predicts a total of 338 (337.8) cues on average. Note that for a cue prediction to be correct, we require a perfect match.

False negatives Table 3.11 on page 30 provides an overview of all cues occurring at least twice in the development test set and the frequency by which they are not detected by the model. As expected and according to Table 1 in Mæhlum et al. (2021), the distribution of cues we see in the development test set, is quite similar to the distribution seen in the dataset as a whole.

We note that *u-* 'un-/dis-/non-' is quite frequent, but still relatively often not detected. By manual inspection, we identify false negatives in words such as *uvanlig* 'unusual', *ulike* 'different', *unaturlig* 'unnatural', *ugifte* 'unmarried', *urørt* 'untouched', *unyansert* 'unvarnished', *unødig* 'unnecessary', *umælende* 'speechless' and *uspennende* 'unexciting'.

At least the two latter ones are not commonly used in Norwegian, so this could explain the inability of the model to label them correctly. The common word *ulike* 'different' occurs several times among the false negatives. Perhaps this can be explained by the word occurring frequently as a non-cue as well.

The model also struggles with recognizing the somewhat less frequent suffix *-løs(e)* '-less'. Words with this suffix occur quite frequently as non-cues, too.

For non-affixal, one-word cues, there are generally very few false negatives. *Ingen* 'no'/'nobody' seems to create some confusion, as does the very infrequent words such as *mangelen* 'the lack'.

False positives False positive cue predictions are quantified in Table 3.12 on page 30. For each false positive cue, we report its proportion of the total number of false positive cues. Furthermore, the False Discovery Rate

(FDR)[10] is reported to quantify how frequently a prediction of a given cue represents a false positive.

Unsurprisingly, the frequent *u-* 'un-/dis-/non-' and *ikke* 'not' make up about 50 % of the total false positive cues. The non-affixal cues *ikke* 'not' and *uten* 'without' have the lowest False Discovery Rate as seen in the rightmost column of Table 3.12 on the next page. These results should be seen in the context of the ambiguity rate of the various cues present in the corpus (Mæhlum et al., 2021), as this is an indicator of the potential for false positives. Note that although according to Mæhlum et al. (2021) the ambiguity rate of *uten* 'without' is 0, our system produces a few false positives for this word, in cases where it is not annotated as a cue. The same might be the case for other cues.

Regarding false positive affixal cues, manual inspection reveals a few cases of over-generalizing to words such as *urmenneskelig* 'primordial', *ussel* 'lousy' and *ustrakt* (*utstrakt* 'stretced out' with a typo). We observe some false positives for words containing *u-* 'un-/dis-/non-' or *-løs(t/e)* '-less' that occur as cues in other sentences (*ulik*, 'different' *barnløs* 'childless', *harmløs* 'harmless'). It is possible that some of these are not actual errors, but are caused by incorrect annotations.

We do find some indisputable annotation errors, such as annotating only *ingen* 'no' in *ingen måte* 'no way'. Such cases might confuse the model. We see one case where our model predicts *ingen* and *måte* as separate cues instead of one multi-word cue, and this could be due to inconsistencies in the gold standard.

Scope errors

The development test set contains one scope for each of the 342 annotated cues, out of which 3 are so-called implicit (empty) scopes. On average, our model predicts 338 (337.8) scopes, one for each predicted cue. The average token length of gold scopes is 7.04. Our model predicts only slightly shorter scopes, with an average length of 7.01. In 54.71 % of the cases where the cue prediction is correct, the predicted scope matches the gold scope exactly.

In Table 3.13 on page 31, we look at the distribution of incorrect scopes by cue, counting any deviation from the gold standard as an incorrect scope. Except in the case of *-løst* '-less', the percentages are generally high for affixal cues. The same is the case for *ingen måte* 'no way', *ingenting* 'nothing' and *ingen* 'no'/'nobody'. *Ikke* 'not', *uten* 'without' and *aldri* 'never' get their scopes correctly resolved in the majority of cases.

So far, we have described the scope errors made by the system from a quantitative point of view. However, the qualitative aspect is important as well. In the following, we perform a manual inspection of the scope resolution errors and discuss interesting patterns.

[10]The formula of False Discovery Rate is FP/(FP+TP). We multiply it by 100 to obtain the score as a percentage, increasing the readability of the numbers.

Cue	Trans.	Total	FN (%)
ikke	not	184	2.0
u-	un-/dis-/non-	66	15.2
uten	without	26	0.0
ingen	none/nobody	12	8.3
-løse	-less	11	29.1
aldri	never	10	0.0
-løs	-less	7	42.9
ingenting	nothing	4	0.0
ingen måte	no way	3	26.7
-løst	-less	3	0.0
mangelen	the lack	2	100.0

Table 3.11: For all cues with > 1 occurrence in the gold standard, we report the number of occurrences as cues in the development test set and % of false negatives (FN). Values are averaged across 5 runs corresponding to 5 instances of the model trained with different seeds. Words have been lowercased before counting.

FP cue	Trans.	% of all FP	FDR
u-	un-/dis-/non-	35.2	13.8
ikke	not	14.1	2.0
uten	without	7.8	7.1
-løse	-less	7.8	15.4
ingen	none/nobody	7.0	13.0
-løs	-less	3.9	12.5
mangelen	the lack	3.9	33.3
ingen måte	no way	3.1	21.1
hverken eller	neither nor	3.1	100.0
nei	no (interj.)	2.3	100.0
utenom-	extra-	2.3	100.0
la være	let be	2.3	100.0
slipper	doesn't have to	2.3	100.0
in-	in-	1.6	100.0
være	be	0.8	100.0
måte	way	0.8	100.0

Table 3.12: Distribution of false positive cues, in terms of percentage of all false positives and False Discovery Rate (FDR). Numbers are averaged across 5 runs corresponding to 5 instances of the model trained with different seeds. Words have been lowercased before counting.

Cue	Trans.	% incorrect scopes
ikke	not	37.2
u-	un-/dis-/non-	67.9
uten	without	36.9
ingen	no/ nobody	50.9
-løse	-less	76.8
aldri	never	32.0
-løs	-less	75.0
ingenting	nothing	55.0
ingen måte	no way	66.7
-løst	-less	20.0

Table 3.13: Scope predictions for true positive cues with > 1 occurrence. Unless there is a perfect match with the gold scope, it is counted as an error. The raw frequency for each cue can be derived from Table 3.11 on the preceding page. Numbers are averaged across 5 runs corresponding to 5 instances of the model trained with different seeds. Words have been lowercased before counting.

Too short predicted scopes We have already stated that our model generally produces scopes that are slightly shorter than the true scopes. The examples (3.3) - (3.8) all have in common that one or more elements are missing from the predicted scopes.

In the following examples, where an affixally negated adjective is part of a noun phrase, we understand that the scope should be the whole noun phrase except for the negation affix. The model labels only the adjective itself as belonging to the scope in (3.3), which is wrong according to the gold standard. In (3.4), the part to the left of the negated adjective is left out.

(3.3) Gold: *... som [et]* ***u****[vanlig påkostet innslag i en*
Pred: *... som et* ***u****[vanlig] påkostet innslag i en*
... like an unusually lavish feature in a
kulturmønstring for ungdom] .
kulturmønstring for ungdom .
culture.gathering for youth .
'...like an unusually lavish feature in a cultural gathering for young people.'

(3.4) Gold: *I dag er eventyrene pastellfarget av [Disneys*
Pred: *I dag er eventyrene pastellfarget av Disneys*
In day are fairytales.the pastel.colored by Disney's
*harm]**løse** og romantiske [univers] .*
*[harm]**løse** og romantiske [univers] .*
harmless and romantic universe .
'Today, the fairytales are pastel colored by the harmless and romantic Disney universe.'

The case is different in (3.5); here, the negated adjective is the predicate of the sentence. The model has resolved the adjective and the subject as part of the scope, but the verb *var* 'was' is missing.

(3.5) Gold: *... var bygd over [en prosessorarkitektur] som [var]*
Pred: *... var bygd over [en prosessorarkitektur] som var*
... was built over a processor.architecture that was
*ukjent og **u**[vanlig] .*
*ukjent og **u**[vanlig].*
unknown and unusual .
'...was built with an unknown and unusual processor architecture.'

Example (3.6) shows a free-standing negation cue modifying the object of the sentence. Our model correctly includes the subject in the negation scope, but leaves out the main verb.

(3.6) Gold: *[Han holder] begravelse etter begravelse , men **ingen***
Pred: *[Han] holder begravelse etter begravelse , men **ingen***
He holds funeral after funeral , but no
[barnedåper] .
[barnedåper] .
child.baptisms .
'He holds funeral after funeral, but no baptisms.'

As for the non-affixal cues, we notice a pattern in examples (3.7) and (3.8); the parts of the sentence to the left of the cue that belong to the scope, are not detected. This hints at the model having difficulties with discontinuous scopes, especially when there is a long distance between the cue and the left part of the scope. Mæhlum et al. (2021) mention a high error rate for discontinuous scopes as a problem in their modeling approach as well.

(3.7) Gold: *[Her inviteres vi til å kikke og le] av , **ikke***
Pred: *Her inviteres vi til å kikke og le av , **ikke***
Here are.invited we to to look and laugh of , not
[med] .
[med] .
with .
'Here, we are invited to look and laugh at, not with.'

(3.8) Gold: *[« Rio » er] et feelgood-album ... og* ***uten*** *[låter*
Pred: *« Rio » er et feelgood-album ... og* ***uten*** *[låter*
« Rio » is a feelgood-album ... and without songs
som havner i « Den store sangboka »] .
som havner i « Den store sangboka »] .
that end.up in « The big songbook » .
'« Rio » is a feelgood-album...and without songs that end up in « The Great Songbook ».'

Too long predicted scopes In contrast to the scope resolution errors already discussed, there are also cases where additional elements are included in the predicted scopes. This applies to the examples (3.9) - (3.13).

Example (3.9) illustrates a case of main verb negation where the whole clause is correctly predicted as part of the scope. However, the preceding clause is falsely included as well.

(3.9) Gold: *... når Lucifer har valgt seg ut en brud*
Pred: *... [når Lucifer har valgt seg] ut [en brud*
... when Lucifer has chosen himself out a bride
[gir han] ***ikke*** *[opp så fort] .*
gir han] ***ikke*** *[opp så fort] .*
gives he not up so fast.
'...when Lucifer has chosen a bride for himself, he does not give up that easily.'

In (3.10), both the expletive subject and the adverb *nå* 'now' are labeled as scope words. The inclusion of an adverb that should be kept outside the scope is also shown in (3.11) (*ennå* 'yet'). Example (3.12) illustrates the tendency that the relative subjunction *som* 'that, which' is sometimes falsely included inside scopes.

(3.10) Gold: *Men det [er] nå* ***ikke*** *[utseendet som er*
Pred: *Men [det er nå]* ***ikke*** *[utseendet som er*
But it is now not appearance.the that is
viktig når vi snakker spillkonsoller] ...
viktig når vi snakker spillkonsoller] ...
important when we talk.about game.consoles ...
'But it is not appearance that matters when we talk about game consoles...'

(3.11) Gold: *... tro på [en tekst] som ennå* ***ikke*** *[finnes] ?*
Pred: *... tro på [en tekst] som [ennå]* ***ikke*** *[finnes] ?*
... believe in a text that yet not exists ?
'...believe in a text that does not exist yet?'

(3.12) Gold: *... [en innlevelse] som* ***ikke*** *[gjør skam på den*
Pred: *... [en innlevelse som]* ***ikke*** *[gjør skam på den*
... an empathy that not make shame on the
flotte latinamerikanske musikktradisjonen] .
flotte latinamerikanske musikktradisjonen] .
great Latin.American music.tradition.the .
'...emotion that the great Latin American music tradition would not be ashamed of.'

Last, we have a case of an implicit gold scope in (3.13). Here, our system predicts a non-empty scope containing the preceding, and only other, word.

(3.13) Gold: *Slett* ***ikke*** *!*
Pred: *[Slett]* ***ikke*** *!*
Plain not !
'Certainly not!'

3.3 Experiments with language models

On the basis of the errors discussed in 3.2.3, we find it interesting to experiment further with different language models. In particular, we are interested in improving the results for affixal negation, as well as for discontinuous scopes. It is also desirable to improve on scope resolution in sentences containing expletive subjects and the relative subjunction.

3.3.1 Language models

There are various transformer-based language models available for Norwegian, as well as multilingual models. We choose the following language models for our experiments, all of which we access through Hugging Face[11]:

NorBERT-2 This is the one that has been used in all our previous experiments. A brief description of the model is provided in 3.2.1.

NB-BERT-base NB-BERT-base[12] (Kummervold et al., 2021) is a model trained by the National Library of Norway (NLN). The model is initiated from multilingual BERT-base, with 12 transformer encoder layers (Devlin et al., 2019), and has a vocabulary size of approximately 120,000 (Kummervold et al., 2021), i.e. significantly larger than the vocabulary of NorBERT-2. According to Nielsen (2023), the number of trainable parameters is 178 million. The model was trained on a corpus of 18 billion words consisting both of texts from NLN and other resources from the web (Kummervold et al., 2022).

[11] https://huggingface.co/
[12] https://huggingface.co/NbAiLab/nb-bert-base

NB-BERT-large NB-BERT-large[13] (Kummervold et al., 2021) is another, larger language model developed at NLN. The number of encoder layers of NB-BERT-large is 24, i.e. twice the number of layers in NB-BERT-base, corresponding to a $\text{BERT}_{\text{large}}$ architecture, and the training data is the same as for NB-BERT-base (Nasjonalbiblioteket, n.d.). The number of trainable parameters is also doubled as compared to the base version (355 million), according to Nielsen (2023). In contrast to NorBERT-2 and NB-BERT-base, the vocabulary of NB-BERT-large is *uncased* (Nasjonalbiblioteket AI lab, 2021) with a size of 50,000 tokens (Nielsen, 2023).

XLM-RoBERTa-base XLM-RoBERTa-base[14] (Conneau et al., 2020) is a multilingual language model developed by Facebook AI. It is trained on 2.5 TB of text in 100 languages, i.e. far more data than the Norwegian models previously discussed, and has a vocabulary size of 250,000 tokens (Conneau et al., 2020). The vocabulary is created by Byte Pair Encoding tokenization (Sennrich et al., 2016).

XLM-RoBERTa-large XLM-RoBERTa-large[15] (Conneau et al., 2020) shares its characteristics with the corresponding base-model, except for some architectural differences such as the number of layers (24 compared to 12) and 550 million trainable parameters, compared to 270 million in the base-version (Conneau et al., 2020).

Except for the language models, we keep our model settings and hyperparameter values as before. The larger language models require us to reduce the training batch size to avoid memory errors. We choose 16 as batch size for these models. For comparison, we experiment with batch size 16 for the other language models as well, in addition to the original size of 32.

3.3.2 Results

The evaluation results of the experiments with various language models and batch sizes are presented in Tables 3.14 to 3.16 on page 37. The tables correspond to the following evaluation methods, respectively: raw predictions against the original gold standard, raw predictions against the adjusted, word-level gold standard, and predictions modified through affix matching against the original gold standard.

Overall, the system using nb-bert-large as its language model performs best, although the CUE score is somewhat higher for the nb-bert-base model trained with batch size 16. There is quite a large improvement compared to the best model from the initial experiments, which is the norbert2 model with batch size 32. We observe an increase in all three metrics. FN increases the most, from 63.62 to 68.10 (4.48), while ST

[13] https://huggingface.co/NbAiLab/nb-bert-large
[14] https://huggingface.co/xlm-roberta-base
[15] https://huggingface.co/xlm-roberta-large

increases by 1.80 from 85.00 to 86.80 and CUE by 1.27 from 92.07 to 93.34. This is based on the *Original+RE* evaluation in Table 3.16 on the facing page.

Training instability

Regarding the system using the xlm-roberta-large language model, we observe instability in training. For some seeds, the models reach a good F_1 score during the first epochs, but the F_1 score then drops to 0 in the next epoch. As we use early stopping and save the best model, this is not necessarily a problem. However, for one seed, the scope resolution model predicts empty scopes for all negation cues. This explains the poor ST and FN results for this model. We assume that fine-tuning xlm-roberta-large requires careful tuning of the hyperparameter values to achieve stable results. This is outside the scope of this thesis.

Training time

We train our models using a GPU and observe that the training time of the best model is approximately 41 minutes. This is roughly 2.5 times the training time of the models using norbert2 as their language model (approximately 17 minutes). For the second best models (nb-bert-base), training is about 20 % slower than for the norbert2 models.

Prediction time

Predictions are run on a CPU using the development test set. Concerning the prediction times, we find the same pattern as for training times. With norbert2 it takes about 3.2 minutes, compared to 4.1 minutes for the nb-bert-base models. For the best system, the one using the large version of nb-bert, the duration is 16.7 minutes, i.e. approximately 5 times longer than the norbert2 system and 4 times longer than the nb-bert-base system. Since the difference in both training and prediction time between the best model and the second best models (nb-bert-base) is quite clearly in favor of nb-bert-base, using the latter could be an option in cases where time and resources are limited.

Evaluation on the held-out test set

We evaluate our best model on the held-out test set and present the results for all three evaluation methods in Table 3.17 on page 38. Based on evaluation against the original gold standard after extraction of negation affixes, our CUE score (93.73) matches the best-performing model of Mæhlum et al. (2021) (93.7). Our ST (87.57) and FN (73.29) scores are higher than theirs, by 0.86 and 6.49, respectively.

3.3.3 Error analysis

In this part, we provide a quantitative and qualitative error analysis of our new best system. As before, we use the development test set and conduct

		Original		
Lang.model	**Batch**	CUE	ST	FN
norbert2	32	79.45 (0.31)	82.03 (0.27)	57.42 (1.27)
	16	79.36 (0.32)	81.79 (0.45)	57.34 (1.25)
nb-bert-base	32	79.51 (0.29)	83.04 (0.79)	58.94 (1.15)
	16	**80.01** (0.75)	83.21 (0.95)	**59.11** (1.15)
nb-bert-large	16	79.48 (0.45)	**83.49** (0.42)	58.82 (0.88)
xlm-roberta-base	32	79.03 (0.54)	82.44 (0.88)	57.89 (1.89)
	16	79.21 (0.30)	82.10 (0.53)	58.88 (1.17)
xlm-roberta-large	16	79.39 (0.61)	65.44 (36.65)	44.85 (24.63)

Table 3.14: Results of models with various language models and batch sizes when evaluated against the original gold standard of the development test set. The metrics from the 2012 *SEM shared task are used. We report the average across 5 runs.

		Adjusted		
Lang.model	**Batch**	CUE	ST	FN
norbert2	32	92.07 (0.10)	85.00 (0.23)	63.62 (2.14)
	16	92.40 (0.29)	84.89 (0.45)	64.93 (0.95)
nb-bert-base	32	93.22 (0.17)	86.29 (0.78)	67.20 (2.25)
	16	**93.70** (0.38)	86.47 (0.85)	66.75 (1.89)
nb-bert-large	16	93.34 (0.40)	**86.80** (0.29)	**68.10** (0.55)
xlm-roberta-base	32	92.26 (0.61)	85.53 (0.86)	66.43 (2.05)
	16	92.37 (0.26)	85.19 (0.53)	66.87 (1.59)
xlm-roberta-large	16	92.56 (0.68)	67.90 (38.02)	51.14 (28.30)

Table 3.15: Results of models with various language models and batch sizes when evaluated against the adjusted gold standard of the development test set. The metrics from the 2012 *SEM shared task are used. We report the average across 5 runs.

		Original+RE		
Lang.model	**Batch**	CUE	ST	FN
norbert2	32	92.07 (0.10)	85.00 (0.23)	63.62 (2.14)
	16	92.40 (0.29)	84.89 (0.45)	64.93 (0.95)
nb-bert-base	32	93.22 (0.17)	86.29 (0.78)	67.20 (2.25)
	16	**93.70** (0.38)	86.47 (0.85)	66.75 (1.89)
nb-bert-large	16	93.34 (0.40)	**86.80** (0.29)	**68.10** (0.55)
xlm-roberta-base	32	92.26 (0.61)	85.53 (0.86)	66.43 (2.05)
	16	92.37 (0.26)	85.19 (0.53)	66.87 (1.59)
xlm-roberta-large	16	92.56 (0.68)	67.90 (38.02)	51.14 (28.30)

Table 3.16: Results of models with various language models and batch sizes when performing affix extraction on the predicted cues and evaluating against the original gold standard of the development test set. The metrics from the 2012 *SEM shared task are used. We report the average across 5 runs.

	CUE	ST	FN
Original	81.69 (0.39)	84.46 (0.56)	64.47 (1.37)
Adjusted	93.73 (0.16)	87.57 (0.49)	73.29 (1.56)
Original+RE	93.73 (0.16)	87.57 (0.49)	73.29 (1.56)

Table 3.17: Results of evaluating the best system on the held-out test set according to the three different evaluation methods. The metrics from the 2012 *SEM Shared Task are used. We report the average across 5 runs.

the analysis according to the same template as in 3.2.3. We will see how the errors of this system compare to the previous best system.

Cue errors

Averaged over 5 seeds, our system correctly predicts 326 (326.4) out of 342 gold cues, leaving only 16 (15.6, 4.6 %) false negatives. The number of false negatives is hence almost halved compared to the previous best system. We find 32 (32.2) false positive cues, somewhat more than we saw in the old system. The total number of predicted cues is 359 (358.6).

False negatives Table 3.18 on the facing page compares the current and previous best system with respect to the percentage of false negatives per cue occurring at least twice in the development test set. For the affixal cue *-løst* '-less', the number has increased from 0.0 to 20.0 %, but for all other cues where false negatives were found in the old model, the proportion is decreased in the new model.

In the analysis of our first system, we listed some words with the negation prefix *u-* 'un-/dis-/non-' that the model did not identify as cues. We observe that our new system manages to correctly label the more infrequent of those words, e.g. *uspennende* 'unexciting' and *umælende* 'speechless'. This is an interesting finding; it indicates that the detection of infrequent cues can benefit from the use of a language model trained on a larger amount of data. It is likely that the model gets exposed to a larger number of rare words when the size of the training set is increased. In general, the number of false negative *u-* 'un-/dis-/non-' cues is reduced by more than 50 %.

False positives An overview of all words falsely predicted as cues by the new best system is provided in Table 3.19 on page 40, in terms of word type, proportion of all false positives and False Discovery Rate. We note that the false positives are spread out across a large number of word types in this system as compared to the previous one. Among the most interesting cases, we find the English words 'hopeless' and 'nothing' and the nominalized adjective *fantasiløshet* 'unimaginativeness'. The ones that we find most

Cue	Trans.	Total	FN (old)	FN (new)
ikke	not	184	2.0	0.7
u-	un-/dis-/non-	66	15.2	7.0
uten	without	26	0.0	0.0
ingen	none/nobody	12	8.3	5.0
-løse	-less	11	29.1	12.7
aldri	never	10	0.0	0.0
-løs	-less	7	42.9	14.3
ingenting	nothing	4	0.0	0.0
ingen måte	no way	3	26.7	20.0
-løst	-less	3	0.0	20.0
mangelen	the lack	2	100.0	40.0

Table 3.18: For all cues with > 1 occurrence in the gold standard, we report the number of occurrences as cues in the development test set and % of false negatives (FN) in the new best system. We include the results from Table 3.11 for comparison. Values are averaged across 5 runs corresponding to 5 instances of the model trained with different seeds. Words have been lowercased before counting.

difficult to explain are *stumme* 'mute', *tvilsomme* 'questionable', *plutselige* 'sudden' and *slukøret* 'crestfallen'.

Scope errors

The average length of predicted scopes is 6.82. In other words, our new system systematically predicts somewhat shorter scopes compared to the gold standard (7.04). Our previous best system produced scopes closer to this length (7.01), but had poorer ST and FN scores.

Furthermore, we see that given a correct cue prediction, 59.01 % of scopes exactly match the gold standard. This is an increase of 4.3 percentage points compared to the old system.

We include Table 3.20 on page 42 to visualize the distribution of incorrect scopes across the various cues and make a comparison to the previous best system in this regard. Only a perfect match between predicted and gold scope is counted as correct. We see that the percentage of incorrect scope predictions is reduced for the three most common cues *ikke* 'not', *u-* 'un-/dis-/non-' and *uten* 'without', as well as for a selection of the less frequent cues. Three cues show an increased proportion of incorrect scopes. In particular, we note that the new system generally outperforms the old one with respect to scope resolution for affixal cues.

Expletive subjects In the predictions of the old system, we observed some cases of the wrongful inclusion of expletive subjects inside scopes. Our new system correctly leaves it out in some cases where the old system

FP cue	Trans.	% of all FP	FDR
u-	un-/dis-/non-	26.7	12.3
ikke	not	12.4	2.1
-løse	-less	8.7	22.6
uten	without	6.2	7.1
in-	in-	5.0	100.0
ingen	none/nobody	5.0	12.3
(h)verken eller	neither nor	4.3	100.0
nei	no (interj.)	3.1	100.0
-løs	-less	3.1	14.3
mangelen	the lack	3.1	33.3
utenom-	extra-	3.1	100.0
hopeless	-	2.5	100.0
la være	let be	2.5	100.0
fantasiløshet	unimaginativeness	1.9	100.0
ingen måte	no way	1.9	20.0
slipper	doesn't have to	1.2	50.0
minus	minus	1.2	100.0
løst	-less	1.2	14.3
miste	lose	1.2	100.0
slukøret	crestfallen	0.6	100.0
stumme	mute	0.6	100.0
nothing	-	0.6	100.0
i	in	0.6	100.0
måte	way	0.6	100.0
plutselige	sudden	0.6	100.0
ute	out	0.6	100.0
tvilsomme	questionable	0.6	100.0
være	be	0.6	100.0

Table 3.19: Distribution of false positive cues in the new best system, in terms of percentage of all false positives and False Discovery Rate (FDR). Numbers are averaged across 5 runs corresponding to 5 instances of the model trained with different seeds. Words have been lowercased before counting.

failed, such as the one in example (3.10). In other cases, like example (3.14), the new system makes mistakes.

(3.14) Gold: ... *det* *[var]* *rett* *og* *slett* ***ikke*** *[godt* *nok]* .
Pred: ... *[det* *var]* *rett* *og* *slett* ***ikke*** *[godt* *nok]* .
... it was straight and plain not good enough .
'...it was simply not good enough.'

The relative subjunction As seen previously in example (3.12), the old model was inclined to include the relative subjunction *som* 'that, which' as part of scope when it would have been correct not to. We see several cases of this with the new model as well. The annotations are not entirely consistent in this regard, hence we believe it difficult for a model to learn these patterns perfectly.

Sentential adverbs Our improved system still predicts sentential adverbs as part of scopes in some cases. One example is the previously seen (3.11) (*ennå* 'yet'). Another is *definitivt* 'definitely' in example (3.15). We note that there are examples where the annotations violate the principle to exclude these from the scopes.

(3.15) Gold: *For* *[av* *kunstnerisk* *art* *er* *motivasjonen* *for* *"*
Pred: *For* *[av* *kunstnerisk* *art* *er* *motivasjonen* *for* *"*
For of artistic nature is motivation.the for "
Zoolander *2* *"]* *definitivt* ***ikke*** .
Zoolander *2* *"* *definitivt]* ***ikke*** .
Zoolander 2 " definitely not .
'...Because the motivation for "Zoolander 2" is definitely not of artistic nature.'

Scopes of affixal cues Here, we observe that the model still struggles with examples (3.3), making the same mistake as the old model.

Discontinuous scopes Our new system makes the same scope resolution error as the old one in example (3.8). In the case of example (3.6), however, it includes the subject and hence predicts the scope correctly.

3.3.4 Conclusion

We conclude that our new system outperforms the old one with respect to recall on cues. This comes at the expense of a somewhat higher frequency of false positive cues. The system generally makes the same types of scope resolution mistakes as the old one. Nevertheless, it makes fewer errors.

A goal of the language model experiments was to improve on affixal negation. We have seen that the number of false negative affixal cues is reduced. In addition, the relative frequency of incorrect scope predictions has decreased, especially for the negation prefix *u-* 'un-/dis-/non-'.

Cue	Trans.	% incorrect (old)	% incorrect (new)
ikke	not	37.2	35.3
u-	un-/dis-/non-	67.9	50.2
uten	without	36.9	33.8
ingen	no/ nobody	50.9	52.4
-løse	-less	76.8	82.7
aldri	never	32.0	38.0
-løs	-less	75.0	71.2
ingenting	nothing	55.0	45.0
ingen måte	no way	66.7	33.3
-løst	-less	20.0	10.0

Table 3.20: Scope predictions in the new best system for true positive cues with > 1 occurrence. We include the results from Table 3.13 for comparison. Unless there is a perfect match with the gold scope, it is counted as an error. The raw frequency for each cue can be derived from Table 3.18. Numbers are averaged across 5 runs corresponding to 5 instances of the model trained with different seeds. Words have been lowercased before counting.

Based on our manual review of the scope resolution errors, we are unable to draw any conclusions as to whether the system has improved on the other problem areas specifically, e.g. discontinuous scopes. This would require a more systematic analysis.

Chapter 4

Reviewing the NoReC$_{neg}$ annotation

Annotation is fundamental in training supervised machine learning models. As part of the annotation process, a set of clear annotation guidelines should be formulated. When there is more than one annotator working on a project, this is crucial to ensure that all of them have a common understanding of the annotation task. Providing the guidelines as a public document is important in order to make the project transparent to researchers and other interested parties. This will allow for the application of the guidelines to new data in other annotation projects. In addition, it opens the door for others to critically review the original guidelines, which is what we aim to do with the annotation instructions for the NoReC$_{neg}$ dataset (Mæhlum et al., 2021) in this chapter.

We will look at the guidelines and identify possible areas of doubt and deficiency. Furthermore, we will inspect the actual annotations made by the annotators and evaluate their adherence to the guidelines and annotation trends in the cases not sufficiently covered by the guidelines. We note that what we refer to when these are mentioned throughout the chapter, is the document containing the complete set of guidelines, which is available on GitHub[1]. The goal of this work is to prepare guidelines for annotation of new data in the medical domain, which will be the topic of the subsequent chapter of this thesis.

4.1 Review of the annotation guidelines

In this section, our focus is on the parts of the guidelines that we find to be unclear or deficient. We examine the general trends in the annotation of the various cases and discuss these findings.

[1]The latest version per May 13, 2023 (from Jun 1, 2021) has been used in this thesis: https://github.com/ltgoslo/norec_neg/blob/main/annotation_guidelines/guidelines_neg.md

4.1.1 Affixal negation

In this part, we have a closer look at affixal negation. This includes identification of underspecified guidelines as well as difficult cases in which we are uncertain if negation should be marked.

Lexicalized or not?

The $\text{NoReC}_{\text{neg}}$ annotators were instructed to only annotate affixal cues in words that also exist in the lexicon without the negative affix, and furthermore, the negative affix is required to actually *negate* this word in order to be annotated (Mæhlum et al., 2021).

Typical examples of correctly annotated cases are seen in (4.1) and (4.2). *Ulykkelig* 'unhappy' clearly is the negation of *lykkelig* 'happy'. *Poseløs* 'bagless' obviously indicates that the nominal part of this word, *pose* 'bag', is absent. Note that because these are copulative sentences with an affixally negated adjective as the predicate, the scope is the whole clause except additional adjectives and adverbs (*ensom* 'lonely' in (4.1)).

(4.1) *[Jane er] ensom og* ***u****[lykkelig] ...*
Jane is lonely and unhappy ...
'Jane is lonely and unhappy...'

(4.2) *[Robotstøvsugeren er pose]****løs*** *...*
Robot.dust.sucker.the is bagless ...
'The robot vacuum cleaner is bagless...'

Some cases are trickier than those above. Example (4.3) contains *fine* (singular: *fin*), a frequently used adjective with a positive meaning ranging from 'nice, pretty' to 'noble', 'posh' and 'precise' (*fin* n.d.). Whether *ufine* 'foul' is the exact negation of any of these meanings is not obvious to us, but it is annotated as such. However, we think that the sentence becomes less natural if *ufine* is replaced by *fine* and thus would have preferred not to annotate negation in this case. Since (4.3) is the only occurrence of *ufin(e)* 'foul' in the dataset, neither supporting examples nor counterexamples can be found.

In example (4.4), negation is annotated for the word *uvøren* 'reckless'. To our knowledge, this word is highly lexicalized, although perhaps not completely, i.e. *vøren* is very rarely used without the prefix *u-* 'un-/dis-/non-'. Based on this argument, one could have chosen not to annotate it. We observe, however, that the total of two occurrences are marked as negation.

(4.3) *... forsøkte å øke andelen elektrisk kjøring med litt*
... tried to increase share.the electric driving with little
u*[fine metoder] .*
foul methods .
'... tried increasing the proportion of electric driving by somewhat foul means .'

(4.4) *... Jan Vardøens debutfilm føles* ***u****[vøren] på flere måter*
... Jan Vardøen's debut.film feels reckless on several ways
'...Jan Vardøen's first movie feels reckless in several ways'

We also look at one specific word, the adjective *ulike* 'dissimilar, various, several', which occurs 50 times in the dataset. We group the use of this word in two, where the first one is illustrated by examples (4.5) and (4.6). Here, it is used to emphasize that a set of objects are different from each other.

(4.5) *Dette møtet mellom to meget* ***u****[like personligheter]*
This meeting.the between to very dis.similar personalities
'This meeting between to very different personalities'

(4.6) *[Venninnene er]* ***u****[like] ...*
Friend.FEMININE.PLURALthe are dis.similar ...
'The friends are different...'

The other main category is where *ulike* 'dissimilar, various, several' is used to express that there are *several* objects, not putting weight on the difference between them. We assess the two sentences (4.7) and (4.8) to be typical examples of this group. Note that the former is annotated as if this were negation, while the latter has no annotation.

(4.7) *... han møter* ***u****[like ansatte i*
... he meets dis.similar employees in
trygdeetaten som fremstår som roboter ...]
social.security.administration.the who appear as robots ...
'... he meets several employees of the Social Security Administration who come across as robots...'

(4.8) *Vi møter Amor sittende i et løsrevet bilsete , med*
We meet Amor sitting in a apart.torn car.seat , with
ulike bildeler strødd utover gulvet ...
various car.parts strewn througout floor.the ...
'We meet Amor sitting in a detached car seat, with various car parts scattered all over the floor...'

By inspecting the dataset, we see that there are several cases that are annotated as negation although they seem to belong to this second group. However, there are cases such as (4.9) where it is not quite clear whether they belong to the first or second group. We think that the best way to handle this problem is to only annotate negation in *ulike* when it is certain that its meaning emphasizes difference between objects, as in examples (4.5) and (4.6).

(4.9) ***U****[like former for ulykkelighet]*
Dis.similar forms of unhappiness
'Different/various forms of unhappiness'

Affixally negated adjectives used as adverbs

Scope resolution for most cases of affixal negation is covered by the $\text{NoReC}_{\text{neg}}$ guidelines, but the case where an affixally negated adjective is used as an adverb is not explicitly mentioned. Interestingly, this is handled in different ways in previous work, either annotating the whole sentence as the scope (Morante et al., 2011), or only the negated adverb (Liu et al., 2018). Note that these two works only mention cases where the adverb modifies a *verb*. As we will see shortly, an affixally negated adverb can also modify an adjective or another adverb.

By inspecting the $\text{NoReC}_{\text{neg}}$ dataset (Mæhlum et al., 2021), we observe that the annotations are highly consistent in following Liu et al. (2018), who argue that their approach is more correct. We support their assessment that only the adverb describing the action or event taking place should be considered negated. This is illustrated by example (4.10), one of many sentences correctly annotated according to this interpretation. The sentence in (4.11) is one of the rare exceptions to this trend, including the whole clause in the scope. We observe the same trend for affixally negated adverbs when they modify a whole clause rather than only a verb. An adverb commonly used in this way is *utvilsomt* 'undoubtedly'.

(4.10) *... blikket feide* ***u****[sikkert] over det svartmalte lokalet*
... gaze.the swept insecurely over the black.painted local
nederst i Maridalsveien .
lowermost in Maridal.'s.road .
'...his gaze swept insecurely over the black-painted room at the bottom of Maridalsveien.'

(4.11) *[« Forgetting Sarah Marshall » minner]* ***u****[sjenert om*
« Forgetting Sarah Marshall » reminds unabashedly about
komedier som ...]
comedies like ...
'« Forgetting Sarah Marshall » unabashedly reminds us of comedies such as...'

To illustrate affixally negated adverbs modifying adverbs and adjectives, we include examples (4.12) and (4.13). Generally, the pattern from above can be observed in these cases as well, i.e. only the negated adverb is inside the scope. Yet we find one occurrence of an adverb with a negation affix modifying an adjective and scoping over the noun phrase it is part of, as seen in (4.14).

(4.12) *Boka er* ***u****[vanlig] fint illustrert .*
Book.the is unusually nicely illustrated .
'The book is unusually nicely illustrated .'

(4.13) *Seks hjerter viser derimot til en* ***u****[sedvanlig]*
Six hearts show on.the.other.hand to an unusually
positiv opplevelse over all forventning .
positive experience over all expectation .
'Six hearts, however, indicates an unusually positive experience exceeding every expectation.'

(4.14) *... spesialiteten er tross alt [skam]****løst*** *[emosjonelle låter]*
... specialty.the is despite all shamelessly emotional songs
med en underliggende desperasjon ...
with an underlying desperation ...
'...after all, their specialty is shamelessly emotional songs with an underlying desperation...'

"The whole NP" as scope

For affixal negation, we understand from the guidelines that the scope will always be the noun phrase in which the affixal negation is contained, except when an affixally negated adjective is the predicate of a copula sentence. At first, this might seem all clear. When we look at the actual sentences of the dataset, however, we realize that it is not always obvious what "the whole NP" refers to. In addition, there are cases that make us wonder if it is really desirable to include the full noun phrase in the scope. In the following, we will discuss various patterns where these questions arise and provide examples obtained from the annotated dataset.

Determiners We look at how the annotators treat various types of *determiners* in noun phrases with affixal negation. Determiners include possessives, demonstratives and quantifiers (Hagemann, 2020). We notice an example in the guidelines indicating that the indefinite article should be included when the noun phrase is in the scope, but other types of determiners are not addressed at all.

Determiners often occur directly to the left of an affixally negated adjective. We observe that the indefinite and definite articles are considered as part of scope in most cases, as in examples (4.15) and (4.16). There are also some cases where they are kept outside scope, see examples (4.17) and (4.18). As far as we can see, there is no reason to treat these pairs of cases differently.

(4.15) *En dag finner han [en hjelpe]****løs*** *[fugleunge] ...*
One day finds he a helpless bird.child ...
'One day he finds a helpless baby bird...'

(4.16) *... unge mennesker i [den håp]****løse*** *[overgangsfasen*
... young humans in the hopeless transition.phase.the
mellom barn og voksen] .
between child and adult .
'...young people in the hopeless transition from child to adult.'

(4.17) *... en* ***u****[klar historie] og mange små feil gjør at jeg*
... an unclear history and many small errors do that I
ikke kan anbefale å bruke penger på dette spillet
not can recommend to use money on this game.the
'...due to an unclear story-line and lots of minor errors, I cannot recommend buying this game'

(4.18) *... den [barn]****løse*** *[dronningen] , spilt av Salma Hayek ...*
... the childless queen.the , played by Salma Hayek ...
'...the childless queen, played by Salma Hayek...'

The number of intervening tokens between the determiner and the rest of the scope is variable. In many cases of a large distance, the determiner is annotated as inside the scope, as in (4.19), while in cases such as (4.20), it is not.

(4.19) *... [en] svakt formulert og [temperaments]****løs***
... a weakly formulated and temperless
[film-andakt uten verken mening eller forstand] .
film-devotion without either meaning or sense .
'...a poorly formulated and temperless devotional movie, neither meaningful nor sensible.'

(4.20) *... de velkjente og tradisjonelle , for ikke å si*
... the well-known and traditional , for not to say
tradisjonsrike , velprøvde og nærmest ***u****[slitelige*
tradition.rich , well.tried and almost inexhaustible
julesangene] .
Christmas.songs.the .
'...the well-known and traditional, not to say rich in tradition, tried-and-true and nearly inexhaustible Christmas carols .'

For quantifiers other than the articles, most observed cases are cardinal numbers such as *tre* 'three' and *seks* 'six'. As far as we can see, all the cases with cardinal numbers annotate the determiner as belonging to the negation scope. An example can be seen in (4.21). In addition, we find one occurrence of the quantifier *noen* 'some', where it is included in the scope, and two occurrences of *alle* 'all'. One of these is annotated as inside scope and is shown in (4.22).

(4.21) *[De 14] " ****u****[kjente landssvikerne] " som profiterte ...*
The 14 " unknown country.traitors " who profited ...
'The 14 "unknown traitors of the nation" who profited...'

(4.22) *De fire spillbare figurene har [alle]* ***u****[like egenskaper]*
The four playable figures have all dis.similar properties
'The four playable characters have all different properties'

For possessives (formerly known as possessive pronouns), the annotation pattern resembles what we observe for the articles. They are annotated as in example (4.23) in most cases. We observe a few instances such as example (4.24), where they are left outside the scope.

(4.23) *... [hans]* ***u****[proffe opptreden] får negative*
... his unprofessional behavior gets negative
konsekvenser
consequences
'...his unprofessional behavior has negative consequences'

(4.24) *Smartingene ... som viste sitt [verdi]****løse*** *[ansikt] da*
Smart.people.the ... who showed their worthless face as
det omsider smalt ...
it finally banged ...
'The smart-asses who showed their worthless faces as it finally exploded...'

We identify one sentence with the determiner *andre* 'other', which we would classify as a demonstrative. As we can see in (4.25), it is considered as inside the scope.

(4.25) *På ett punkt skiller imidlertid Urbanite XL seg ut*
On one point divides however Urbanite XL itself out
*fra [andre tråd]****løse*** *[modeller] :*
from other wireless models :
'At one point, however, Urbanite XL differs from other wireless models:'

In summary, there is a general tendency to include the determiner inside the scope no matter the type of determiner (possessive, demonstrative, quantifier). To us, it is not quite clear why this is done. For instance, in the case of the quantifier *alle* 'all' in (4.22), the scope of the phrase *alle ulike egenskaper* 'all different properties' becomes *alle like egenskaper* 'all similar properties'. It seems more intuitive to set the scope to *like egenskaper* 'similar properties', since the quantifier is not really affected by the negation.

Genitive phrases We find several cases of genitive noun phrases where the head of the phrase is a noun phrase modified by an adjective containing a negation affix. These phrases seem similar to the previous examples with determiners, but since determiners are not complete phrases (Hagemann, 2020), they are syntactically different. We provide two examples: In (4.27), the genitive noun phrase is not included in the negation scope, in contrast to (4.26).

(4.26) *[Pi Patels]* ***u****[trettelige oppfinnsomhet] er imponerende .*
Pi Patel's tireless ingenuity is impressive .
'Pi Patel's tireless ingenuity is impressive.'

(4.27) *Flere av verdens største stjerner setter melodi til*
Several of world.the.'s biggest stars set melody to
*Hank Williams' **u**[kjente tekster] .*
Hank Williams' unknown texts .
'Several of the world's greatest stars add melody to Hank Williams' unknown texts.'

Among the examples we have seen, both annotation patterns are well represented. The genitive phrase is included in scope slightly more often than not.

Postmodifiers of noun phrases Prepositional phrases and relative clauses can act as modifiers of a noun or noun phrase. In both cases, they appear to the right of the nominal head and can be referred to as *postmodifiers* (Nordquist, 2020). The question of whether to include these when the head of the phrase is inside the scope, is not addressed by the guidelines. We inspect the annotations with respect to the inclusion or exclusion of these constructs inside scopes. This is done by looking mostly at examples with the cue *u-* 'un-/dis-/non-' since this is by far the most common affixal cue.

First, we study cases with relative clauses as postmodifiers. In general, there is some variation as to whether these clauses are included in the scope of the affixal cue. We believe this could be related to the distinction between restrictive and non-restrictive relative clauses. Ideally, a non-restrictive relative clause would be preceded by a comma in Norwegian, but this is probably not always done in practice.

For example, in (4.28), the relative clause 'that is the trademark of the band' is closely related to the head of the noun phrase, i.e. we are not talking about *any* 'irresistible urge', but the one that the band is famous for. The same intuition applies to example (4.29); this is the specific arrogance represented by a certain person. This is made more obvious by the absence of the relative subjunction *som* 'that, which'.

(4.28) *« (It's Not War) Just The End Of Love » får [det]*
« (It's Not War) Just The End Of Love » gets the
***u**[imotståelige suget som er bandets varemerke] ...*
irresistible suck that is band.the.'s trademark ...
'«(It's Not War) Just The End Of Love» gets the irresistible urge that is the trademark of the band...'

(4.29) *... [det tanke]**løse** [standshovmodet hun representerer]*
... the thoughtless social.position.arrogance she represents
'...her thoughtless arrogance towards people of lower social class'

Two sentences where the relative clause is excluded from the scope are shown in (4.30) and (4.31). In comparison to the examples (4.28) and (4.29), the relative clauses feel more loosely connected to the head of the noun phrase, i.e. the information they carry seem to be supplementary rather than crucial to the meaning of the noun phrase as a whole.

(4.30) ... *[en]* ***u****[tiltalende]* , *måpe-kyndig [buskkvast-nerd] som*
... an unappealing , gape-capable bush.broom-nerd who
snakker ironisk til oss ...
speaks ironic to us ...
'...an unappealing "bush broom nerd" capable of gaping who speaks to us ironically...'

(4.31) ... *[et]* ***u****[oversiktlig kaos] der alle potensielt er*
... an un.clear chaos there all potentially are
hverandres fiender , men også allierte .
each.other.'s enemies , but also allied .
'...an unorganized chaos where everyone potentially is an enemy, but also an ally.'

There are, however, also examples that do not comply with this principle of restrictive vs. non-restrictive. Notwithstanding that we consider the relative clause as a piece of additional information in (4.32), it is included in the scope of the negation.

(4.32) ... *[en]* ***u****[elegant] og forstoppa [komedie som etter*
... an unelegant and constipated comedy which after
hvert trenger farse-klyster for å komme til enden] .
every needs farce-enema for to come to end.the .
'...an unelegant and constipated comedy that eventually needs farcical enema to get to the end.'

Next, we look at prepositional phrases. These phrases are quite frequently included inside the scopes, but there are also several cases where they are left out. Examples (4.34), (4.35), (4.37) and (4.39) belong to the first group, while (4.33), (4.36) and (4.38) belong to the second. Parallel to what we saw in the examples with relative clauses, there seem to be differences between the individual cases regarding the importance of the prepositional phrase in giving meaning to the NP as a whole. In (4.35), *ubetinget suksess* 'unconditional success' makes more sense on its own than *uvanlig blanding* 'unusual mix' in (4.33), where an essential bit of information is added by the subsequent prepositional phrase. Despite this, the prepositional phrase is not part of the scope in (4.33), whereas it is included in (4.35).

Note that some of the examples (4.33) through (4.39) perhaps should not have been considered negation at all according to our previous discussion on lexicalized adjectives with the *u-* 'un-/dis-/non-' prefix, and this might also be the case in other examples. In this part, however, we are only focused on how the scopes are resolved.

(4.33) ... *en svært* ***u****[vanlig blanding] av jazz og hinduistisk*
... a very unusual mix of jazz and hinduistic
inspirert åndelig musikk .
inspired spiritual music .
'...a very unusual mix of jazz and Hindu inspired spiritual music.'

(4.34) *[De] stort sett* ***u****[morsomme longørene mellom låtene]*
The largely seen unfunny longueurs between songs.the
'The mostly unfunny longueurs between the songs'

(4.35) *Boksen har vært [en]* ***u****[betinget suksess for Get]*
Box.the has been an unconditional success for Get
'The box has been an unconditional success for Get'

(4.36) *... en halvkokt og* ***u****[morsom komedie] om selvhøytidelige*
... a half.boiled and unfunny comedy about self-important
regissører ...
directors ...
'... a half-cooked and unfunny comedy about self-important directors...'

(4.37) *...* ***u****[like former for dronefilming] ...*
... dissimilar forms for drone.filming ...
'...different types of drone filming...'

(4.38) *... [et]* ***u****[overveid skifte] i manusforfattere etter den*
... an unconsidered change in manuscript.writers after the
andre filmen ...
second movie.the ...
'...an ill-considered change of screenwriters after the second movie...'

(4.39) *... [en]* ***u****[nødvendig uthaling av tiden mellom hvert*
... an unnecessary out.dragging of time.the between every
oppdrag] .
mission .
'...an unnecessary way to drag out time between missions.'

Furthermore, we notice a few examples like (4.40). Here, the noun phrase *penger og oppmerksomhet* 'money and attention' serves to specify the noun *mengder* 'amounts' and is annotated as part of the scope. A preposition such as *med* 'with' or *av* 'of' could have been used, but is not necessary.

(4.40) *Beyoncé Knowles vil raske med seg* ***u****[horvelige*
Beyoncé Knowles will grab with herself un.appropriate
mengder penger og oppmerksomhet] ...
amounts money and attention ...
'Beyoncé Knowles will get hold of enormous amounts of money and attention...'

Finally, we include (4.41) to show that a noun can also be modified by an infinitive phrase. We think that the infinitive phrase directly modifies the noun *måte* 'way' and thus it is reasonable to include it in the scope as the annotator has done in this case.

(4.41) *[En]* ***u****[formell]* *og* *morsom* *[måte* *å* *spise* *på]* ...
An informal and fun way to eat on ...
'An informal and fun way to eat...'

Additional adjectives and adverbs

For cases of affixal negation, we understand from the guidelines that the scope can be either the full clause or limited to a noun phrase. We have already expressed some confusion and pointed to annotation inconsistencies in the cases where the scope is supposed to be the full noun phrase. The inclusion or exclusion of additional adjectives and adverbs not directly involved in negation is relevant to such noun phrase scopes, but also in affixal negation scoping over a full clause.

First, we will look at some examples of affixal negation with noun phrase scopes and explain why we think the guidelines should have been more specific regarding what it means to include the whole noun phrase in the scope. We refer to the examples below for illustration purposes. In (4.42), if we accept that *subjektiv skyld for sin krigsinnsats* 'subjective guilt for their war effort' is negated, this would equate to saying that 'non-subjective guilt' is semantically equivalent to 'subjective non-guilt' or 'subjective innocence', which is definitely not the case. The latter interpretation is the intended one for this example. Examples (4.43) and (4.44) show that adverbs of degree like *totalt* 'totally' and *ganske* 'quite' strongly interfere with negation; '**not** totally known' is something else than 'totally **un**known', and '**not** quite complicated' can be interpreted as different from 'quite **un**complicated'.

(4.42) ... *i* *egen* *fornemmelse* *av* *[subjektiv]* ***u****[skyld* *for* *sin*
... in own sensation of subjective un.guilt for their
krigsinnsats] ...
war.effort ...
'...in their own sense of subjective innocence for their war efforts...'

(4.43) *For* *ikke* *å* *snakke* *om* *det* *å* *komme* *over* *[et* *totalt]*
For not to talk about that to come over a totally
u*[kjent* *kjempetalent]* ...
unknown giant.talent ...
'Not to talk about discovering a totally unknown, huge talent...'

(4.44) *Innerst* *inne* *er* *Warriors* *egentlig* *[et]* *ganske*
Innermost inside is Warriors actually a quite
u*[komplisert* *slåssespill]* .
uncomplicated fighting.game .
'At the core, Warriors is actually a fairly uncomplicated fighting game.'

Actually, this problem is touched upon in the $\text{NoReC}_{\text{neg}}$ paper, and to our understanding, the rule on leaving out additional adverbs and

adjectives from affixal scopes (Mæhlum et al., 2021) applies both to nominal and clausal scopes. We still think this should have been explained in greater detail providing more examples, and it should certainly be added to the guidelines.

There are also cases with inclusion of adverbs or adjectives in affixally negated *sentences* in the dataset. Example (4.45) illustrates this. When including *ekstremt* 'extremely' in the scope , it definitely changes the meaning of the negated proposition. '**Not** extremely pleasant' and '**not** pleasant' certainly have different meanings, where the last and most negative one is the intended meaning in this case.

(4.45) *[Filmatisk er scenene ekstremt]* ***u****[behagelige å være*
Filmatically are scenes.the extremely unpleasant to be
vitne til] ...
witness to ...
'Cinematically, the scenes are extremely unpleasant to watch...'

In the following, we include examples that are annotated correctly in accordance with the previous discussion. The adverbs *mer* 'more' in (4.46) and *så* 'so' in (4.47) are excluded from the negation scopes, and so is the adjective *vanlige* 'usual' in (4.46).

(4.46) *Mer* ***u****[frelste] , vanlige [kinogjengere] vil ikke sterkt føle*
More unsaved , regular movie.goers will not strongly feel
savnet .
longing.the .
'The more unsaved, regular moviegoers will not miss it much.'

(4.47) *[De unge soldatene er] så* ***u****[erfarne] at ...*
The young soldiers.the are so un.experienced that ...
'The young soldiers are so inexperienced that...'

Our general impression from all the examples we have found is that additional adjectives are often left out of scope. When it comes to additional adverbs, there is more variation. This applies both to noun phrase scopes and clausal scopes.

Copula verbs

Mentioned as copular verbs in the guidelines are only *å være* 'to be' and *å bli* 'to become'. However, we observe some cases of affixal negation with other verbs that adhere to the annotation guidelines for copula sentences. *Kopula* 'copula' is defined by Store norske leksikon *(eng: The Great Norwegian Encyclopedia)* as verbs that connect the subject to the predicate (Hagemann, 2023). Using this definition, the term can be extended to include verbs such as *virke* 'seem' and *fremstå* 'appear'. In examples (4.48) and (4.50), the annotators seem to have used such an extended definition, yet this is not consistent throughout the dataset, as exemplified by (4.49) and (4.51).

(4.48) ... *[bandet virker]* ***u****[fokusert]* *og likegyldig* ...
... band.the seems unfocused and indifferent ...
'...the band seems out of focus and indifferent...'

(4.49) *Det virker* ***u****[oppfinnsomt]* .
It seems uninventive .
'It seems uninventive.'

(4.50) ... *[Jolies regi fremstår]* ... ***in****[effektiv]* ...
... Jolie's direction appears ... inefficient ...
'...Jolie's direction appears inefficient...'

(4.51) ... *fremstår imidlertid utvalget heller* ***u****[inspirert]* .
... appears however selection.the rather uninspired .
'... the selection however appears rather uninspired.'

4.1.2 *uten* 'without' as a cue

Uten 'without' is a free-standing, syntactic negation cue. It belongs to the prepositions and is thus used differently than other common syntactic cues such as the sentential adverbs *ikke* 'not' and *aldri* 'never', and *ingen* 'no, nobody', which can function both as a determiner and as a pronoun.

The cue *uten* 'without' is not specifically discussed in the guidelines, and we believe this to have caused some confusion for the annotators, as we observe several annotation inconsistencies in these cases. Although not addressed in their own guidelines, we note that 'without' is briefly described in the annotation guidelines (Morante et al., 2011) of the English ConanDoyle-neg corpus (Morante and Daelemans, 2012), upon which the $\text{NoReC}_{\text{neg}}$ guidelines are partly based (Mæhlum et al., 2021). There, it is suggested that the scope of 'without' be the phrase that it introduces, i.e. the noun phrase or clause to the right of the cue. We will keep this in mind as we study the different usages of *uten* 'without' and report for each case how it is treated by the annotators.

uten + NP as adverbial phrase

A prepositional phrase with *uten* 'without' as its head and a noun phrase (NP) as the prepositional object can function as an adverbial in a sentence. This is the case in the examples (4.52) and (4.53). They also illustrate two different ways of resolving the scope of *uten* 'without'. As we can see from the first sentence, the annotated scope spans across the whole clause, whereas in the second one, the scope is limited to the prepositional object. In the majority of cases that we have identified in the dataset, scope resolution adheres to the pattern seen in (4.53), which is also in accordance with Morante et al. (2011). However, as (4.52) shows, there are some exceptions.

(4.52) *... [Turboneger gikk av scenen]* ***uten*** *[tegn på ereksjonssvikt]* .
... Turboneger went off stage.the without signs of erection.failure .
'...Turboneger went off stage with no signs of erectile dysfunction.'

(4.53) ***Uten*** *[store geberder] gjennomlever han en livsomveltende krise* .
Without big gestures through.lives he a life-changing crisis .
'He lives through a life-changing crisis without any big fuss.'

In (4.52), the proposition has been interpreted as a negation of 'Turboneger went off stage *with* signs of erectile dysfunction.' The annotation of example (4.53), on the other hand, seems to be based on a different logic; the prepositional object refers to what is *actually* being described as absent, and the rest of the sentence, corresponding to the proposition 'He lives through a life-changing crisis', is not made untrue in any sense by the presence of the negation cue.

According to the guidelines, the whole sentence should be inside the scope if the main verb is modified by the negation. It seems likely that this rule has been interpreted as applicable in (4.52), since one could say that the negation *modifies* the main verb. However, given the example provided with this guideline, we suspect that the rule is meant to apply only to negation cues that actually *invert* the truth value of the action or event referred to by the verb. If that is the case, the guidelines should have included a separate rule defining how to treat the cases with *uten* 'without'.

uten + NP as the predicate of a copula sentence

The dataset contains a handful of sentences with copula verbs (*å være* 'to be', *å bli* 'to become') where the predicate is a prepositional phrase with *uten* 'without' as the head and a noun phrase as the complement. Two such examples are shown in (4.54) and (4.55). There is a trend that these are annotated in the same way as example (4.54), i.e. the subject, copula verb and prepositional object of *uten* 'without' are included in the scope. Example (4.55) illustrates a less frequent case, where the annotator has included the subject, but left out the verb. Interestingly, both these practices are contrary to the suggestion of Morante et al. (2011) regarding the cue 'without', although they do not target copula sentences specifically.

(4.54) *Den største skuffelsen er likevel at [musikken er]* ***uten*** *[vinger] ...*
The biggest.the disappointment.the is nevertheless that music.the is without wings ...
'Nevertheless, the biggest disappointment is that the music does not have wings ...'

(4.55) *Som historielekse er [den] derimot ikke helt* ***uten***
As history.homework is it however not wholly without
[små svakheter] .
small weaknesses .
'As a history assignment, however, it has a few flaws .'

Regarding predicative complements in copula sentences, the $\text{NoReC}_{\text{neg}}$ (Mæhlum et al., 2021) guidelines put forward that if such elements are negated by a negative item, the whole phrase (sentence) will be inside the scope. They make an exception to this rule in the case where the predicate is an affixally negated noun phrase. The predicative complement in these cases is the prepositional phrase consisting of *uten* 'without' and the noun phrase. To be accurate, it is not the predicative complement that is negated, but the prepositional object inside the predicative complement.

However, there is a similarity between these cases and copula sentences with an affixally negated adjective as the predicate, and in those cases, the guidelines imply that the whole clause is the scope. We visualize this similarity by the made-up examples (4.56) and (4.57). Morante et al. (2011) emphasize that cases with the same meaning should be analyzed in the same manner. Given this argument, the scope resolution in (4.54) and (4.57) seems reasonable.

(4.56) *[Han er vilje]****løs***
He is will.less
'He has no will'

(4.57) *[Han er]* ***uten*** *[vilje]*
He is without will
'He has no will'

We make another observation in copula sentences where a 'without + NP' predicate is combined with another negation cue. In most of these cases, the annotators have either overlooked the negation with *uten* 'without' entirely, as in (4.58), or included only the complement of the preposition as the scope, as in (4.59).

(4.58) *Og kritikken er ikke uten en viss berettigelse ...*
And critic.the is not without a certain justification ...
'And the criticism is not without some justification...'

(4.59) *Forholdet mellom de to hovedrollene er ikke* ***uten***
Relation.the between the two main.roles is not without
[spenning og skepsis] ...
excitement and scepticism ...
'The relation between the two lead roles is not free of excitement and scepticism...'

***uten* + NP as postmodifier of a noun phrase**

Prepositional phrases headed by *uten* 'without' quite frequently occur as postmodifiers of noun phrases in the dataset, providing additional information about the noun phrase head. In these cases, the prepositional complement is also a noun phrase. Three annotated examples from the dataset, (4.60), (4.61) and (4.62), are provided for illustration.

(4.60) *Musikk* ***uten*** *[vinger]*
Music without wings
'Music without wings'

(4.61) *. . . ansatte i trygdeetaten som*
. . . employees in social.security.administration.the who
fremstår som [roboter] ***uten*** *[medmenneskelighet] .*
appear as robots without compassion .
'. . . employees of the Social Security Administration who come across as robots without compassion.'

(4.62) *... utryddelse ... eksekvert av [ideologiske fanatikere]* ***uten***
... extinction ... executed by ideological fanatics without
[samvittighet] , ...
conscience , ...
'... extermination ... executed by ideological fanatics with no conscience, ...'

What these examples also visualize, are the two main scope resolution patterns observed for these constructs in the dataset. In example (4.60), only the prepositional object of *uten* 'without' is contained in the scope. In contrast, the part of the noun phrase to the left of the cue is included as well in examples (4.61) and (4.62). In the latter, this involves an additional adjective to the left of the noun phrase head.

Our inspection of the dataset shows that the cases including the whole noun phrase are far more numerous than the cases limiting the scope to the prepositional object of *uten* 'without'. We note that this annotation practice is contrary to the principle applied by Morante et al. (2011). In the $\text{NoReC}_{\text{neg}}$ (Mæhlum et al., 2021) guidelines, we find nothing neither to support nor to contradict this annotation scheme. What we do notice, however, is that this practice is quite similar to the guideline concerning affixal negation in noun phrases, stating that the scope should be the noun phrase. By paraphrasing *ideologiske fanatikere uten samvittighet* 'ideological fanatics without conscience' as *samvittighetsløse ideologiske fanatikere* 'conscienceless ideological fanatics', the similarity becomes apparent. Another paraphrase is *ideologiske fanatikere har ingen samvittighet* 'ideological fanatics have no conscience'. This case is different due to the presence of a verb (*har* 'has'), but one might use the argument that one would include the subject *ideologiske fanatikere* 'ideological fanatics' in the scope in this case, and thus it should be included in the scope as the head of the noun phrase in the original sentence as well.

uten + infinitive phrase / subordinate clause

Above, we provided examples of prepositional phrases with *uten* 'without' as adverbials in a sentence. Here, we will look at more such cases, but this time with different prepositional complements. In the $\text{NoReC}_{\text{neg}}$ dataset (Mæhlum et al., 2021), there are 46 occurrences of *uten* 'without' followed by an infinitive phrase, and 43 cases of the same cue followed by a subordinate clause introduced by the subjunction *at* 'that', i.e. this is quite common. In English, both these constructs could be replaced by 'without' combined with a present participle phrase.

In the clear majority of cases, the annotated scope consists of the whole subordinate clause or the whole infinitive phrase, including the subjunction or infinitive marker, respectively. This is parallel to the tendency seen for cases with *uten* 'without' and noun phrase complements and thus also complies with Morante et al. (2011). Examples (4.63) and (4.64) illustrate these typical cases. We also observe a few sentences where the whole main clause is included in the scope, as well as a minor number of cases where the subjunction *at* 'that' is excluded.

(4.63) *Vil du betjene BeoPlay A3* ***uten*** *[at den velter] ...*
Will you operate BeoPlay A3 without that it tips ...
'If you want to use BeoPlay A3 without it tipping over...'

(4.64) *... kan stå på egne bein* ***uten*** *[å være avhengig av*
... can stand on own legs without to be dependant of
de kjente karakterene] .
the known characters.the .
'...can stand on its own without relying on the famous characters.'

4.1.3 Lexical negation

Certain verbs and nouns in Norwegian have negation as part of their meaning, e.g. *forsvinne* 'disappear', *la være* 'refrain from' and *mangel* 'lack' (Mæhlum et al., 2021). This form of negation is referred to as lexical negation by Jiménez-Zafra et al. (2020). In the following, we provide an overview of lexical negation cues observed in the dataset and discuss the scope resolution patterns we notice for these cues.

Lexical negation cues

In Table 4.1 on the next page, we list words and phrases we have identified as *possible* lexical negation cues in the dataset. This list includes all lexical negation cues that are annotated in the dataset. These were identified by extracting the set of all annotated cues and ruling out those fitting the definition of syntactic negation (Jiménez-Zafra et al., 2020) and morphological negation, also known as affixal negation (Jiménez-Zafra et al., 2020). We consider exception items such as *(med) unntak (av)* '(with) exception (of), except' and *bortsett (fra)* 'except (for)' as syntactic negation.

Word/phrase	Translation	PoS	Occurs as cue?
blotte (for)	(make) devoid (of)	Verb	Yes
mangel	lack	Noun	Yes
mangle	lack	Verb	Yes
fjerne	remove	Verb	No
hindre	prevent	Verb	No
forhindre	prevent	Verb	No
forsvinne	disappear	Verb	Yes
fravær	absence	Noun	Yes
fraværende	absent	Verb (Adj)	Yes
la være	refrain from	Verb	Yes
ribbe (for)	strip (of)	Verb	Yes
strippe (for)	strip (of)	Verb	Yes
savn	lack	Noun	Yes
savne	miss	Verb	Yes
slippe	not have to	Verb	Yes
bort	away	Adverb	Yes
borte	gone	Adverb	Yes
miste	lose	Verb	Yes
nekte	refuse, deny	Verb	Yes
ha til gode	still not have	Verb	Yes
utebli	not appear	Verb	Yes
avstå (fra)	abstain (from)	Verb	Yes
unngå	avoid	Verb	Yes

Table 4.1: Possible lexical negation triggers present in the $\text{NoReC}_{\text{neg}}$ (Mæhlum et al., 2021) dataset, including their English translation and part of speech. Nouns are given in their indefinite, singular form. Verbs are given in the infinitive, or in the occurring form if an infinitive does not exist. The rightmost column indicates whether or not the expression occurs as an *annotated* cue in the dataset. We do not include exception items as part of this list.

By manual inspection of the dataset, we observe three possible lexical negation cues that are never marked as cues, either because they were missed or not considered as negation by the annotators. These are *fjerne* 'remove', *hindre* 'prevent' and *forhindre* 'prevent'. We include these in Table 4.1 on the facing page as well.

The reader should note that some of the expressions in Table 4.1 on the preceding page have multiple meanings, where only one implies negation or something close to negation. This applies to the verb *slippe*, which can be used in the meaning of 'not have to', '(to) drop', '(to) let go of' and '(to) release (an album or similar)'. It is also the case with the adverb *borte*, which can mean 'gone', but can also be used in the expression *der borte* 'over there'. Another example is *fjerne*, which could mean either '(to) remove' or 'far away', however, these are different parts of speech, i.e. verb and adjective, respectively.

The guidelines do not provide a complete list of cues for lexical negation. Thus, discretion must be used to decide what should be considered lexical negation and not. Also, cues such as *(å) mangle* '(to) lack' can be ambiguous as to whether or not a negated reading is implied (Mæhlum et al., 2021), so each possible cue occurrence requires individual assessment.

As previously mentioned, three words from Table 4.1 on the facing page never occur as cues. There is a difference in the perspective or the degree of activity as opposed to passivity, but we still think that *fjerne* 'remove' is quite similar to *forsvinne* 'disappear' and that *(for)hindre* 'prevent' resembles *unngå* 'avoid'.

Example (4.65) illustrates the use of *fjerne* 'remove'. According to our understanding, it would have been a possibility to consider *mystikken* 'the mystery' as negated in this sentence.

(4.65) *Team Ninja har fjernet mystikken fra og respekten*
Team Ninja has removed mystery.the from and respect.the
for Samus ...
for Samus ...
'Team Ninja has removed the mystery from and the respect for Samus...'

The verbs *hindre* 'prevent' and *forhindre* 'prevent' can be used in different contexts with slightly different meanings. In some cases, their meaning is more in the direction of 'make difficult, cause something to be slowed down, delayed' or 'be in somebody's way'. Other uses of these verbs are more similar to negation; according to an acknowledged Norwegian dictionary, *forhindre* can mean 'make something not happen' (*forhindre* n.d.), and *hindre* can be used similarly (*hindre* n.d.). A constructed example is *En katastrofe ble forhindret* 'A disaster was prevented'. Here, it is quite natural for us to think of *en katastrofe* 'a disaster' as negated.

Worth noting is also that there might be an aspect of modality to *(for)hindre* 'prevent'. Modality is a complex topic that relates to subjectivity, reliability, perspective and the degree of speculation expressed

in a statement, and it is known to interact with negation (Morante and Sporleder, 2012), which can make it difficult to distinguish negation from modality. The verbs *(for)hindre* 'prevent' express an aspect of not wanting something to happen. From our previous example 'A disaster was prevented', we understand that the potential disaster was an unwanted event. Volition, or the degree to which an event or state is wanted or unwanted, is a common parameter in several annotation efforts targeting modality (Morante and Sporleder, 2012). Based on this and the decision in $\text{NoReC}_{\text{neg}}$ to only annotate negation when it can be separated from modality (Mæhlum et al., 2021), we have a possible explanation why *(for)hindre* 'prevent' does not occur as an annotated cue. We include three examples ((4.66) - (4.68)) of how *(for)hindre* 'prevent' is used in the dataset. Inspection of the annotations also show that *slippe* 'not have to' and *unngå* 'avoid' are quite frequently *not* annotated as cues, which might be because there is a similar modal aspect to these verbs as well.

(4.66) *De forsøker å gi historien nødvendig vekt for å*
They try to give history.the necessary weight for to
forhindre at den stiger til værs som en varmluftsballong
prevent that it rises to weather.'s as a warm.air.balloon
'They attempt to give the necessary weight to the story to prevent it from rising into the air like a hot air balloon...'

(4.67) *... skjermen ... , som har fått en matt hine*
... screen.the ... , which has gotten a matte membrane
utenpå , antagelig for å hindre gjenskinn .
outside , presumably for to prevent reflection .
'...the screen...,which is covered by a matte membrane, presumably to prevent reflections.'

(4.68) *Det er det eneste som hindrer meg fra å gi denne*
It is the only that hinders me from to give this
filmen en sekser på terningen .
movie.the a six on die.the .
'It is the only thing that prevents me from giving this movie a top rating.'

A quick search through the dataset tells us that for many of the lexical negation cues from Table 4.1 on page 60, the annotations are not consistent as to whether negation is marked or not. We keep in mind the high ambiguity rate for such cues (Mæhlum et al., 2021), but we still have the impression that negation triggered by these cues is sometimes overlooked. We provide a selection of examples numbered (4.69) - (4.74), marking possible missed cues in bold. Although we believe these cases to represent missing annotations, it could be that at least some of them have been assessed by the annotators and evaluated as not qualifying as negation. For instance, we imagine there could have been a discussion as to whether *mangelen* 'the lack' in example (4.69) represents an actual absence of discipline or just *limited* discipline.

(4.69) *Denne* ***mangelen*** *på disiplin gjør dem til "boyband" -*
This lack.the on discipline does them to "boyband" -
bransjens mest sympatiske .
industry.the.'s most sympathetic .
'This lack of discipline makes them the boyband industry's most likeable.'

(4.70) *Og spiser du ikke suppa di , svinner du hen*
And eat you not soup.the yours , fade you away
og blir ***borte*** *.*
and become gone .
'And if you do not eat your soup, you will fade away and be gone.'

(4.71) *Synes du det høres naivt ut ,* ***har*** *du* ***til gode*** *å*
Think you it sounds naïve out , have you to good to
overvære et dj-sett der alt klaffer :
over.be a DJ.set there all flaps :
'If you think it sounds naïve, you have never heard a perfect DJ set:'

(4.72) *I* ***fravær*** *av berøringsskjerm , navigerer man seg rundt*
In absence of touch.screen , navigates one onself around
med knappene på urkassen .
with buttons.the on watch.case.the .
'In absence of a touch screen, one navigates with the buttons on the watch case.'

(4.73) *Men skriveren er helt* ***strippet*** *for funksjoner ...*
But writer.the is wholly stripped for functions ...
'But the printer is totally devoid of functionality...'

(4.74) *... designerne ser ut som om de hadde* ***mistet*** *noe av*
... designers.the see out as if they had lost some of
inspirasjonen ...
inspiration.the ...
'...the designers look as if they had lost some of their inspiration...'

Scope resolution for lexical negation

We understand from the guidelines that particles or prepositions associated with lexical negation cues should be regarded as part of neither the cue nor the corresponding scope. Other than this, the resolution of scopes in lexical negation is not discussed directly. In theory, the guidelines regarding subjects and objects modified by negation seem to apply in cases of verbal lexical negation cues. However, it is not clear to us if these guidelines are designed to be applied to lexical negation since (1) none of the provided examples include this type of cues, and (2) it seems less intuitive to include the whole clause or sentence in the scope when the main verb is the cue and thus is outside the scope. From Table 4.1 on page 60, we know that there is

a large number of possible lexical negation cues. We believe that the correct way to resolve the scope depends on the cue. In many cases of verbal lexical negation cues, a central question seems to be if the whole sentence should be in the scope or not. When examining the relevant examples in the dataset, we observe a tendency to include the whole clause inside the scope in the cases where a lexical negation verb is the main verb of the sentence. A selection of such sentences are shown in examples (4.75) - (4.80).

(4.75) *Og [det]* ***unngår*** *[heller ikke "Thanks for Sharing"] .*
And that avoids also not "Thanks for Sharing" .
'And "Thanks for Sharing" does not avoid that either.'

(4.76) *... [han] rammes av produsentens imperialistgen og* ***mister*** *[kontrollen] .*
... he is.affected of producer.the.'s imperialist.gene and loses control.the .
'... he is affected by the producer's imperialist gene and loses control.'

(4.77) *[Filmen er]* ***blottet*** *for [humor] .*
Movie.the is stripped of humor .
'The movie is devoid of humor.'

(4.78) *[Jeg]* ***har*** *fortsatt* ***til gode*** *[å se at en spiller som har scoret , ... , får dårligere enn 7] .*
I have still to good to see that a player who has scored , ... , gets worse than 7 .
'I still have not seen a player who scored a goal get rated lower than 7.'

(4.79) *... [den]* ***uteble*** *[i stor grad i denne delen av historien] .*
... it was.absent in large degree in this part.the of history.the .
'...it was largely absent in this part of the story.'

(4.80) *-Hva med blekkspruten , undret [Fredag] , som [hadde]* ***avstått*** *fra [å smake på den] .*
-What with squid.the , wondered Friday , which had abstained from to taste on it .
'-How about the squid , wondered Fredag , who had abstained from tasting it.'

As we can see from Table 4.1 on page 60, not all lexical negation cues are verbal; there are also nouns and adverbs. Additionally, some cues are used in the present participle form of the verb, like *manglende* 'lacking' and *fraværende* 'absent'.

In examples (4.81), (4.82) and (4.83), we illustrate the use of three related cues. *Mangler* 'lacks' is used as the main verb in (4.81). The corresponding

noun *mangelen* 'the lack' is found in (4.82), and the present participle form *manglende* 'lacking' is a negation cue in (4.83). We note that in the example with the main verb cue, (4.81), the whole clause is annotated as the scope. In the examples with the nominal and present participle cues, the scope is limited to the noun phrase following the cue. Furthermore, we include example (4.84) to show that present participle cues can be used as predicates. Based on this example and the previous one, present participle cues seem to be treated similarly to affixal cues; if it is the predicate, the whole sentence is the scope (example (4.84)), and if not, the scope is only the NP directly modified by the adjective (example (4.83)).

(4.81) *[Munken bistro]* ***mangler*** *[litt på detaljene]* .
Monk.the bistro lacks a.little on details.the .
'Munken bistro lacks a little in the details.'

(4.82) *... i skarp kontrast til* ***mangelen*** *på [en interessant historie]*
... in sharp contrast to lack.the on an interesting history
'... in stark contrast to the lack of an interesting story'

(4.83) *Det ... illustrerer allikevel* ***manglende*** *[støtte fra tredjeparter]* .
It ... illustrates anyway lacking support from third.parties .
'Nevertheless, it illustrates a lack of support from third parties.'

(4.84) *... men [magien er] fullstendig* ***fraværende*** .
... but magic.the is complete absent .
'... but the magic is completely absent.'

4.2 Clear deviations from the guidelines

In the previous part, section 4.1, we looked at cases not sufficiently covered by the guidelines, and we saw that the handling of these was generally not consistent. Here, we will discuss problem areas where we observe obvious annotation errors although the guidelines leave little doubt as to how these cases should be annotated.

4.2.1 Elements to be excluded from scopes

One main group of errors that we observe has in common that certain elements are falsely included in the negation scopes. In the following, we will describe and provide examples of these mistakes.

Prepositions and particles in fixed expressions

Mæhlum et al. (2021) state that prepositions and particles in expressions such as *mangel på* 'lack of' and *fravær av* 'absence of' should not be considered part of negation cues, and according to their examples, they

should not be included in the corresponding scopes either. There is only a handful of such cues in the dataset. In a majority of the annotated sentences with *mangel(en) på* '(the) lack of', among them example (4.85), *på* 'of' is actually considered as a part of the scope by the annotators, while this is never the case with *av* 'of' in *fravær av* 'absence of'.

(4.85) *... med en total* ***mangel*** *[på karisma] .*
... with a total lack of charisma .
'...with a total lack of charisma.'

Relative subjunction

The guidelines explicitly mention that when the (main) verb of a subordinate clause is negated, the scope should span the whole subordinate clause except for the subordinating conjunction. They illustrate this using an example sentence with the relative subjunction *som* 'that, which'. Initially, we observe some cases where *som* 'that, which' is falsely included inside scopes. To get an approximation of the prevalence of this error, we search for *som ikke* ' that/which/who not ' in the annotated dataset and get 74 matches. Out of these, we count 20 sentences where the scopes seem to have been correctly resolved except for the inclusion of *som* 'that, which'. Among the other 54 matches, the majority are exactly correct scopes, i.e. *som* 'that, which' is excluded. The rest consists of a few other annotation errors and un-annotated cases. Examples (4.86) and (4.87) represent the group where scopes are correctly resolved except for the inclusion of the relative subjunction.

(4.86) *Kanskje er jeg blendet av [overskrifter som]* ***ikke*** *[har med*
Perhaps am I blinded of headlines that not have with
film å gjøre] ...
film to do ...
'Perhaps I am blinded by headlines unrelated to movies...'

(4.87) *... , [ting som]* ***ikke*** *[rakk å bli med i 2012] ?*
... , things that not reached to become with in 2012 ?
'..., things that did not make it into 2012?'

Expletive subjects

According to the guidelines, annotators should leave expletive subjects outside scopes, that is when *det* 'it/there' is used not as a pronoun actually referring to an entity, but in a setting such as 'It is not raining today'. In some cases, it can be unclear whether *det* 'it/there' is expletive or not, especially when looking at a single sentence without its surrounding context. When performing a case-insensitive search with the phrase *'det er ikke'*, 'it/there is not' through the whole dataset, we get 47 matches, where probably not all are cases of expletive subjects. In ten cases where we are sure that the subject is expletive, it is mistakenly included in the scope of

ikke 'not', as illustrated by example (4.88). In other words, this type of mistake is quite frequent. However, our impression from the inspection of the data is that this case is treated correctly more often than not. Still, the relatively large amount of wrongly annotated cases is a likely explanation for our models' varying behavior concerning these cases, as discussed in chapter 3.

(4.88) *... [det er]* ***ikke*** *[sikkert du skjønner hele*
... it is not secure you understand whole.the
historien med en gang] .
history.the with one time .
' ... it is not certain that you will understand the whole story at once.'

***verken eller* 'neither nor' inside scope**

The guidelines clearly state that when *verken eller* 'neither nor' is triggered by another negation cue, they should be excluded from the scope of this cue. (4.89) is an example of such a case. The reader should note, however, that both *verken* 'neither' and *eller* 'nor' are inside the annotated scope. We note that this seems to be the general trend in the annotation of these constructs, i.e. this guideline is violated. We identify a couple of cases such as the one in example (4.90), where the part containing *verken eller* 'neither nor' is separated from the rest by a comma. In both sentences, the entire part following the comma is left out from the scope. According to our understanding of the guidelines, the presence of the comma should not make this case any different from example (4.89). Thus *i konkret* 'in concrete' and *overført betydning* 'transferred meaning' should have been part of the scope in (4.90).

(4.89) *... [en dimensjon kunstneren]* ***ikke*** *[er verken bevisst*
... a dimension artist.the not is neither conscious
eller klar over] ...
or clear over ...
' ... a dimension that the artist is neither conscious nor aware of...'

(4.90) *... det [er] fortsatt* ***ikke*** *[mulig å bli blendet av*
... it is still not possible to be blinded by
skjermen] , verken i konkret eller overført betydning .
screen.the , neither in concrete or transferred meaning .
' ... it is still not possible to be dazzled by the screen, neither concretely nor figuratively.'

4.2.2 Affixal negation

In this part, we highlight annotation mistakes related to affixal negation. We provide an example of what should not be regarded as affixal negation and discuss scope resolution errors.

Annotating negation in lexicalized words

In 4.1.1, we discussed a few examples with the prefix *u-* 'un-/dis-/non-' that we were not certain whether to annotate or not. There are also cases where we are confident that negation should not have been marked. The adjective *ubetalelig* in example (4.91) can be read literally as 'unpayable' or 'not possible to buy with money' but is only used in the meanings 'very funny' and 'priceless'. Removing the affix *u-* 'un-/dis-/non-' does not result in the negation of any of these meanings, but in a word that does not exist in the lexicon. Therefore, we find it surprising that all three occurrences of *ubetalelig* 'very funny, priceless' in the dataset are annotated as negation.

(4.91) *... [de beste øyeblikkene her er]* ***u****[betalelige]* .
... the best moments here are un.payable .
'...the best moments are priceless .'

Scope resolution errors

Our general impression from what we have seen of the dataset, is that scope resolution of affixal cues is mostly done according to the guidelines, but there are various types of exceptions.

One group of observed errors is found in copulative sentences where the predicate is an affixally negated adjective. Examples are (4.92) and (4.93), where the subject and copula verb, *blir* 'becomes' and *er* 'is', respectively, are missing from the scope.

(4.92) *Her blir plottet både* ***u****[troverdig] og vinglete ...*
Here becomes plot.the both un.credible and wobbly ...
'Here, the plot becomes both hard to believe in and unsteady...'

(4.93) *Filmens farsskikkelse er også tragisk , [hjelpe]****løs*** *og trengende ...*
Movie.the.'s father.figure is also tragic , helpless and needy ...
'The father figure in the movie is also tragic, helpless and in need ...'

Another type of error occurs in copula sentences where the predicate is a noun phrase containing an adjective negated by an affix. Here, the scope is supposed to span the noun phrase only, but occasionally, the whole clause is included in the scope. One of these cases is shown in (4.94).

(4.94) *[Hovland er] likevel [en]* ***u****[beregnelig type] ...*
Hovland is however an unpredictable type ...
'However, Hovland is an unpredictable guy...'

4.2.3 Negation raising

We look at the practice concerning 'negation raising', the phenomenon characterized by a negating item being moved to a higher level of the syntactic tree. For Norwegian, this is equivalent to a leftward movement within the sentence. In the cases we use to illustrate this, see (4.95) and (4.96), 'not' negates the subordinate clause 'it is funny'. 'It' is omitted from the scope due to the guideline on expletive subjects mentioned earlier in this chapter. In (4.95), the negation cue is found inside the subordinate clause it negates, but in (4.96), it has been moved ("raised") to the main clause. Still, it is the main verb of the subordinate clause that is negated, not the main verb of the main clause.

(4.95) I think (that) it [is] **not** [funny] .

(4.96) I do **not** think (that) it [is funny] .

Regarding negation raising, the NoReC$_{\text{neg}}$ (Mæhlum et al., 2021) guidelines adhere to the revision of the ConanDoyle-neg guidelines (Morante et al., 2011) made by Liu et al. (2018), stating that cases like (4.95) and (4.96) are to be annotated in the same manner, i.e. limiting the scope to the subordinate clause.

We get a rough overview of the prevalence of negation raising in the dataset by searching for the different conjugated forms of the verbs *synes* 'think', *tenke* 'think', *tro* 'believe' and *mene* 'think, mean', which are known to trigger negation raising. Among the results, we find seven typical examples of this phenomenon, and out of these, only one is correctly annotated. The remaining six sentences are annotated as though the verb of the main clause was negated. Two examples of incorrectly resolved scopes are included in (4.97) and (4.98).

(4.97) *[Harry Hole tror]* *imidlertid* ***ikke*** *[at saken kan være så enkel] ...*
Harry Hole believes however not that case.the can be so simple ...
'Harry Hole does not think the case is that simple...'

(4.98) *[Kaninen syns]* ***ikke*** *det [hjelper med en gulrot] ...*
Rabbit.the thinks not it helps with a carrot ...
'The rabbit does not think that a carrot helps...'

Additionally, we discover another group of sentences, such as example (4.99), which we believe to also be cases of negation raising. Here, there is no subordinate clause, but an infinitive phrase to which the negation belongs.

(4.99) *Men [forfatter Linde Hagerup har]* ***ikke*** *[tenkt å gjøre det enkelt] .*
But writer Linde Hagerup has not thought to do it simple .
'But the writer Linde Hagerup is not going to make it easy.'

Furthermore, we identify a few other verbs we believe are subject to negation raising as well. One is *virke* 'seem', as seen in (4.100), and another is *se ut til* 'seem' in (4.101).

(4.100) *Uheldigvis for Samsung [virker] det* ***ikke*** *[som om*
Unfortunately for Samsung seems.it that not as if
Bada har fått noe fotfeste blant utviklere ennå] .
Bada has gotten any foothold among developers yet .
'Unfortunately for Samsung, Bada does not seem to have gained a foothold among developers yet.'

(4.101) *... hvordan [så mange på tysk område etter krigen]* ***ikke***
... how so many on German area after war.the not
[så ut til å kjenne ansvar og skyld] ...
saw out to to feel responsibility and guilt ...
'...how so many in the German area after the war seemed to not feel any responsibility or guilt...'

As a final addition to this topic, we would like to mention that the $\text{NoReC}_{\text{neg}}$ guidelines specifically state that the same scope resolution rule applies where elements are elided from the subordinate clause. We have observed only two such cases in the dataset, however, in both of them, this seems to be violated through the inclusion of the main clause subject and verb in the scope. We refer to examples (4.102) and (4.103) for a visualization of this.

(4.102) *Nei , [jeg synes]* ***ikke*** *[det] .*
No , I think not it .
'No, I do not think so.'

(4.103) *Og da [mener jeg]* ***ikke*** *[flashy studioer med blinkende*
And then mean I not flashy studios with flashing
gulv og artister på storskjerm] .
floors and artists on big.screen .
'By that I do not mean flashy studios with flashing floors and artists on the big screen.'

From all this, we understand that scope resolution in sentences with negation raising must have been a challenging task for the annotators. According to our intuition, this phenomenon can be difficult to recognize, perhaps especially in cases where the subordinating conjunction is omitted, as in example (4.98), and not least when the subordinate clause is incomplete, as in (4.102) and (4.103). Also worth noting are the sentences where negation is "raised" from an infinitive phrase to the main clause. This specific case is not mentioned by Liu et al. (2018). However, as it appears to be very similar to the clear negation raising cases, we believe it should be treated similarly.

4.2.4 Missing annotations

We perform a quick search through the dataset for the common cues *ikke* 'not', *aldri* 'never' and *ingen* 'no, nobody'. Although in most cases, these cues are recognized as triggers of negation by the annotators, we discover several cases where they have been left unannotated. We provide some examples in (4.104) - (4.108), in addition to one example with the more infrequent cue *ei* 'not' in (4.109). Unannotated cues are in bold.

(4.104) *Man kan* ***ikke*** *unngå å glise seg tvers gjennom*
One can not avoid to grin oneself across through
surrehuerockeren "Air Bud" ...
buzz.head.rocker.the "Air Bud" ...
'It is impossible not to grin through the scatterbrain rocker "Air Bud"...'

(4.105) *Bekjennelseslitteratur av dette kaliber er ikke* ***u****kjent .*
Confession.literature of this caliber is not unknown .
'Confessional literature of this caliber is not unfamiliar.'

(4.106) *... og hun har* ***aldri*** *gjort en musikal siden .*
... and she has never done a musical since .
'...and she has not done musicals since.'

(4.107) *... det finnes* ***ingen*** *kunstneriske grunner til at denne*
... there exists no artistic reasons to that this
lite spektakulære filmen skal vises i 3D ...
little spectacular movie shall be.shown in 3D ...
'...there are no artistic reasons to show this unspectacular movie in 3D...'

(4.108) *United 93 var det fjerde kaprede flyet som* ***aldri*** *traff*
United 93 was the fourth hijacked plane that never hit
sitt mål 11. september 2001 .
its target 11th September 2001 .
'United 93 was the fourth of the hijacked planes that never hit its target on 9/11.'

(4.109) *... om man vil eller* ***ei*** *...*
... if one will or not ...
'...whether you want to or not...'

Performing a full review of the dataset with respect to all missed negations, is outside the scope of this thesis. Hence, we will not be able to quantify the extent of this problem. Nevertheless, we consider it important to shed light on this, since inconsistencies in the annotations will affect models trained and evaluated on this dataset.

4.2.5 Cue is annotated, but scope is missing

Through our work with $\text{NoReC}_{\text{neg}}$ (Mæhlum et al., 2021), we accidentally discover some sentences where cues are annotated, but scope is set to empty. We examine all cases of empty scopes in the entire dataset, and like the authors, we find 37 occurrences. By inspection of these, we count 13 actual *implicit scopes* (Mæhlum et al., 2021), one mistake that seems to be due to annotating *hverken eller* 'neither nor' as two separate cues, and 23 cases of non-implicit scopes that are not annotated. We believe this can either be caused by the annotators actually failing to mark the scope, or, perhaps more likely, failing to draw a 'Negates' relation from the cue to the corresponding scope, hence making it appear as if there were no scope. We do not know how the raw annotated files were processed by Mæhlum et al. (2021), so this is simply guesswork on our part.

Luckily, the extent of the problem is limited to less than 1 % of the total number of cues in the dataset (23 out of 2,672 (Mæhlum et al., 2021) cues). Still, it is worth noting since it creates noise and lowers the quality of the dataset. Our examples (4.110) - (4.112) show that this mistake has been made with a variety of cues.

(4.110) *"Rødsonen" er* ***u****tvilsomt et av høydepunktene i*
"Red.zone.the" is undoubtedly one of highlights.the in
serien .
series.the .
' "The red zone" is undoubtedly one of the series' highlights.'

(4.111) *Det er dog særlig to ting som Toyota* ***aldri***
There are though especially two things that Toyota never
ser ut til å lære .
see out to to learn .
'However, there are especially two things Toyota never seem to learn.'

(4.112) *Skjønt , oppfinnsomhet er kanskje* ***ikke*** *det*
Although , ingenuity is perhaps not the
viktigste kvalitetskriteriet her .
important.most quality.criterion.the here .
'Although ingenuity may not be the most important quality criterion here.'

Chapter 5

Annotating negation in a biomedical dataset

In this chapter, we will work with a Norwegian corpus of biomedical journal articles. Our aim is to annotate it with respect to negation in order to use it for testing our models from chapter 3 and assessing their portability into the medical domain. This evaluation will be performed in the next chapter.

First, a description of the dataset is provided. We emphasize the differences between its original negation annotations and the annotation scheme used for the $\text{NoReC}_{\text{neg}}$ (Mæhlum et al., 2021) dataset, which was used as training data for our models. Next, we discuss the procedure of annotating the biomedical corpus according to the $\text{NoReC}_{\text{neg}}$ guidelines, including the challenges met and choices made in this process. We would like to remind the reader that the complete guidelines are accessible on GitHub[1] and that any mention of them refers to this document. Finally, we analyze the dataset we have annotated with respect to negation and discuss similarities and differences in comparison to $\text{NoReC}_{\text{neg}}$.

5.1 The Norwegian GastroSurgery Biomedical Negation Corpus (NGSBNC)

The dataset we use was originally collected by Budrionis et al. (2018), containing 170 articles from Tidsskriftet for Den Norske Legeforening[2] *(eng: The Journal of the Norwegian Medical Association)*. Every publication found in the dataset belongs to the field of gastrointestinal surgery. Since the texts are part of a scientific journal, we must assume that they have been carefully edited prior to publication. The dataset was annotated for negation as part of a master's thesis from the Department of Computer and System Sciences at the University of Stockholm and designed to be used with a Norwegian version of the system known as NegEx (Chapman et al.,

[1] The latest version per May 13, 2023 (from Jun 1, 2021) has been used in this thesis: https://github.com/ltgoslo/norec_neg/blob/main/annotation_guidelines/guidelines_neg.md

[2] Webpage: https://tidsskriftet.no/

2001), which is designed to recognize negations in discharge summaries. The following description is based on this thesis (Sadhukhan, 2021):

Sadhukhan (2021) have performed preprocessing steps on the original dataset, such as removal of headings, literature references and sentences consisting of less than three tokens, as well as duplicate lines and unnecessary whitespace characters. This reduced the size of the dataset from 5,477 sentences and 88,819 tokens to 3,304 sentences and 55,683 tokens.

2,330 sentences were selected for double annotation by two native Norwegian language users, both experienced with clinical text. The annotations of one annotator, a physician, were used as the gold standard and have been made available as the *Norwegian GastroSurgery Biomedical Negation Corpus*[3], which we will refer to using the abbreviation NGSBNC. Inter-annotator agreement (IAA) was measured as the F-score between the gold annotator and the other annotator, initially yielding only 0.59. IAA F-score increased to 0.84 after changing the annotation task to use system-generated annotations with which the annotators would either agree or disagree, as well as letting the annotators see each other's work.

The authors emphasize that the annotators expressed a large degree of uncertainty in their annotations. Furthermore, they note that the type of text contained in their corpus differs qualitatively from clinical text, for which the NegEx (Chapman et al., 2001) algorithm was originally developed, e.g. by being more generic. This might have made it more difficult for the annotators to determine what should be considered as negation and not and thus explain the initial low F-score for IAA.

5.1.1 Negation annotation by Sadhukhan vs. NoReC$_{neg}$

The annotation effort made by Sadhukhan (2021) largely differs from the NoReC$_{neg}$ (Mæhlum et al., 2021) project. Sadhukhan (2021) annotates negation only in relation to certain clinical medical expressions, more specifically terms belonging to the categories symptoms, findings and diseases. Examples of symptoms annotated as negated are *kvalm* 'nauseous', *kastet opp* 'thrown up' and *magesmerter* 'abdominal pain'. Belonging to the 'findings' category, we find *blødning* 'bleeding', *resttumor* 'residual tumor' and *stafylokokker* 'staphylococci', among others. We count *peritonitt* 'peritonitis' and *coloncancer* 'colon cancer' as terms related to disease. Other than this, Sadhukhan (2021) do not clearly describe the guidelines used in their annotation effort. However, they define negation detection as the detection of so-called 'pertinent negatives', defined by Chapman et al. (2001, pp. 301 - 302) as "findings and diseases explicitly or implicitly described as absent in a patient". As far as we understand, the task of the annotators has been to mark all negated expressions according to this definition.

Due to the strict view of what constitutes relevant negation, the corpus created by Sadhukhan (2021) contains sentences where negated

[3] https://github.com/DebaratiSJ/NegEx-on-Norwegian-biomedical-text/tree/main/Gold%20standard%20biomedical%20corpus

expressions are present but not annotated since they are not related to any of the mentioned categories. In NoReC$_{neg}$, on the other hand, no thematic restrictions are imposed on the annotation. Example (5.1) illustrates one case where negation is not annotated by Sadhukhan (2021) and also shows how it would have been annotated by Mæhlum et al. (2021):

(5.1) *Sadhukhan:* *Dette er imidlertid ikke dokumentert .*
NoReC$_{neg}$: *[Dette er] imidlertid* ***ikke*** *[dokumentert] .*
This is however not documented .
'However, this has not been documented.'

Also when considering only the sentences where negation *is* annotated by Sadhukhan (2021), annotations are often fundamentally different from NoReC$_{neg}$. In NGSBNC, there is no concept of 'cue' and 'scope'. Negated terms are marked by an opening tag (<NEGATED>) and a closing tag (</NEGATED>). We will refer to this as 'scope' and mark it similarly in our examples even though this term is not used. The corresponding 'cue' is not marked in any way, so the reader must infer from context which expression has triggered the negation. In a large number of cases, this seems to be negation triggers known from chapter 3, such as *ikke* 'not', *uten* 'without' and *ingen* 'no'/'nobody'.

The examples (5.2) - (5.5) illustrate some typical contrasts between the two annotation schemes. Where the annotations in NGSBNC are very specific as to what word or term is the target of negation, NoReC$_{neg}$ tends to include whole phrases, clauses or sentences in the scope.

(5.2) *Sadhukhan:* *Det ble ikke påvist andre [blødningskilder] .*
NoReC$_{neg}$: *Det [ble]* ***ikke*** *[påvist andre blødningskilder] .*
It became not proven other bleeding.sources .
'Other sources of bleeding were not identified.'

(5.3) *Sadhukhan:* *Inngrepet ble gjennomført uten at det*
NoReC$_{neg}$: *Inngrepet ble gjennomført* ***uten*** *[at det*
Operation.the became carried.out without that it
fremkom tegn på [blødning] ...
fremkom tegn på blødning] ...
appeared signs of bleeding ...
'The surgery was performed without any sign of bleeding...'

(5.4) *Sadhukhan:* *Tre pasienter uten [opplyst diabetestype]*
NoReC$_{neg}$: *[Tre pasienter]* ***uten*** *[opplyst diabetestype]*
Three patients without enlightened diabetes.type
ble ikke tatt med .
ble ikke tatt med .
became not taken with .
'Three patients without a stated diabetes type were not included.'

(5.5) *Sadhukhan:* ... *vi* *observerte* *ingen* *[komplikasjoner]* *hos* *mor*
NoReC$_{neg}$: ... *[vi* *observerte]* ***ingen*** *[komplikasjoner* *hos* *mor]*
... we observed no complications at mother
'...we observed no complications for the mother'

When annotating negation, there is a discussion regarding whether to annotate non-factual negation or not. Morante and Daelemans (2012) exclude negation in conditional, interrogative and imperative sentences from their annotation scheme, whereas NoReC$_{\text{neg}}$ do not limit their annotations by factuality because negation is assumed to be important regardless of this in tasks such as sentiment analysis (Mæhlum et al., 2021). It seems to be the case that Sadhukhan (2021) include cases with some uncertainty in their annotations. From a clinical perspective, it seems reasonable to only annotate the factual cases if the motivation is to train a negation system to identify which symptoms, findings, diagnoses etc. are *actually* not present in a patient. A system trained on data annotated according to the NoReC$_{\text{neg}}$ scheme could not have been used directly for this purpose; it would be necessary to be able to separate between factual and non-factual negation.

5.2 Annotating negation in the dataset

In the following, we describe the procedure of reannotating NGSBNC (Sadhukhan, 2021) according to the annotation guidelines of Mæhlum et al. (2021) and our review of these from chapter 4. The driving force behind this effort is the possibility to directly evaluate the negation models that we have trained (see chapter 3), as well as to assess the transferability of these models to the medical domain. First, we explain the technical details of the annotation process and the preprocessing we applied to the original corpus. Then, we account for the assumptions we have made in the annotation process. We introduce the name NorMed$_{\text{neg}}$ for the version of the dataset annotated by us, which we will use from here onwards.

5.2.1 Annotation tool and setup

We used the open-source annotation tool Brat (Stenetorp et al., 2012) when annotating. The Mæhlum et al. (2021) authors kindly gave us access to the configuration files from their project, providing us with the same setup. The corpus has been divided into sentences beforehand, and information on which sentences belong to the same document is not included. Thus, we keep this setup and annotate one sentence at a time. This fits well with the fact that the annotations according to NoReC$_{\text{neg}}$ do not cross sentence boundaries (Mæhlum et al., 2021).

Since the scopes annotated by Sadhukhan (2021) seem to nearly always be a part of the scopes annotated by us, we choose to mark them as the 'Focus' of its corresponding negation cue. Resolving the so-called 'Focus' of negation is a well-known task in NLP (Morante and Blanco, 2012).

Although this task is not performed in our work, we include the 'Focus' to preserve the information conveyed by the annotations made by Sadhukhan (2021) next to our own cue and scope annotations.

All sentences have been annotated by the author of this thesis. These annotations serve as the gold standard. In addition, a subset of sentences has been annotated by a second annotator for the computation of inter-annotator agreement. We elaborate on this in 5.2.2.

5.2.2 Inter-Annotator Agreement

We make a random selection of sentences from $\text{NorMed}_{\text{neg}}$ to be annotated by a second annotator in order to enable computation of inter-annotator agreement. 30 sentences are randomly selected for the purpose. Note that these are chosen from the set of *negated* sentences only. We avoid sampling sentences from the entire dataset, since this would generate much more work for this annotator by requiring a larger number of sentences to be reviewed in order to get a good basis for IAA computation. The annotator to whom we assign this task also participated in the $\text{NoReC}_{\text{neg}}$ project and thus is familiar with the original annotation guidelines already. They are also informed of our additional assumptions and modifications of these as described in chapter 5, subsection 5.2.4.

Inter-annotator agreement is computed using two measures: F_1 and κ. Each measure is computed for cues (0.941 F_1, 0.939 κ) and scope tokens (0.943 F_1, 0.927 κ). These metrics are token-based, i.e. each token of a sentence is represented by 0 (not annotated) or 1 (annotated). For affixal negation, if a subtoken of a word is annotated as a cue, the whole word is represented by a 1 when computing cue IAA. Subtokens inside scopes are treated correspondingly in the computation of scope IAA. In general, our results indicate high agreement between the annotators.

5.2.3 Preprocessing and cleaning of the dataset

Here, we describe the steps we made in order to clean the dataset. These include some initial preprocessing, removal of duplicate sentences and English sentences, and finally, manual correction of poor sentence splitting.

Initial preprocessing

As mentioned in 5.2.1, we preserve the original negation tags through our annotations, but the physical tags (<NEGATED> and </NEGATED>) are removed from the sentences. We also extract punctuation as separate tokens using the Python library spaCy[4]. Each sentence is then assigned a unique identifier.

[4] https://spacy.io/

Before	After
(1) For hele gruppen er bedring i sykdomsfri (2) overlevelse på 2—4 % som vil si circa (3) 2 % bedring i overlevelse etter fem år	For hele gruppen er bedring i [sykdoms]**fri** [overlevelse] på 2—4 % som vil si circa 2 % bedring i overlevelse etter fem år

Table 5.1: Preprocessing: combining fragments of a sentence ('Before') to a complete sentence, shown with its annotations ('After').

Removal of duplicates

Although Sadhukhan (2021) removed duplicates from the dataset, we still found several occurrences of identical sentences. We extracted a list containing tuples of matching sentences, where each tuple had a variable size N, and then removed N-1 of the identical occurrences of the sentence. This reduced the number of sentences from 2,330 to 2,281.

Removal of English sentences

During the annotation process, we found that the corpus contains a considerable number of non-Norwegian sentences, all of which were in English. We ran all sentences of the corpus through a language detector from the spacy_langdetect[5] pipeline and removed those that were recognized as English with a probability score above 0.8 (80 %). A quick manual review indicated that setting the threshold to this value led to few mistakes, i.e. few false positives and false negatives. This reduced the number of sentences from 2,281 to 2,060.

Correcting for suboptimal sentence splitting

We observe that the original sentence tokenization of NGSBNC (Sadhukhan, 2021) has some flaws. Occasionally, what is considered a 'sentence' in the corpus is only a part of a full sentence. In some cases, this directly affects our annotations by making it impossible to mark the whole scope of a negation cue when a sentence is split into multiple parts. In our cleaned version of the dataset, we decide to correct this manually, as illustrated by Table 5.1. Our corrections do not always result in full sentences, since we are sometimes not able to find all parts of the sentence.

As part of the cleaning process, we also split 'sentences' that actually consist of *multiple* sentences into their individual parts. An example of this is seen in Table 5.2 on the next page. We do not perform splitting in cases where we understand that headings have been included as part

[5] https://github.com/Abhijit-2592/spacy-langdetect

Before	After
Oppgitte interessekonflikter : Ingen Se også side 3055 Som lokalt residiv etter kirurgi for endetarmskreft regnes alle residiver i bekken og perineum	(1) Oppgitte interessekonflikter : (2) **Ingen** (3) Se også side 3055 (4) Som lokalt residiv etter kirurgi for endetarmskreft regnes alle residiver i bekken og perineum

Table 5.2: Preprocessing: splitting a collection of sentences, originally regarded as one ('Before'), into individual sentences, shown with their annotations ('After').

Before	After
KommentarSpontan hemoperitoneum2129Spontan hemoperitoneum defineres som blod i peritonealhulen av ikke-traumatisk årsak	Kommentar Spontan hemoperitoneum Spontan hemoperitoneum defineres som blod i peritonealhulen av **ikke-**[traumatisk årsak]

Table 5.3: Preprocessing. 'Before': page numbers and missing spaces. 'After' (annotated): removed page numbers and added missing spaces.

of the subsequent sentence, or in cases where pieces of text and numbers we believe must originate from a table are regarded as one sentence. This decision is made because we know that this 'problem' frequently occurs in datasets like ours and is difficult for sentence tokenizers to handle perfectly.

By inspecting the original format of the articles[6] from which the sentences are obtained, we understand that some of the sentences contain page numbers. We believe these should not have been included initially and remove them. Additionally, we add space between words where this is obviously missing. An example of these modifications is shown in Table 5.3.

All the mentioned modifications are applied only to the sentences that we identify to contain negation. It would require a significant amount of manual work to correct the non-negated sentences as well.

Two versions of the dataset

Removal of duplicates and English sentences followed by manual correction of poor sentence segmentation in negated sentences results in a new version of the dataset, which will be used for evaluation and further experiments in chapter 6. The original version, where only the initial preprocess-

[6] Available through searching here: https://tidsskriftet.no/sok

ing has been applied, is kept for an attempted comparison of our models to the model developed by Sadhukhan (2021). While the number of sentences in the original version is 2,330, the improved version ends at 2,086 after correction of sentence tokenization errors in negated sentences. We make the improved version available on GitHub under the name NorMed$_{neg}$.[7]

5.2.4 Assumptions added to the NoReC$_{neg}$ guidelines

In chapter 4, we discussed several cases where the annotation guidelines of the NoReC$_{neg}$ dataset could have been specified better. In the following, we state how we decide to treat these cases in the process of annotating NorMed$_{neg}$. We also provide some additional assumptions regarding questions arising from specific observations in the dataset. The reader can assume that we follow the original guidelines unless contrary information is given.

Affixal negation and lexicalization

We adhere strictly to the principle of only annotating affixal negation if the affixed word is clearly a negation of the word without the negation affix. For example, we do not mark *ulike* 'dissimilar, various, several' as negation in cases where it can be interpreted as meaning 'various, several'.

Affixally negated adjectives used as adverbs

In these cases, only the remaining part of the adverb is regarded as negated by the affix and thus annotated as the scope. This is regardless of whether the affixally negated adverb modifies a verb, an adjective or another adverb.

"The whole NP" as scope

Determiners Since, for affixal negation, all types of determiners are frequently included inside noun phrase scopes in the NoReC$_{neg}$ annotations, we choose to consistently adhere to this. We think that the scope annotations would have given a more precise description of what is actually negated if the determiner was left outside, but we do not want to deviate too much from their practice, since it forms a basis of comparison both in terms of corpus statistics and modeling results.

Genitive phrases In noun phrases such as *Pasientens **u**vanlige sykdom* 'The patient's **un**usual disease', we decide to include the genitive phrase in the scope, i.e. the scope will be *Pasientens vanlige sykdom* 'The patient's usual disease' rather than only *uvanlige sykdom* 'unusual disease'. We do this because there was an overweight of such cases in the NoReC$_{neg}$ dataset, and because this case resembles the case with possessive determiners.

[7] https://github.com/marieef/NorMed_neg/

Postmodifiers of noun phrases Relative clauses that belong to noun phrases modified by an adjective with a negation affix are included as part of the scope only when they are clearly restrictive in meaning, i.e. non-restrictive ones are not considered part of the scope. An example of a restrictive relative sentence is shown in (5.6). In this specific example, the relative sentence begins with the relative adverb *der* 'where'. (5.7), on the other hand, contains a relative sentence that we do not count as restrictive.

(5.6) *... behandling av ufrivillig [barn]**løse** [par der*
... treatment of involuntarily childless couples where
mannen har Klinefelters syndrom] .
man.the has Klinefelter's syndrome .
'... treatment of involuntarily childless couples where the male has Klinefelter's syndrome.'

(5.7) ***U**[standardiserte betaverdier] som angir størrelse og*
Unstandardized beta.values that indicate size and
retning på den aktuelle effekten ...
direction on the relevant effect ...
'Unstandardized beta values indicating size and direction of the relevant effect ...'

Prepositional phrase postmodifiers are treated similarly, i.e. we only include them inside the scope when we perceive them as essential to the meaning of the noun phrase head. An example of a prepositional phrase postmodifier included in the scope is shown in (5.8). A counterexample can be seen in (5.9); here, we think that the prepositional phrase *i PubMed* 'in PubMed' can even be thought of as an adverbial in the sentence.

(5.8) *I mange kreftceller er det **u**[kontrollert distribusjon av*
In many cancer.cells is there uncontrolled distribution of
arvemateriale] ...
heir.material ...
'In many cancer cells, there is an uncontrolled distribution of genetic material...'

(5.9) *I tillegg ble det utført [et] **ikke-**[systematisk*
In addition became it performed a non-systematic
litteratursøk] i PubMed .
literature.search in PubMed .
'In addition, a non-systematic literature search was performed in PubMed.'

We treat postmodifiers of noun phrases this way because we think it is reasonable to distinguish between necessary/restrictive and unnecessary/non-restrictive, and from what we saw in chapter 4, we believe that the annotators at least partly have tried to make this distinction as well.

Additional adjectives and adverbs in affixal negation

In affixal negation, we leave out additional adjectives and adverbs from negation scopes, following our understanding of Mæhlum et al. (2021). This applies both to noun phrase scopes and clausal scopes. We make a few exceptions in phrases such as ***a**[typiske kliniske funn]* '**a**[typical clinical findings]' and ***u**[vanlig kirurgisk tilstand]* '**un**[usual surgical condition]', where we consider *kliniske funn* 'clinical findings' and *kirurgisk tilstand* 'surgical condition' as fixed expressions.

Copula verbs

We consider all verbs that connect the subject to the predicate as copula verbs. This includes *å være* 'to be' and *å bli* 'to become', which are most typically regarded as copula verbs, but also verbs such as *å fremstå (som)* 'to appear (as)', *å virke* 'to seem', *å synes* 'to seem', *å oppfattes (som)* 'to be perceived (as)' and *å bli ansett som* 'to be regarded as'. This means that in sentences with one of these verbs as the main verb and an affixally negated adjective as the predicate, the full clause will be the scope.

Negation affixes derived from Greek and Latin

Texts belonging to the medical domain tend to contain domain-specific terminology, much of which is derived from Greek and Latin (Dalianis, 2018). This includes words containing the prefixes *a-*, *ab-*, *an-*, *im-* and *in-* . *In-* is the only one of these to be found in NoReC$_{\text{neg}}$, but note that we have not had the capacity to control whether there are any unannotated occurrences of these cues. While all the mentioned negation affixes occur in NorMed$_{\text{neg}}$, *a-* and *in-* are the most common ones. Examples (5.10) and (5.11) illustrate two sentences containing these cues.

(5.10) *Imidlertid [er halvparten av pasientene]* ***a**[febrile ved*
However are half.part.the of patients.the afebrile by
diagnosetidspunktet] .
diagnosis.time.the .
'However, half of the patients are afebrile at the time of diagnosis.'

(5.11) *Kirurgi av lokalt residiv ved* ***in**[kurable fjernmetastaser]*
Surgery of local recurrence by incurable distant.metastases
'Surgery for local recurrence in incurable distant metastases'

In this annotation effort, we choose to include all of the aforementioned prefixes as affixal negation, given that the following criteria apply:

1. The case is actually a matter of negation and not something similar.
2. The word negated by the affix can be used by itself.

Based on criterion 1, we exclude affixes such as *anti-* in *antikoagulasjon* 'anticoagulation', since we interpret this as meaning something more than 'only' the negation of *koagulasjon* 'coagulation'. This is supported by the fact that $\text{NoReC}_{\text{neg}}$ contains 18 occurrences of this prefix in combination with nouns or adjectives, but none of them are annotated. Note also that the other affixes mentioned can be part of words where they have nothing to do with negation. An example is *in-* in *inflammasjon* 'inflammation'.

The guidelines created by Mæhlum et al. (2021) state that nominalizations of affixally negated adjectives such as *ulykkelighet* 'unhappiness' are not to be annotated. One option would be to apply the same principle to nouns corresponding to adjectives negated by Greek or Latin prefixes. We believe that morphological structure of a word such as *instabilitet* 'instability' to be *((in+stabil)+itet)* '((in+stable)+ity)' and not *(in+(stabil+itet))* '(in+(stable+ity))'. It might also be the case with *inkontinens* 'incontinence' that the noun is derived from the adjective *inkontinent* 'incontinent'. However, these nouns meet the two criteria we set above. Furthermore, the information they represent is interesting from a medical point of view. As an example, for healthcare personnel working in a nursing home, it is valuable to know if the patients are continent or incontinent since incontinence has certain practical implications.

To summarize, we choose not to take into account the order of operations in the word derivation process for these words. Thus, we annotate negation in adjectives such as *instabil* 'unstable', *inkontinent* 'incontinent' and *abnormal* 'abnormal', as well as the corresponding nouns *instabilitet* 'instability', *inkontinens* 'incontinence' and *abnormalitet* 'abnormality'.

Note that when we annotate these nominalizations, we do not include adjectives and adverbs to the left. In (5.12), for instance, *genetisk* 'genetic' is excluded from the scope of the negation triggered by *in-* 'in-'. This is in accordance with the principle of leaving out additional adjectives and adverbs in cases of affixal negation (Mæhlum et al., 2021). Also, we decide not to include prepositional phrase postmodifiers in these scopes. We see that these often resemble adverbials in the sense that the sentence could be paraphrased by moving the prepositional phrase away from the noun phrase: for example, 'There was an instability in certain signaling pathways' paraphrased as 'In certain signaling pathways, there was an instability', in which case only 'stability' would be inside the scope.

(5.12) *Kolorektal kreft kjennetegnes ved genetisk* ***in****[stabilitet] i*
Colorectal cancer is.characterized by genetic instability in
bestemte signalveier .
certain signal.ways .
'Colorectal cancer is characterized by genetic instability in certain signaling pathways.'

The scope of *uten* 'without'

Concerning the negation cue *uten* 'without', we decide to follow the most common practice in each of the cases described in chapter 4. I.e., for any type of prepositional phrase with *uten* 'without' as the head, if it functions as an adverbial in the sentence, the scope will be the complement of the preposition. When a prepositional phrase consisting of *uten* 'without' and a noun phrase serves as a postmodifier of a noun phrase, this whole noun phrase is the scope. In copula sentences where *uten* 'without' + noun phrase is the predicate, we include the subject and verb in the scope, in addition to the prepositional object.

Lexical negation cues

In 4.1.3, we address the handling of lexical negation in NoReC$_{neg}$. They have identified several lexical negation cues that we agree can trigger negation. However, it is necessary to evaluate each specific occurrence individually to decide whether to mark negation.

As for the verbs *forhindre* 'prevent' and *hindre* 'prevent', we decide not to include them as negation triggers due to the modal aspect of volition to these verbs, as mentioned in 4.1.3. Furthermore, we believe this applies to the verb *slippe* 'not have to' and most occurrences of *unngå* 'avoid' and *nekte* 'refuse, deny', so we decide not to annotate these cases.

Regarding *fjerne* 'remove', we judge that it can trigger negation. In examples (5.13) and (5.14), we have two examples with *fjerne* 'remove' and its nominalization *fjerning* 'removal' that we will annotate. We will not annotate this cue in cases where the removal is not complete, as in (5.15).

(5.13) *Man måtte* ***fjerne*** *[758 divertikler] for å hindre ett*
One had.to remove 758 diverticula for to prevent one
dødsfall .
death.fall .
'758 diverticula had to be removed in order to prevent one death.'

(5.14) *Kirurgisk* ***fjerning*** *av [tilfeldige divertikler] førte til*
Surgical removal of random diverticula led to
signifikant flere komplikasjoner ...
significantly more complications ...
'Surgical removal of random diverticula led to significantly more complications...'

(5.15) *... slik at triglyseridene raskere fjernes fra plasma*
... so that triglycerids.the faster are.removed from plasma
'...so that the removal of triglycerids from plasma is faster'

Another case we identify as lexical negation is seen in (5.16). *Opphevet* 'canceled, revoked' does not occur as a cue in NoReC$_{neg}$, but is quite similar in meaning to *forsvinne* 'disappear', which is present as a cue. We also considered whether to mark negation in the sentence in (5.17), but refrained

from it since we are not certain this represents a complete block of the signaling. (5.18) contains a word that could serve as a negation cue (*tap* 'loss'). However, in the sentence in question, there is not a complete loss of tumor cells, but rather a *reduction*.

(5.16) *... **opphevet** [evne til åpning av den anorektale vinkelen] .*
... canceled ability to opening of the anorectal angle .
'...loss of the ability to open the anorectal angle.'

(5.17) *Flere hemmere har som hensikt å blokkere*
Several inhibitors have as purpose to block
VEGF-indusert signalaktivering .
VEGF-induced signal.activation .
'Several inhibitors serve to block VEGF signaling.'

(5.18) *... forstyrrer balansen mellom tilvekst og tap av*
... disturbs balance.the between growth and loss of
tumorceller .
tumor.cells .
'...disturbs the balance between growth and loss of tumor cells.'

We choose to also include as cues in $\text{NorMed}_{\text{neg}}$ certain words that are not present in the vocabulary of $\text{NoReC}_{\text{neg}}$ at all. First, there is the verb *utelukke* 'rule out', of which an example sentence is given in (5.19).

(5.19) *[Blødningstilstander] er viktig å **utelukke** .*
Bleeding.conditions are important to out.close .
'Ruling out bleeding conditions is important.'

Furthermore, there is the verb *ekskludere* 'exclude' and its corresponding noun *eksklusjon* 'exclusion'. We conclude that they should be annotated as cues as well, both when their meaning is similar to *utelukke* 'rule out' and when they mean 'exclude, not include'. We also annotate *utelate* 'exclude' as a cue.

(5.20) *Typisk klinisk bilde samt sikker **eksklusjon** av [andre*
Typical clinical picture and secure exclusion of other
mulige årsaker] kan gi diagnosen ...
possible causes can give diagnosis.the ...
'The diagnosis can be established on the basis of a typical clinical picture and certain exclusion of other possible causes.'

In addition, we include the adjective *negativ* 'negative' as a negation trigger when used as in example (5.21). To our understanding, this is close to equivalent to *Hemofec ga ingen funn* 'Hemofec gave no findings', meaning that what the Hemofec test is designed to detect, i.e. blood in the patient's stool, was not found. Formulated as *Hemofec var negativ* 'Hemofec was negative', we would not consider this a case of negation. In that case, we would interpret *negativ* 'negative' simply as an adjective used to describe the test result.

(5.21) *[Hemofec ga]* ***negative*** *[funn]* .
Hemofec gave negative findings .
'Hemofec was negative.'

Lexical negation scopes

In general Generally, we think that the scope should cover what is described as absent. The question is whether to include more elements.

Nominal cues For nominal lexical cues we follow the practice seen in NoReC$_{neg}$, i.e. in expressions such as *mangel på X* 'lack of X', *eksklusjon av X* 'exclusion of X' and *fjerning av X* 'removal of X', the scope will be the whole phrase represented by X.

Participle cues For present participles used as lexical cues, i.e. *manglende* 'lacking' and *fraværende* 'absent', we generally follow the patterns described for these cues in the part on scope resolution for lexical negation in 4.1.3. An example of this is (5.22). One case, (5.23), is treated differently because it seems reasonable to apply the rule on subjects modified by negation cues (Mæhlum et al., 2021).

(5.22) ***Manglende*** *[stadieinndeling]* *gjør* *det* *vanskelig* *å*
Lacking stadium.division does it difficult to
sammenlikne *resultatene* *...*
compare results.the ...
'Lack of staging makes it hard to compare the results...'

(5.23) *Det* *[foreligger]* ***manglende*** *[data* *for* *reresidivfaren*
It is.present lacking data for re-recurrence.danger.the
etter *R1-reseksjoner]* .
after R1.resections .
'There is a lack of data concerning the risk of re-recurrence after R1 resections.'

We encountered one sentence, (5.24) where this type of cue modifies a noun phrase with a determiner. In this case we choose to include the determiner in the scope based on a similar example in the NoReC$_{neg}$ dataset and similar to how we treat affixal negation in noun phrases.

(5.24) *Dersom* *vi* *antar* *at* *[alle]* ***manglende*** *[svar* *på*
If we assume that all lacking answers on
glukoseintoleransetest *etter* *fødsel* *er* *negative]* *...*
glucose.intolerance.test after birth are negative ...
'If we assume that all missing glucose intolerance test results after birth are negative...'

Inclusion of the subject We saw in 4.1.3 that the subject is often included in the scope of lexical negation verbs in constructs such as 'X lacks Y', 'X loses Y' etc. Although there are not many sentences of this form in $\text{NorMed}_{\text{neg}}$, as a main rule, we include the subject where applicable. This is illustrated by (5.25).

(5.25) *[Ekstern validering] er viktig , noe [denne studien]*
External validation is important , something this study
mangler .
lacks .
'External validation is important, something that this study lacks.'

Inclusion of adverbials Adverbials are included where we assess them as a natural part of the scope. An example of this is seen in (5.26). In (5.27), *(fra) deltakelse* '(from) participation' is not included in the scope.

(5.26) *[Svar]* ***manglet*** *[hos henholdsvis 43 % og 35 %] .*
Answers lacked at respectively 43 % and 35 % .
'Responses lacked in 43 % and 35 %, respectively.'

(5.27) *... [pasienter som er under antikoagulasjonsbehandling] har*
... patients who are under anticoagulation.treatment have
vært ***utelukket*** *fra deltakelse .*
been excluded from participation .
'...patients who are undergoing anticoagulation treatment have been excluded from participation.'

The scope of *fjerne*, *ekskludere* and similar verbs Because verbs like *fjerne* 'remove', *ekskludere* 'exclude' and *utelukke* 'exclude' do not occur as cues in $\text{NoReC}_{\text{neg}}$, we have no examples to use as a basis for the scope annotation. Thus, we decide to limit the scope to what is actually described as absent, i.e. the object argument of the verb. Different from verbs such as *mangle* 'lack', the subject argument of *fjerne* 'remove' generally does not represent the entity that does not have the negated item, but rather who is responsible for the absence of this item. Annotation examples are shown in (5.28), (5.29) and (5.30).

(5.28) *... kan man ikke* ***utelukke*** *[at det finnes fokal*
... can one not rule.out that it exists focal
spermatogenese]
spermatogenesis
'... one cannot rule out that focal spermatogenesis exists'

(5.29) *[Nukleolene] ble* ***fjernet*** *fra oocytter hos gris og*
Nucleoli.the became removed from oocytes at pig and
mus ved mikrokirurgi
mouse by microsurgery
'The nucleoli were removed from oocytes in pigs and mice through microsurgery'

(5.30) *To meget viktige momenter blir å* ***ekskludere***
Two very important points become to exclude
[pasienter hvor makroskopisk cancer må etterlates ved
patients where macroscopic cancer must be.left.behind at
operasjonen] ...
operation.the ...
'Two very important elements will be to exclude patients where macroscopic cancer must be left behind during the operation...'

Factuality

All types of negation, including both factual and non-factual propositions, are annotated in $\text{NoReC}_{\text{neg}}$ (Mæhlum et al., 2021). Although not stated explicitly in their paper, we interpret this as applying to all types of negation, i.e. affixal, lexical and syntactic.

Negation in compounds?

We do not annotate negation in compounds. Examples of compounds are *urininkontinens* 'urinary incontinence', *inkontinenstilstanden* 'the incontinence condition' and *serotoninmangel* 'serotonin deficiency'. This decision is made to maintain simplicity and because it would create problems when evaluating our models. As described in chapter 3, our models are trained on the word level. This means that for affixal negation, they do not recognize affixes directly, but rather the whole word containing the negation affix. The cue and scope part of words containing affixal negation is inferred by the use of regular expressions that recognize certain known negation prefixes and suffixes. It is not straightforward to detect negation affixes in the middle of a words, such as *in-* 'in-' in example (5.31). In (5.32), automatic affix extraction would infer *kontinenstilstanden* 'the continence condition' as the scope. The problem is, the word *inkontinenstilstanden* 'incontinence condition' is a compound consisting of *inkontinens* 'incontinence' and *tilstanden* '(the) condition' and thus it is incorrect to count the head of the compound as part of the scope. Last, we have the compound in example (5.33), which would be possible to recognize if *mangel* 'lack' was added to the list of suffixes. This, however, denotes *reduced levels* of serotonin and not a total lack or absence of the substance. Hence, it is not a case of negation at all, and this seems to be the case in all compounds with *mangel* 'lack' as the head. As we have not observed any other compounds with negation, we consider the exclusion of such constructs from our annotation effort to be a minor problem.

(5.31) *Risikoen er perforasjon av urethra eller skade av*
Risk.the is perforation of urethra or damage of
*urethralsfinkter med urin****in****kontinens .*
urethral.sphincter with urine.incontinence .
'The risk is perforation of the urethra or damage to the urethral sphincter leading to urinary incontinence.'

(5.32) *... de ... lar* ***in****kontinenstilstanden påvirke hverdagen*
... they ... let incontinence.condition.the affect everyday.the
i stor grad .
in large degree .
'...they ... let their incontinence affect everyday life to a large extent.'

(5.33) *Testen ble gjennomført både på gruppen som hadde*
Test.the became conducted both on group.the that had
*serotonin****mangel*** *og på kontrollgruppen .*
serotonin.lack and on control.group.the .
'The test was carried out on both the group with serotonin deficiency and the control group.'

Adverbials interfering with negation

In the guidelines, it is stated that sentential adverbs such as *dessverre* 'unfortunately' are not to be included in negation scopes. It is also mentioned that modal expressions occurring in combination with a negation cue will be kept outside the scope. We will add the assumption that this also applies to adverbials such as the one in example (5.34), *i enkelte tilfeller* 'in some cases'. We arrive at this conclusion by paraphrasing the sentence in (5.34): (1) 'In some cases, it is not the case that a certain cause is detected'. Note the difference between this and (2) 'It is not the case that in some cases, a certain cause is detected'. The paraphrase in (1) is in accordance with the original sentence, while (2) is not, and thus we decide to keep the adverbial outside the scope.

(5.34) *I enkelte tilfeller [påvises]* ***ingen*** *[sikker årsak] .*
In individual cases is.detected no certain cause .
'In some cases, no certain cause is detected.'

Including the subject/object in the scope

In cases such as (5.35) and (5.36), we include the part of the sentence described by the negated adjective, i.e. the subject or object, inside the scope. We do not have much data from $\text{NoReC}_{\text{neg}}$ to base this on, but to us it seems reasonable to do so.

(5.35) *... kan ... preoperativ bestråling ... gjøre [slik reseksjon]*
... may ... preoperative radiation ... do such resection
u*[nødvendig] .*
unnecessary .
'... preoperative radiation may make such resection unnecessary.'

(5.36) *Mens [sykehusdriften] før reformen ofte ble*
While hospital.operation before reform.the often became
kritisert for å være ***in****[effektiv] ...*
criticized for to be inefficient ...
'While the running of the hospital before the reform was often criticized for inefficiency...'

5.3 Corpus statistics

In this section, we analyze $\text{NorMed}_{\text{neg}}$, using mostly quantitative methods. Where we make interesting and contrasting findings in our dataset and $\text{NoReC}_{\text{neg}}$, we attempt to discuss these in light of the differences between the underlying domains, i.e. medical research literature and review articles. In Tables 5.5 to 5.7 on page 93, we include numbers reported by Mæhlum et al. (2021) and use the data analysis script from their GitHub repository to ensure that our results are comparable to theirs.[8]

5.3.1 Vocabulary

We perform a lexical analysis of the two corpora to find out to what extent domain-specific words are present in their vocabularies. This is done by counting the frequencies of all present lemmas using a Norwegian NLP pipeline from spaCy[9], which includes a lemmatizer.[10] Note that punctuation, words recognized as numbers and stopwords[11] are disregarded. Furthermore, we lowercase all lemmas prior to counting, since we observe that the lemmatizer is not consistent in this regard. We also take parts of speech into account, e.g. *få* 'get, few' gives rise to three different lemmas corresponding to the three parts of speech it can have, i.e. auxiliary verb, main verb and adjective.[12] In total, 5610 unique lemmas were identified in $\text{NorMed}_{\text{neg}}$, and 23773 in $\text{NoReC}_{\text{neg}}$.

In Table 5.4 on page 92, we have listed the most frequent lemmas in each dataset in descending order. In $\text{NorMed}_{\text{neg}}$, we recognize several words with strong links to the medical domain, such as *pasient* 'patient', *behandling* 'treatment' and *sykehus* 'hospital'. As for $\text{NoReC}_{\text{neg}}$, fewer domain-specific lemmas are found among the most common ones, but there are some, e.g. *film* 'movie', *spill* 'game' and *historie* 'history, story'. We also notice a difference in the distribution of parts of speech among the top lemmas. In $\text{NorMed}_{\text{neg}}$, 13 of them are common nouns; in $\text{NoReC}_{\text{neg}}$, only 6. The

[8] Our version of their script, which calculates sentence-level statistics for *negated* sentences as well: https://github.com/marieef/master-thesis_code/blob/main/data_analysis/data_analysis.py

[9] https://spacy.io/

[10] The code we used is available in our GitHub repository: https://github.com/marieef/master-thesis_code

[11] We used a list of Norwegian stopwords from NLTK: https://www.nltk.org/.

[12] The PoS tagset used is obtained from Universal Dependencies and consists of 17 tags. An overview of the tagset is provided here: https://universaldependencies.org/u/pos/

number of verbs, excluding auxiliary verbs, sums to 4 in $\text{NorMed}_{\text{neg}}$ and 9 in $\text{NoReC}_{\text{neg}}$.

5.3.2 Sentence-level analysis

Table 5.5 on page 93 compares sentence-level statistics in $\text{NorMed}_{\text{neg}}$ and $\text{NoReC}_{\text{neg}}$. In terms of number of sentences, the latter is more than five times larger than the former. Remember, however, from earlier that the sentence segmentation in $\text{NorMed}_{\text{neg}}$ is far from optimal and has been improved manually only for the sentences actually containing negation. The suboptimal sentence tokenization will affect the number of sentences in our corpus, but since the errors are both of the kind that splits a sentence into several pieces and of the kind that considers multiple sentences *one* sentence, we assume that the number 2,086 is not too far from the 'true' number of sentences. Out of these sentences, 448 (21.5 %) are *negated*, i.e. they contain at least one negation. This is quite close to the proportion of negated sentences in $\text{NoReC}_{\text{neg}}$, which is 20.6 %.

Sentence tokenization errors also have an effect on the numbers reported as average and maximum sentence length for $\text{NorMed}_{\text{neg}}$ in Table 5.5 on page 93. As we can see, the longest sentence of our dataset is 2.5 times longer than the longest one in $\text{NoReC}_{\text{neg}}$. The mean length of 17.7 tokens in $\text{NorMed}_{\text{neg}}$ is just slightly longer than 16.8, which is the average sentence length in $\text{NoReC}_{\text{neg}}$. However, we compute the average and max length of *negated* sentences as well and see that negated sentences are generally longer in $\text{NoReC}_{\text{neg}}$ (23.1 tokens) than in $\text{NorMed}_{\text{neg}}$ (20.6 tokens).

Furthermore, we examine the distribution of the two variants of written Norwegian. While $\text{NoReC}_{\text{neg}}$ contains a small amount of documents in the minority variant, Nynorsk, manual inspection of $\text{NorMed}_{\text{neg}}$ tells us that only the majority variant, Bokmål, is present.

5.3.3 Cues

Table 5.6 on page 93 contains key numbers describing the frequency of negation cues in $\text{NorMed}_{\text{neg}}$ and $\text{NoReC}_{\text{neg}}$. There is a total of 510 cues in $\text{NorMed}_{\text{neg}}$, out of which 35.5 % are affixal cues. Since this is quite remarkably more than the corresponding proportion in $\text{NoReC}_{\text{neg}}$ (24.9 %), we consider this an interesting finding. However, we keep in mind that our annotations include Greek and Latin negation affixes in nominalizations, as discussed in 5.2.4. These do not occur in $\text{NoReC}_{\text{neg}}$, but would probably not have been annotated if they did, according to their rule on not annotating affixal negation in nominalized adjectives (Mæhlum et al., 2021). Furthermore, we note that the percentage of negated sentences containing multiple negation cues is quite similar in the two corpora, with 12.5 % in $\text{NorMed}_{\text{neg}}$ and 13.0 % in $\text{NoReC}_{\text{neg}}$ (Mæhlum et al., 2021). The maximal number of tokens in a cue is 2 in $\text{NorMed}_{\text{neg}}$, whereas in $\text{NoReC}_{\text{neg}}$, it is 3. Cues containing more than one token are rare and thus, the average token-level cue length is 1 in both datasets. In $\text{NorMed}_{\text{neg}}$,

NorMed$_{neg}$				NoReC$_{neg}$			
Lemma	**PoS**	**Trans.**	**Fq.**	**Lemma**	**PoS**	**Trans.**	**Fq.**
pasient	NOUN	patient	299	mye	ADJ	much	988
hos	ADP	at, by	194	god	ADJ	good	850
behandling	NOUN	treatment	167	få	VERB	get	541
gi	VERB	give	120	hel	ADJ	whole	490
vise	VERB	show	107	gjøre	VERB	do	485
god	ADJ	good	101	stor	ADJ	big	469
annen	DET	other	100	film	NOUN	movie	467
residiv	NOUN	recurrence	95	se	VERB	see	458
kvinne	NOUN	woman	86	annen	DET	other	449
mye	ADJ	much	80	komme	VERB	come	381
stor	ADJ	big	77	litt	ADJ	little	359
år	NOUN	year	73	ta	VERB	take	353
få	VERB	get	72	år	NOUN	year	353
liten	ADJ	small	71	liten	ADJ	small	341
lokal	ADJ	local	70	ny	ADJ	new	324
høy	ADJ	high, tall	69	gå	VERB	walk, go	301
ofte	ADJ	often	65	gi	VERB	give	294
pankreatitt	NOUN	pancreatitis	65	the*	PROPN	the	279
kirurgisk	ADJ	surgical	64	all	DET	all	276
sykehus	NOUN	hospital	64	gang	NOUN	time, corridor	266
burde	AUX	should	62	to	NUM	two	258
viktig	ADJ	important	62	måtte	AUX	have to	252
årsak	NOUN	cause	61	nok	ADV	presumably**	239
to	NUM	two	60	lang	ADJ	long	235
tidlig	ADJ	early	59	tid	NOUN	time	233
type	NOUN	type	56	første	ADJ	first	229
kirurgi	NOUN	surgery	55	spill	NOUN	game	224
mann	NOUN	man	55	spille	VERB	play	208
lege	NOUN	doctor	54	historie	NOUN	history, story	183
perforasjon	NOUN	perforation	53	bruke	VERB	use	171
påvise	VERB	detect	53	kanskje	ADV	maybe	171

Table 5.4: List of the most frequent lemmas in NorMed$_{neg}$ and NoReC$_{neg}$ (Mæhlum et al., 2021), sorted by descending frequency. The Norwegian lemma, its English translation and the raw frequency ('Fq.') is reported for each dataset. For NorMed$_{neg}$ and NoReC$_{neg}$, the list includes all lemmas with at least 53 and 171 occurrences, respectively. *: 'the' is English, **: *nok* has more meanings, e.g. 'enough', 'yet'.

	Sentences					
Dataset	#	neg	avg	max	avg_{neg}	max_{neg}
$NorMed_{neg}$	2,086	448 (21.5%)	17.7	259	20.6	259
$NoReC_{neg}$	11,346	2,332 (20.6%)	16.8	103	23.1	103

Table 5.5: Corpus statistics – sentences containing negation. We compare $NorMed_{neg}$ to $NoReC_{neg}$. For each dataset, we report the total number of sentences (#), the raw frequency of negated sentences, as well as average and maximum length of *all* sentences and *negated* sentences, respectively.

	Cues					
Dataset	#	avg	max	disc	mult	affix
$NorMed_{neg}$	510	1	2	1 (0.0%)	56 (12.5%)	181 (35.5%)
$NoReC_{neg}$	2,672	1	3	21 (0.8%)	304 (13.0%)	665 (24.9%)

Table 5.6: Cue statistics for $NorMed_{neg}$ and $NoReC_{neg}$. We report the raw frequency, average and maximum length in tokens, frequency of discontinuous cues, sentences with multiple cues (mult) and frequency of affixal cues. Relative frequencies are in parentheses. For 'mult', this is the proportion of *negated* sentences containing more than one cue.

	Scopes					
Dataset	#	avg	max	disc	t.disc	null
$NorMed_{neg}$	503	5.7	40	283 (56.3%)	80 (15.9%)	7 (1.4%)
$NoReC_{neg}$	2,635	6.9	53	1,842 (69.9%)	566 (21.5%)	37 (1.4%)

Table 5.7: Scope statistics for $NorMed_{neg}$ and $NoReC_{neg}$. We report the raw frequency, average and maximum length in tokens, as well as frequency of discontinuous, true discontinuous and implicit scopes (null). Relative frequencies are in parentheses.

		NorMed$_{neg}$		NoReC$_{neg}$	
Cue	**Trans.**	Freq$_{raw}$	Freq$_{\%}$	Freq$_{raw}$	Freq$_{\%}$
ikke	not	178	34.9	1,364	51.0
u-	un-/dis-/non-	82	16.1	514	19.2
in-	in-	48	9.4	2	0.0
uten	without	46	9.0	190	7.1
ingen	none/nobody	36	7.1	134	5.0
ikke(-)-	non-	22	4.3	22	0.8
a-	a-	11	2.2	0	0.0
fjerne	remove	11	2.2	0	0.0
utelukke	exclude	11	2.2	0	0.0
mangle	lack (V)	10	2.0	43	1.6
-fri	-free/-less	7	1.4	13	0.5
aldri	never	5	1.0	95	3.6
fjerning	removal	4	0.8	0	0.0
-løs	-less	4	0.8	123	4.6
eksklusjon	exclusion	3	0.6	0	0.0
ab-	ab-	3	0.6	0	0.0
mangel	lack (N)	3	0.6	8	0.3
utelate	exclude	3	0.6	0	0.0
nei	no (interj.)	3	0.6	10	0.4

Table 5.8: All cues with at least 3 occurrences in NorMed$_{neg}$, sorted by descending frequency in NorMed$_{neg}$. Each cue is reported in the base form (infinitive for verbs (V), singular indefinite for nouns (N)) and is accompanied by its English translation. The four rightmost columns report the frequency as a cue in NorMed$_{neg}$ and NoReC$_{neg}$: raw frequency as a cue and relative frequency in the pool of cues, given in %.

there is one single occurrence of a discontinuous cue, which rounds to 0.0 % of all cues. This phenomenon is infrequent in NoReC$_{neg}$ as well, but here, they make up 0.8 %.

A more detailed picture is painted by Table 5.8. Here, we list all cues occurring at least three times in NorMed$_{neg}$, along with their raw and relative frequencies in NorMed$_{neg}$ and NoReC$_{neg}$. It is worth noting that the cue *ikke* 'not' makes up just above one third of all cues in NorMed$_{neg}$, while in NoReC$_{neg}$ (Mæhlum et al., 2021), it is more common than all other cues combined (51.0 %). The relative frequencies of *uten* 'without' and *ingen* 'none/nobody' are slightly higher in NorMed$_{neg}$, while *aldri* 'never' has a higher relative frequency in NoReC$_{neg}$.

As previously mentioned, the two corpora differ in the amount of affixal negation, and we know from 5.2.4 on page 82 that we annotate some negation affixes that do not exist in NoReC$_{neg}$. Thus, we include Table 5.9 on the facing page to examine the distribution of the various affixal cues more closely. It lists all affixal cues present in the union of the two datasets,

		NorMed$_{neg}$		NoReC$_{neg}$	
Cue	**Trans.**	Freq$_{raw}$	Freq$_{\%}$	Freq$_{raw}$	Freq$_{\%}$
a-	a-	11	6.1	0	0.0
ab-	ab-	3	1.7	0	0.0
an-	an-	3	1.7	0	0.0
-fri	-free/-less	7	3.9	12	1.8
ikke(-)-	non(-)-	22	12.2	7	1.1
im-	im-	1	0.6	0	0.0
in-	in-	48	26.5	2	0.3
-løs	-less	4	2.2	123	18.5
mis-	mis-	0	0.0	1	0.2
-tom	-tom	0	0.0	1	0.2
u-	un-/dis-/non-	82	45.3	514	77.3
utenom-	extra-	0	0.0	2	0.3

Table 5.9: List of all affixal cues present in the dataset annotated by us (NorMed$_{neg}$) and NoReC$_{neg}$ (Mæhlum et al., 2021), accompanied by raw and relative (%) frequency as cues in the two datasets. Relative frequency is computed from the total number of *affixal* cues in each dataset. All inflected forms of the same suffix are represented by the masc./fem. singular form.

reports for each cue the raw and relative frequency and visualizes several large differences between the datasets. We note that *u-* 'un-/dis-/non-' by far is the most common affixal cue in both corpora, but still makes up a much larger proportion of the affixal cues in NoReC$_{neg}$, 77.3 %, against 45.3 % in NorMed$_{neg}$. The second most common affixal cue in NorMed$_{neg}$, the Latin prefix 'in-', is barely present in NoReC$_{neg}$. Additionally, we note that the suffix *-løs* '-less' occurs rarely in NorMed$_{neg}$, but accounts for nearly one fifth of all affixal cues in NoReC$_{neg}$. The proportion of affixal cues comprised by the prefix *ikke(-)-* 'non(-)-', with or without a hyphen, is 11 times larger in NorMed$_{neg}$ than in the other dataset.

In Table 5.10 on the next page, we provide an overview of all cues found in NorMed$_{neg}$ that are absent as cues in NoReC$_{neg}$, accompanied by an example from NorMed$_{neg}$. As we can see, the examples are in general thematically connected to clinical medicine and medical research.

Cue	Example	Example (trans.)
a-	Menn med **a**zoospermi	Men with **a**zoospermia
ab-	**ab**normalt fungerende	**ab**normally functioning
an-	**an**euploide svulster	**an**euploid tumors
ekskludere	**ekskludere** pasienter	**exclude** patients
eksklusjon	**eksklusjon** av andre årsaker	**exclusion** of other causes
fjerne	**fjernede** polypper	**removed** polyps
fjerning	**fjerning** av levermetastaser	**removal** of liver metastases
im-	De fleste er **im**motile	Most are **im**mobile
negativ	**negative** funn	**negative** findings
oppheve	**opphevet** evne til åpning	**lost** ability to open
utelate	kontrollgrupper **utelates**	control groups are **omitted**
utelukke	**utelukke** premalignitet	**exclude** premalignancy

Table 5.10: This table contains a list of all cues found in $\text{NorMed}_{\text{neg}}$ that are *not* present as cues in $\text{NoReC}_{\text{neg}}$. Verbs are given in the infinitive, nouns in the singular indefinite form and adjectives in the singular masc./fem. form. The second column represents a short example from $\text{NorMed}_{\text{neg}}$ and the third its English translation.

5.3.4 Scopes

Table 5.7 on page 93 presents statistics of negation scopes and tells us that the total number of annotated scopes in $\text{NorMed}_{\text{neg}}$ is 503. We remember that the number of cues is 510, meaning that there are seven negation cues with an empty, so-called *implicit* scope. The relative frequency of empty scopes among all scopes is identical in $\text{NorMed}_{\text{neg}}$ and $\text{NoReC}_{\text{neg}}$, 1.4 %. In the latter, the maximum scope length and the average scope length are both larger than in $\text{NorMed}_{\text{neg}}$. It is especially interesting that the difference in average scope length is as large as 1.2: 5.7 tokens in $\text{NorMed}_{\text{neg}}$ and 6.9 in $\text{NoReC}_{\text{neg}}$. However, we know that in many cases of negation, the scope covers the whole sentence. For that reason, we must see this in relation to the fact that *negated* sentences in $\text{NoReC}_{\text{neg}}$ generally have a larger number of tokens, as shown in Table 5.5 on page 93.

There is a smaller proportion of discontinuous scopes in $\text{NorMed}_{\text{neg}}$, but the fraction of *true* discontinuous scopes among *all* discontinuous scopes in $\text{NorMed}_{\text{neg}}$ (28.3 %) is quite similar to the corresponding fraction in $\text{NoReC}_{\text{neg}}$ (30.7 %).

in 5.3.3, we identified certain patterns concerning the distribution of various cues. We are interested in knowing whether there are similar patterns for words that occur inside negation scopes. To analyze the content of scopes, we use the method described in 5.3.1 to extract the most frequent lemmas from the parts of the datasets annotated as negation scopes. Table 5.11 on page 98 illustrates the result.

For $\text{NorMed}_{\text{neg}}$, we recognize several of the lemmas in Table 5.11 from the list of the most frequent lemmas in the complete dataset (Table 5.4 on page 92). Among these are *pasient* 'patient', *behandling* 'treatment',

mann 'man', *kvinne* 'woman', *residiv* 'recurrence', *lege* 'doctor' and *kirurgi* 'surgery'. It is not surprising that frequent lemmas are also frequently found inside negation scopes. Therefore, we pay special attention to lemmas in Table 5.11 on the following page that are *not* among the most common ones in the dataset, i.e. their frequencies inside negation scopes are large relative to their frequencies in the dataset as a whole. Members of this group are the nouns *kontin* (incorrect lemmatization of *kontinens*) 'continence', *syndrom* 'syndrome', *effekt* 'effect' and *symptom* 'symptom'. There are also a few proper nouns: *Norge* 'Norway' and *Klinefelter* 'Klinefelter', which is the name of a diagnosis. The verbs *finnes* 'exist' and *foreligge* 'be present' belong to the same group, and so do the adjectives *sikker* 'certain' and *systematisk* 'systematic'.

As for $\text{NoReC}_{\text{neg}}$, the majority of the most frequent scope lemmas occur in Table 5.4 on page 92 as well. Several of these are common verbs, adverbs and adjectives, which are not specific to the review article domain. Others, such as *spill* 'game' and *historie* 'history, story' reveal to a larger extent information about the thematic content of the reviews. Eight lemmas occur relatively often inside scopes compared to their frequency rank in the corpus as a whole. Here, we find the noun *låt* 'tune', probably common in music reviews, as well as more generic lemmas: the adverbs *heller* 'rather, also', *lenge* 'long' and *alltid* 'always', the verbs *klare* 'manage', *vite* 'know' and *finne* 'find', and the adjective *mulig* 'possible'.

5.3.5 Summary and further discussion of the major differences between $\text{NorMed}_{\text{neg}}$ and $\text{NoReC}_{\text{neg}}$

Our study of the vocabularies of the two datasets shows that the most common lemmas to a large extent are domain-specific in $\text{NorMed}_{\text{neg}}$ compared to $\text{NoReC}_{\text{neg}}$. Knowing that the reviews in $\text{NoReC}_{\text{neg}}$ belong to eight quite different domains (Mæhlum et al., 2021) and thus are heterogenous with regard to their content, this is not unexpected. Concerning the observation that nouns are common among the most frequent lemmas in $\text{NorMed}_{\text{neg}}$, but not in $\text{NoReC}_{\text{neg}}$, we hypothesize that this is related to the previous statement about domain-specific words and heterogeneity in $\text{NoReC}_{\text{neg}}$; we believe that a larger proportion of nouns are domain-specific compared to other parts of speech such as verbs, and that a large proportion of domain-specific words belong to the noun class. An examination of the lemmatized corpora also shows that nouns are relatively more frequent in $\text{NorMed}_{\text{neg}}$ than in $\text{NoReC}_{\text{neg}}$; common nouns make up 48.9 % of lemmatized tokens after removal of punctuation, numerical-like words and stop words in $\text{NorMed}_{\text{neg}}$, while the corresponding proportion is only 35.9 % in $\text{NoReC}_{\text{neg}}$. This indicates a difference in writing style and is in accordance with our experience that descriptive texts written in a formal language tend to contain a large amount of nouns and few adjectives and verbs compared to texts belonging to more creative and subjective genres.

Negation is identified in just above one fifth of sentences in both datasets. We note that $\text{NorMed}_{\text{neg}}$ thus contains more negation than a comparable source such as the part of the BioScope Corpus consisting

NorMed$_{neg}$				NoReC$_{neg}$			
Lemma	**PoS**	**Trans.**	**Fq.**	**Lemma**	**PoS**	**Trans.**	**Fq.**
kontin*	NOUN	continence	33	hel	ADJ	whole	104
pasient	NOUN	patient	27	mye	ADJ	much	103
finnes	VERB	exist	15	heller	ADV	rather, also	85
sikker	ADJ	certain	15	god	ADJ	good	80
gi	VERB	give	13	få	VERB	get, receive	77
årsak	NOUN	cause	12	like	ADV	as	73
annen	DET	other	12	film	NOUN	movie	69
behandling	NOUN	treatment	12	gjøre	VERB	do	65
hos	ADP	at, by	10	stor	ADJ	big	62
vise	VERB	show	9	se	VERB	see	56
påvise	VERB	detect	9	ta	VERB	take	52
mann	NOUN	man	9	annen	DET	other	49
syndrom	NOUN	syndrome	9	komme	VERB	come	44
effekt	NOUN	effect	8	spill	NOUN	game	40
systematisk	ADJ	systematic	8	gå	VERB	walk, go	40
foreligge	VERB	be present	8	all	DET	all	38
norge	PROPN	Norway	8	klare	VERB	manage	33
god	ADJ	good	8	gi	VERB	give	32
klinefelter	PROPN	Klinefelter	8	mulig	ADJ	possible	30
kirurgi	NOUN	surgery	7	lenge	ADJ	long	29
symptom	NOUN	symptom	7	vite	VERB	know	29
stor	ADJ	big	7	ny	ADJ	new	28
residiv	NOUN	recurrence	7	år	NOUN	year	27
lege	NOUN	doctor	7	låt	NOUN	tune	27
passe	VERB	fit, match	7	the**	PROPN	the	27
finne	VERB	find	7	nok	ADV	presumably***	26
avhengig	ADJ	dependent	7	finne	VERB	find	25
forskjell	NOUN	difference	7	lang	ADJ	long	24
kvinne	NOUN	woman	7	historie	NOUN	history, story	24
spermie	NOUN	sperm cell	7	alltid	ADV	always	24
pankreatitt	NOUN	pancreatitis	7	gang	NOUN	time, corridor	24

Table 5.11: List of the most frequent lemmas inside scopes in NorMed$_{neg}$ and NoReC$_{neg}$, sorted by descending frequency. For each dataset, we report the Norwegian lemma, its English translation and the raw frequency ('Fq.'). *: *kontinens* 'continence' has been lemmatized as *kontin*, **: 'the' is English, ***: *nok* has more meanings, e.g. 'enough', 'yet'.

of biological papers, where 13.76 % of sentences are negated, and much more than the part consisting of *clinical* medical text (Vincze et al., 2008). According to Dalianis (2018), in Swedish clinical text, the amount of negated expressions and sentences has been quantified to 13.5 % (Dalianis and Skeppstedt, 2010). We note that the reported frequency of negation is affected by different definitions of negation in different corpora. In BioScope (Vincze et al., 2008), negation is interpreted as the non-existence of something.

Regarding negation cues, we find dissimilar distributions in the two datasets. *Ikke* 'not' is undoubtedly the most frequent cue in both, but proportionately less frequent in $\text{NorMed}_{\text{neg}}$. Furthermore, in comparison to $\text{NoReC}_{\text{neg}}$, $\text{NorMed}_{\text{neg}}$ contains a substantial proportion of affixal negation, a large number of which is found in affixes associated with words derived from Greek and Latin, which are known to be common in texts from the medical domain (Dalianis, 2018). The distribution of affixal cues in $\text{NoReC}_{\text{neg}}$ is much more homogenous, with *u-* 'un-/dis-/non-' and *-løs* '-free/-less' accounting for close to 96 %. Both $\text{NorMed}_{\text{neg}}$ and $\text{NoReC}_{\text{neg}}$ contain negation cues that are not present in the intersection of cues in the datasets. We focus on those unique to $\text{NorMed}_{\text{neg}}$ and find that there are several.

Negation scopes are generally shorter in $\text{NorMed}_{\text{neg}}$, and we believe this might be at least partly explained by the fact that negated sentences are shorter in this dataset than in $\text{NoReC}_{\text{neg}}$. Concerning discontinuous scopes, these are common in both datasets, but clearly more frequent in $\text{NoReC}_{\text{neg}}$. We also study the content of scopes and identify in both datasets lemmas that occur frequently inside scopes although not among the most frequent lemmas in the datasets as a whole.

Chapter 6

Negation resolution in the biomedical domain

This chapter covers the application of some of the models fine-tuned on $\text{NoReC}_{\text{neg}}$ (Mæhlum et al., 2021) in chapter 3 to the $\text{NorMed}_{\text{neg}}$ dataset annotated as part of chapter 5. We will report the results using the same standardized metrics as before. Additionally, a quantitative and qualitative error analysis of the best-performing model will be conducted. We will also attempt to view our models in a medical context by conducting an adjusted evaluation resembling the approach of Sadhukhan (2021), followed by a discussion on the applicability of the $\text{NoReC}_{\text{neg}}$ annotation scheme in a medical setting. Finally, we explore the effect of fine-tuning our models on $\text{NorMed}_{\text{neg}}$ and present preliminary modeling results.

6.1 Results

This section presents the results of performing cue detection and scope resolution on $\text{NorMed}_{\text{neg}}$. Three of the systems from chapter 3 are tested, each with fine-tuned embeddings from different language models; NB-BERT-large (Kummervold et al., 2021), NB-BERT-base (Kummervold et al., 2021) and NorBERT-2 (Kutuzov et al., 2021). As our test set we use $\text{NorMed}_{\text{neg}}$ in its entirety. Tables 6.1 to 6.3 on pages 102–103 present the results of evaluation against the original gold standard, the adjusted, word-level gold standard and the original gold standard after affix extraction from the predictions. We explain these various evaluation methods in chapter 3, subsection 3.1.4. For the extraction of affixes in the $\text{NorMed}_{\text{neg}}$ evaluation, we include affixal patterns from this dataset as well.[1]

The system with nb-bert-large as its language model overall achieves the best scores. This is in accordance with the results from chapter 3. We study the results in Table 6.3 on page 103 in light of the results from the evaluation on the $\text{NoReC}_{\text{neg}}$ test set in chapter 3, Table 3.17 on page 38. Unsurprisingly, the models perform notably poorer when applied to text

[1] https://github.com/marieef/master-thesis_code/blob/main/format_conversion/extract_affixes.py

	Original		
Lang.model	CUE	ST	FN
norbert2	68.42 (0.80)	77.24 (0.94)	47.04 (3.20)
nb-bert-base	67.44 (1.11)	78.58 (0.34)	49.66 (0.87)
nb-bert-large	**68.60** (0.76)	**79.22** (0.31)	**49.86** (0.73)

Table 6.1: Results of models with various language models when evaluated against the original $\text{NorMed}_{\text{neg}}$ gold standard. The metrics from the 2012 *SEM shared task are used. We report the average across 5 runs.

	Adjusted		
Lang.model	CUE	ST	FN
norbert2	80.44 (1.30)	80.28 (1.05)	56.21 (2.91)
nb-bert-base	81.89 (2.08)	82.26 (0.55)	62.07 (1.26)
nb-bert-large	**82.30** (1.10)	**82.72** (0.35)	**62.32** (0.79)

Table 6.2: Results of models with various language models when evaluated against the adjusted $\text{NorMed}_{\text{neg}}$ gold standard. The metrics from the 2012 *SEM shared task are used. We report the average across 5 runs.

from a domain different from the domain of the training data. For the nb-bert-large model, the CUE metric drops from 93.73 when evaluated on the $\text{NoReC}_{\text{neg}}$ test set, to 82.30 when evaluated on $\text{NorMed}_{\text{neg}}$. The corresponding numbers for the ST measure are 87.57 ($\text{NoReC}_{\text{neg}}$) and 82.72 ($\text{NorMed}_{\text{neg}}$). As for the FN scores, we observe a drop from 73.29 ($\text{NoReC}_{\text{neg}}$) to 62.32 ($\text{NorMed}_{\text{neg}}$).

We remember from chapter 5 that $\text{NorMed}_{\text{neg}}$ contains certain cues that are not present in $\text{NoReC}_{\text{neg}}$, and that we choose not to annotate some of the expressions interpreted as negation cues in $\text{NoReC}_{\text{neg}}$. This will not only influence the CUE score, but also the FN score, which requires both cue and scope to be correct. Thus, it seems reasonable that the $\text{NorMed}_{\text{neg}}$ evaluation leads to a large reduction for these metrics. The ST score, which mirrors the token-wise overlap between gold and predicted scopes, is less affected by the change of domain.

6.2 Error analysis

Within this part, we perform an error analysis on the best model from 6.1 on the preceding page. The different kinds of mistakes are quantified, and examples are provided.

6.2.1 Cue errors

Our system predicts 377 cues out of 510 gold standard cues correctly. The proportion of correctly predicted gold standard cues is thus 73.9 %. There

	Original+RE		
Lang.model	CUE	ST	FN
norbert2	80.36 (1.41)	80.28 (1.05)	56.21 (2.91)
nb-bert-base	81.89 (2.08)	82.26 (0.55)	62.07 (1.26)
nb-bert-large	**82.30** (1.10)	**82.72** (0.35)	**62.32** (0.79)

Table 6.3: Results of models with various language models when evaluated against the original $\text{NorMed}_{\text{neg}}$ gold standard after extracting affixes from the predictions. The metrics from the 2012 *SEM shared task are used. We report the average across 5 runs.

are 133 false negatives, corresponding to 26.1 %. The number of false positive cues is 30 (30.2, 7.4 % of all predicted cues).

False negatives

Table 6.4 on the next page provides an overview of the frequency by which the various negation cues are missed by the model. Only cues occurring more than twice are included.

In general, cues that do not occur in $\text{NoReC}_{\text{neg}}$ are false negative in all cases, such as *utelukke* 'exclude' and *fjerning* 'removal'. However, a few occurrences of *a-* 'a-' and *ab-* 'ab-' are actually correctly recognized as cues although absent in the training data. Affixal cues such as *u-* 'un-/dis-/non-' and *ikke(-)* 'non(-)', and in particular *-fri* '-free/-less' are relatively often not detected. We notice that the model struggles with the word *progredieringsfri* 'progression-free'. Other words found among these false negatives are *uendret* 'unchanged', *uavhengig* 'independent', *unormal* 'abnormal' and *ulikt* 'dissimilar'.

False positives

In Table 6.5 on page 105, we list all false positive cues and report their frequency in the pool of false positives, as well as the false discovery rate.

U- 'un-/dis-/non-' stand for two thirds of the falsely predicted cues. The majority of these errors can be attributed to the word *ulike* 'dissimilar, various, several' in cases where it is used in the meaning 'various, several'. False positives in nominalizations of affixally negated adjectives occur as well (*ufølsomhet* 'insensitivity'). In Table 6.5 on page 105, we observe the same type of errors in nominalizations of adjectives ending with *fri* '-free/-less' (*smertefrihet* 'freedom of pain' and *symptomfrihet* 'freedom of symptoms').

An interesting observation is the prediction of *suboptimal(e)* 'suboptimal' as a word containing a negation cue. *Sub-* 'sub-' does not occur as a cue in $\text{NoReC}_{\text{neg}}$ and has not been considered as such by us, yet in this exact word, one could argue that it leads to the reading 'not optimal'.

Cue	Trans.	Total	FN (%)
ikke	not	178	0.0
u-	un-/dis-/non-	82	21.0
in-	in-	48	92.9
uten	without	46	0.0
ingen	none/nobody	36	0.0
ikke-	non-	19	14.7
a-	a-	11	94.5
utelukke	exclude	8	100.0
-fri	-free/-less	6	53.3
manglende	lacking	5	0.0
aldri	never	5	8.0
fjerning	removal	4	100.0
-løse	-less	4	5.0
eksklusjon	exclusion	3	100.0
fjernet	removed	3	100.0
ab-	ab-	3	93.3
mangler	lacks	3	26.7
fjernes	is.removed	3	100.0
ikke	non- (affixal)	3	20.0
nei	no (interj.)	3	20.0
fjerne	remove	3	100.0
an-	an-	3	100.0

Table 6.4: For all cues with > 2 occurrences in the gold standard, we report the number of occurrences as cues in $\text{NorMed}_{\text{neg}}$ and % of false negatives (FN). Values are averaged across 5 runs corresponding to 5 instances of the model trained with different seeds. Words have been lowercased before counting.

FP cue	Trans.	% of all FP	FDR
u-	un-/dis-/non-	67.5	23.9
slipper	does not have to	4.0	100.0
ikke	not	3.3	0.6
ingen	none/nobody	3.3	2.7
mangler	lacks	3.3	31.3
nei	no (interj.)	3.3	29.4
smertefrihet	freedom of pain	2.0	100.0
verken eller	neither nor	2.0	42.9
unntaket	the exception	2.0	100.0
slippe	not have to	1.3	100.0
suboptimal	suboptimal	1.3	100.0
suboptimale	suboptimal	1.3	100.0
im-	im-	0.7	33.3
in-	in-	0.7	5.6
frie	free	0.7	100.0
mangelfull	deficient	0.7	100.0
negativt	negative	0.7	100.0
symptomfrihet	freedom of symptoms	0.7	100.0
verken	neither	0.7	100.0
cancersuspekt	suspicious for cancer	0.7	100.0

Table 6.5: Distribution of false positive cues, in terms of percentage of all false positives and False Discovery Rate (FDR). Numbers are averaged across 5 runs corresponding to 5 instances of the model trained with different seeds. Words have been lowercased before counting.

Among other false positive cues, we find *mangler* 'lacks', which does not always indicate actual negation. Another case is *slippe(r)* '(does) not have to', which we do not consider negation due to inseparability from modality.

Unfortunately, a small number of cue occurrences seem to have been missed in the annotation process. This applies to one occurrence of each of the cues *ikke* 'not', *nei* 'no' and *unntaket* 'the exception'. The correct number of false positives for these cues would be 0.

6.2.2 Scope errors

We consider the cases with correctly predicted cues and find that 64.2 % of scope predictions are correct. The average scope length of the dataset is 5.6 tokens (including implicit scopes). Predicted scopes are somewhat longer (6.35 tokens). Table 6.6 on the following page provides a quantitative overview of scope errors by cue. We notice that approximately one third of scope predictions are wrong for the frequent cues *ikke* 'not' and *u-* 'un-/dis-/non-'. In the case of *ab-* 'ab-', scope is never correctly resolved, but note that this number is based on one case only. Other cues with high error

Cue	Trans.	% incorrect scopes
ikke	not	35.2
u-	un-/dis-/non-	31.5
in-	in-	25.7
uten	without	42.6
ingen	none/nobody	29.4
ikke-	non-	23.4
a-	a-	0.0
-fri	-free/-less	13.3
manglende	lacking	24.0
aldri	never	18.0
-løse	-less	73.3
ab-	ab-	100.0
mangler	lacks	53.3
ikke	non- (affixal)	23.3
nei	no (interj.)	83.3

Table 6.6: Scope predictions for true positive cues with > 2 occurrences in the gold standard. Unless there is a perfect match with the gold scope, it is counted as an error. The raw frequency for each cue can be derived from Table 6.4 on page 104. Numbers are averaged across 5 runs corresponding to 5 instances of the model trained with different seeds. Words have been lowercased before counting.

percentages are the following: *nei* 'no', *-løse* '-less/-free', *mangler* 'lacks' and *uten* 'without'.

Too long predicted scopes

As mentioned, the scopes produced by the model are in general longer than the gold standard scopes. In the following, we include some examples of sentences where additional words are falsely included inside predicted scopes.

Examples (6.1) and (6.2) illustrate two problems well-known from chapter 3. In (6.1), the relative subjunction *som* 'that, which' should have been left outside the scope. The same applies to the expletive subject *det* 'there' in (6.2).

(6.1) Gold: *... si ifra til sine overordnede om [faglige*
Pred: *... si ifra til sine overordnede om [faglige*
... say from to their superiors about professional
forhold] som ***ikke*** *[er tilfredsstillende] .*
forhold som] ***ikke*** *[er tilfredsstillende] .*
relationships that not are satisfactory .
'...make their superiors aware of unsatisfactory professional conditions.'

(6.2) Gold: *[I mange utviklingsland finnes] det ofte*
Pred: *[I mange utviklingsland finnes det] ofte*
In many development.countries exist there often
ikke *[noe slikt som prehospital akuttmedisin]* .
ikke *[noe slikt som prehospital akuttmedisin]* .
not something such as pre-hospital acute.medicine .
'In many developing countries, there is often no pre-hospital emergency medicine.'

Challenging cases are also found for affixal negation. In the following, this is exemplified by three sentences. (6.3) contains an affixally negated adjective used as an adverb, and the gold standard scope is delimited to the adjective alone. In spite of this, the system predicts the following adjective and the noun phrase head as part of the scope as well. As for example (6.4), the scope of *u-* 'un-/dis-/non-' is resolved to be the whole sentence, when in reality, it should be the noun phrase modified by the negated adjective. In (6.5), a noun phrase postmodifier is erroneously predicted as part of the noun phrase scope.

(6.3) Gold: *... behandling av* ***u****[frivillig] barnløse par der*
Pred: *... behandling av* ***u****[frivillig barnløse par] der*
... treatment of involuntarily childless couples there
mannen har Klinefelters syndrom .
mannen har Klinefelters syndrom .
man.the has Klinefelter's syndrome .
'...treatment of involuntarily childless couples where the male has Klinefelter's syndrome.'

(6.4) Gold: *Søreide viser at denne prøven har* ***u****[sikker*
Pred: *Søreide viser at [denne prøven har]* ***u****[sikker*
Søreide shows that this test.the has uncertain
diagnostisk treffsikkerhet] .
diagnostisk treffsikkerhet] .
diagnostic accuracy .
'Søreide shows that the diagnostic accuracy of this test is uncertain.'

(6.5) Gold: *Artikkelen er basert på ... et* ***ikke-****[systematisk*
Pred: *Artikkelen er basert på ... et* ***ikke-****[systematisk*
Article.the is based on ... a non-systematic
litteratursøk] i PubMed .
litteratursøk i PubMed] .
literature.search in PubMed .
'The article is based on ... a non-systematic literature search in PubMed.'

Too short predicted scopes

There are also cases where parts of gold standard scopes are outside the predicted scopes. One such case is (6.6), where the noun phrase head

should be part of the scope of *uten* 'without'. Another example is (6.7); here, the subject is correctly included in the scope, but so should the verb be. Last, there is (6.8), where the restrictive relative clause is not predicted as part of scope by the system.

(6.6) Gold: *Når [residiv] **uten** [tegn til fjernmetastaser]*
Pred: *Når residiv **uten** [tegn til fjernmetastaser]*
When recurrence without signs to distant.metastases
ble diagnostisert , henviste 23 av 25 lokalsykehus .
ble diagnostisert , henviste 23 av 25 lokalsykehus .
became diagnosed , referred 23 of 25 local.hospitals .
'When recurrence without signs of distant metastases were diagnosed, 23 out of 25 local hospitals made a referral.'

(6.7) Gold: *[Coloncancer ble ansett som] **u**[sannsynlig] ...*
Pred: *[Coloncancer] ble ansett som **u**[sannsynlig] ...*
Colon.cancer became regarded as unlikely ...
'Colon cancer was considered unlikely...'

(6.8) Gold: *... behandling av ufrivillig [barn]**løse** [par der*
Pred: *... behandling av [ufrivillig barn]**løse** [par] der*
... treatment of involuntarily childless couples there
mannen har Klinefelters syndrom] .
mannen har Klinefelters syndrom .
man.the has Klinefelter's syndrome .
'...treatment of involuntarily childless couples where the male has Klinefelter's syndrome.'

6.3 Our models in a clinical context

One of the aims of this thesis is to assess the portability of negation models trained on review articles into the medical domain. This was partly addressed through the application of the models to $\text{NorMed}_{\text{neg}}$ in 6.1 and the associated error analysis in 6.2. However, the question of whether these models are suitable for use in an actual clinical or medical setting remains unanswered and will be approached in this section.

6.3.1 An adjusted comparison to Norwegian NegEx

Here, we make an attempted quantitative comparison of our models to the system created by Sadhukhan (2021). We would like to stress that this is only an approximation, i.e. the results we present cannot be directly compared to theirs.

Assumptions and adjustments

A series of assumptions and adjustments to the annotation scheme of Sadhukhan (2021) were made in this process. These are accounted for in the following.

Definition of negation The NegEx model works by deciding for each occurrence of a term from a predefined list of terms whether the given term occurrence is negated or not (Chapman et al., 2001; Sadhukhan, 2021). As discussed in chapter 5, the negation scopes our models have been trained to identify are fundamentally different. For the predictions of our models, we therefore define a term occurrence as negated if the entire term is contained in a predicted scope. If it is not predicted as part of a scope, we say that it is not predicted as negated.

Gold standard Concerning the gold standard, we identify two main options: Either, the original annotations of the corpus (Sadhukhan, 2021) can be used, or our annotations, building on Mæhlum et al. (2021) and described in chapter 5. The choice of gold standard depends on the aim of the comparison.

The use of the original annotations would be rather unfair to our models, which are trained on a dataset using an annotation scheme very different from the original one. This would probably generate a large number of false positives in cases where predefined clinical terms are not annotated by Sadhukhan (2021), but correctly predicted as part of scope according to the $\text{NoReC}_{\text{neg}}$ scheme.

By using our own annotations we take the properties of our models into account. Although the gold standard is different from Sadhukhan (2021), this still allows us to quantify the ability of our models, given their prerequisites, to correctly recognize clinical terms as negated or not negated, i.e. inside or outside negation scopes. We decide to follow this approach.

List of predefined terms Regarding the use of predefined terms, as far as we understand, Sadhukhan (2021) obtain these from two sources:[2] The first is a list of clinical terms related to medical conditions extracted from NorMedTerm (Pilan et al., 2020), referred to as NorMedTermCondition, and the second is a *custom* list of terms, originating from the terms annotated as negated by Sadhukhan (2021), as well as related terms found through the use of word2vec (Mikolov et al., 2013a; Mikolov et al., 2013b) on the dataset. They also mention the use of terms from the Norwegian ICD-10[3] diagnosis code system, but they neither provide a list of these nor use any other sources of terms than the two previously mentioned in their published code. Also, it is our impression that the NorMedTermCondition actually does contain terms obtained from ICD-10.

During testing, we understand that Sadhukhan (2021) use *all* terms from NorMedTermCondition. As for the list of *custom* terms, however, they seem to only use those that occur in the *development* set. In other words, their model will not be able to detect gold standard negations of *custom*

[2]The files named NorMedTermCondition.txt and myWords.txt in this directory: https://github.com/DebaratiSJ/NegEx-on-Norwegian-biomedical-text/tree/main/Associated%20files

[3]Available here: https://finnkode.ehelse.no/#icd10/0/0/0/-1

terms that are only present in the test set, i.e. these will automatically become false negatives unless they occur in NorMedTermCondition as well. We follow the same approach in our evaluation to make the conditions as similar as possible.

Greedy approach Altogether, the number of terms in the custom list and NorMedTermCondition is approximately 24,000. We notice that a large number of shorter terms occur inside other, longer terms and decide to only count the longest matching term in each case, i.e. a greedy approach.

Count each term occurrence once Furthermore, in the $\text{NoReC}_{\text{neg}}$ annotation scheme, it is the case that a word or group of words can occur inside the negation scope of multiple negation cues within the same sentence. We choose the approach that a negation is registered if a term occurs inside any scope in a sentence, but it is only counted once even if it occurs in multiple scopes.

Counting terms corresponding to word boundaries We decide to only count predefined terms that correspond to word boundaries. *Operert* 'operated (upon)' is an example of a term found in the predefined term list and will be registered as negated if it occurs inside a scope. *Uoperert* 'not operated (upon)' represents an affixal negation of this term, but in such a case, *operert* 'operated (upon)' will not be registered as negated. As far as we can see, this corresponds to the practice of Sadhukhan (2021). Their list of negation triggers does not include negation affixes, but contains the affixed word *usannsynlig* 'unlikely' as an independent trigger, enabling recognition of *coloncancer* 'colon cancer' as negated in *Coloncancer ble ansett som usannsynlig* 'Colon cancer was considered unlikely'. With our approach, this term occurrence would also be regarded as negated, since it is inside the scope of *u-* 'un-/dis-/non-' in the sentence.

By only counting terms that correspond to word boundaries, we also avoid the problem of recognizing strings that overlap with a term without actually representing the term, as could be the case with shorter terms such as *dø* 'die', *MS* 'MS, multiple sclerosis' and *ør* 'lightheaded, dizzy'.

Evaluation metric Sadhukhan (2021) use F-score as their evaluation metric, and since nothing else is stated, we assume this must be a balanced F-score, i.e. F_1. Given that their model recognizes a set of predefined, possibly multi-word terms (Chapman et al., 2001; Sadhukhan, 2021), we suppose their F-score is computed on the term-level, as opposed to the token-level, i.e. either the complete term is matched, or there is no match at all.

Data split The data split used for development and testing in Sadhukhan (2021) has not been provided. However, through personal correspondence with the authors, we have been informed that they used the last 30 % of the sentences for testing. We use the same data split in this comparison.

Lang.model	P	R	F_1
norbert2	0.89 (0.03)	0.54 (0.03)	0.67 (0.02)
nb-bert-base	0.92 (0.04)	0.60 (0.03)	**0.72** (0.03)
nb-bert-large	0.92 (0.02)	0.57 (0.03)	0.70 (0.02)

Table 6.7: Results of systems with various language models when evaluated with respect to predefined terms contained inside our annotated and predicted scopes. We use the same sentences for testing as Sadhukhan (2021) did. The evaluation metric is term-level F_1-score (F_1). Precision (P) and recall (R) are also reported. All values are averaged across 5 runs.

With a greedy approach where each negated term is counted only once, and only terms corresponding to word boundaries are considered, there is a total of 454 term occurrences in the test set. 50 of these are negated according to the gold standard, whereas the remaining 404 are not. Note that due to their stricter definition of negation, Sadhukhan (2021) report only 11 cases of gold standard negation in the test set. When we count their annotations, we find that the actual number is 10.

Results

We adhere to the assumptions described in 6.3.1 and apply the three models from section 6.1 to the test set. The evaluation method is the one known as 'Original+RE' from previous tables, i.e. we perform negation affix extraction on the cases predicted as affixal negation before evaluation is performed.

The results are presented in Table 6.7. In this case, the best-performing system is the one with nb-bert-base as its language model. This model achieves a term-level F_1 score of 0.72, while the nb-bert-large model, which outperformed the others according to the *SEM 2012 (Morante and Blanco, 2012) evaluation in section 6.1, has a score of 0.70. The system with fine-tuned norbert2 embeddings shows the poorest performance with an F_1 score of 0.67. We remember that Sadhukhan (2021) achieved a score of 0.55. However, our approach is simplified and different from theirs; although a predefined term occurs inside one of our scopes, it does not mean that it would be considered a pertinent negative (Chapman et al., 2001; Sadhukhan, 2021). This means that we have more occurrences of gold standard negated terms in our version of the dataset. Another observation is the low scores with respect to recall seen in Table 6.7. These are necessarily low, since several of the terms annotated as negated by Sadhukhan (2021) occur exclusively in the test set and thus are not available for recognition. From the annotation process, we know that all the terms annotated as negated by them are also annotated inside negation scopes by us. In other words, the terms missed by their model will also be missed by our models according to this evaluation practice.

6.3.2 Detecting pertinent negatives with the NoReC$_{neg}$ annotation scheme

In a clinical setting, one is interested in detecting pertinent negatives, i.e. terms related to symptoms, findings and diseases that are described as *absent* (Chapman et al., 2001; Sadhukhan, 2021). Above, we made the simplification that an occurrence of a term inside a negation scope means a negation of the term. The remaining question is whether the annotation scheme developed for NoReC$_{neg}$ by Mæhlum et al. (2021) is useful in negation modeling for the medical or clinical domain: Are the terms identified inside the scopes actually pertinent negatives, and how can we determine this?

In order to answer these questions, we perform a qualitative analysis of the 50 cases of negation identified in 6.3.1. We find that in the majority of these cases, the terms recognized inside our negation scopes cannot be considered as 'negated' or 'absent'. (6.9) and (6.10) illustrate this. Predefined terms are underlined in the examples. Bold typeface is used for cues and square brackets for scopes, as before.

(6.9) *Det [finnes]* ***ikke*** *[data på hvor mange av disse tilfellene*
There exist not data on where many of these cases
som ble utløst av hypertriglyseridemi] .
that became triggered by hypertriglyceridemia .
'There are no data on the number of cases triggered by hypertriglyceridemia.'

(6.10) *Tumorekspresjon av ... er assosiert med økt*
Tumor.expression of ... is associated with increased
*[progredierings]**fri** [overlevelse] ved kolorektal kreft* .
progression.free survival by colorectal cancer .
'Tumor expression of ... is associated with prolonged progression-free survival in colorectal cancer.'

There are also several cases where an identified term *can* be regarded as a pertinent negative. All negations annotated by Sadhukhan (2021) are inside our annotated scopes. This includes two cases we would not count as pertinent negation. Furthermore, three term occurrences that could be regarded as pertinent negatives are contained in our scopes, but unannotated by Sadhukhan (2021). Among these are (6.13) and (6.14), which are discussed in the following paragraph.

In the sentences (6.11) - (6.14), we find examples of terms covered by the NorMed$_{neg}$ annotations that we count as real pertinent negatives. We see that in (6.14), there is an exact overlap between the annotated scope and the predefined term. However, in (6.11) - (6.13), one would need to apply a post-processing strategy to the scopes in order to extract the pertinent negative. (6.11) and (6.13) have in common that the identified term is the head of the object noun phrase. In (6.12), the negated term is the object of the prepositional phrase functioning as a complement of the object noun phrase head.

(6.11) *... [vi observerte]* ***ingen*** *[komplikasjoner hos mor] .*
... we observed no complications at mother .
'...we observed no complications in the mother.'

(6.12) *[Menn viste]* ***ingen*** *[tegn til humørforandring , målt*
Men showed no signs to mood.change , measured
ved POMS] , ...
by POMS , ...
'The men showed no signs of change in mood, measured in terms of POMS,...'

(6.13) *[I fase 2- og 3-studier ved metastatisk sykdom har*
In phase 2- and 3-studies at metastatic illness has
monoterapi med disse preparatene] ***ikke*** *[vist sikker*
monotherapy with these medications not shown certain
objektiv respons] ...
objective response ...
'In phase 2 and 3 studies of metastatic illness, monotherapy with these medications have not shown certain objective response...'

(6.14) *... fant vi økt nivå av survivin og telomerase i*
... found we increased level of survivin and telomerase in
fjernede *[polypper] hos pasienter som senere utviklet*
removed polyps at patients who later developed
kolorektal kreft ...
colorectal cancer ...
'...we found raised levels of survivin and telomerase in polyps removed from patients who later developed colorectal cancer...'

In another group of cases, the identified term is the focus of a negation we consider non-factual. We observe an exact overlap between the annotated scope and the identified term in (6.15). In (6.16), on the other hand, the scope includes a prepositional phrase modifying the noun recognized as a term.

(6.15) *Selv om kirurgisk* ***fjerning*** *av [levermetastaser] er mulig*
Self about surgical removal of liver.metastases is possible
hos en rekke pasienter , ...
at a row patients , ...
'Even if surgical removal of liver metastases is possible in many patients,...'

(6.16) *Diagnosen kan stilles ved* ***manglende***
Diagnosis.the can be.made at lacking
[lipoproteinlipaseaktivitet etter en intravenøs dose med
lipoprotein.lipase.activity after an intravenous dose with
heparin]
heparin
'The diagnosis can be made in case of lack of lipoprotein lipase activity after a dose of intravenous heparin'

Based on the studied examples, we would suggest the application of a syntactic analysis to sentences with term occurrences inside negation scopes. Different cases would need to be considered, some of which we mention below.

When the cue is a lexical negation verb, such as *mangle* 'lack', *fjerne* 'remove' or *utelukke* 'exclude', it makes sense to look at the arguments of the verb. Often, the direct object would represent what is absent, but this depends on the verb, and whether the sentence is in active or passive voice. Inside the relevant verb argument, one could identify the head and apply some heuristic to check whether the head matches or overlaps with an identified term.

When it comes to negation adverbs like *ikke* 'not' and *aldri* 'never', which lead to the negation of the main verb, one opportunity is to use a list of verbs that are likely to occur in descriptions of symptoms, findings and diseases in a patient, e.g. 'be', 'have', 'find' etc. If there is a verb match, one could do as suggested for negative verbs above. In this context, we note that the properties of $\text{NorMed}_{\text{neg}}$ scopes often enable the extraction of a group of words describing more precisely what is absent than the predefined term alone. (6.13) illustrates this point; here, *sikker objektiv respons* 'certain objective response' is more accurate than the term *respons* 'response' itself.

With nominal cues, like *fjerning* 'removal' and *eksklusjon* 'exclusion', a possible heuristic is to check if the cue has a prepositional phrase complement with the preposition *av* 'of', and if so, consider the prepositional object as negated if it matches a relevant term.

Affixal negation would also need to be handled. One simple approach would be to check for a term match in the remaining part of the affixally negated word, e.g. if *symptom* 'symptom' is a term and the suffix *-fri* '-free' indicates negation, *symptom* 'symptom' would be considered negated in the word *symptomfri* 'symptom free'.

A general challenge would be the separation of factual negation from non-factual negation and cases of modality. To exclude cases that do not represent pertinent negation, one could make an attempt at detecting double negations, modal verbs in relation to the main verb and constructs with subordinate conjunctions (*hvis* 'if', *dersom* 'if'). However, there are also other ways to express negations that are not factual, as seen in (6.15) and (6.16). We think that a better option might be to reannotate the dataset with factuality information and train models to perform the task of determining whether negations are factual or not.

Above, we have provided suggestions of how pertinent negatives could be extracted from scopes annotated according to NoReC$_{neg}$. This is by no means an exhaustive list of all possible cases, nor a solution that would provide a perfect system for resolving negation in clinical and medical text.

6.4 Fine-tuning on a subset of NorMed$_{neg}$

In section 6.1, we observed a drop in evaluation scores when porting the models trained on NoReC$_{neg}$ review articles into the medical domain. The natural follow-up question is whether access to domain-specific annotated data can help improve cue detection and scope resolution in biomedical text. We thus conduct one round of fine-tuning of the best-performing model from section 6.1 on NorMed$_{neg}$. The training settings and parameters are kept as before. The only exception is that input token sequences had to be truncated to a length of 128 during training to avoid excessive memory usage.

As the held-out test set, we use the sentences in NorMed$_{neg}$ corresponding to the test set used by Sadhukhan (2021) (643 sentences). The remaining part of NorMed$_{neg}$ (1443 sentences) is split into train (85 %) and development test (15 %). To get an impression of the amount of annotated data needed for fine-tuning to have an effect, we divide the training set into portions of different size. We conduct our experiments with three different sizes of the training set: 10, 40 and 100 %. This corresponds to 123, 491 and 1227 sentences, respectively, and in each case, 21.5 % can be expected to be negated sentences.

The results are presented in Tables 6.8 to 6.10 on the next page. 'M' represents the model from section 6.1, fine-tuned on NoReC$_{neg}$ only. The other models are denoted by a subscript referring to the size of the NorMed$_{neg}$ training set used in further fine-tuning. Especially when studying the 'Adjusted' and 'Original+RE' results, we see that even fine-tuning on a small amount of sentences (10 %) has an effect on all metrics. The effect increases with the size of the training set. When using the complete training set for fine-tuning, we achieve scores of 91.87 (CUE), 88.66 (ST) and 74.07 (FN). This is a clear improvement compared to the results from Table 6.3 on page 103.

	Original		
	CUE	ST	FN
M	73.28 (0.91)	80.44 (0.69)	47.27 (1.99)
M_{10}	74.01 (1.73)	80.41 (0.69)	51.82 (1.87)
M_{40}	74.02 (1.38)	81.49 (0.54)	54.06 (1.37)
M_{100}	**78.16** (0.81)	**85.35** (1.33)	**60.48** (1.30)

Table 6.8: Results of models after further fine-tuning on $\text{NorMed}_{\text{neg}}$ with different training set sizes. The models are evaluated against the original gold standard of the held-out test set. The metrics from the 2012 *SEM shared task are used. We report the average across 5 runs.

	Adjusted		
	CUE	ST	FN
M	84.19 (0.84)	82.94 (0.65)	58.57 (2.19)
M_{10}	86.11 (1.53)	83.23 (0.66)	64.54 (2.18)
M_{40}	88.42 (1.44)	84.86 (0.57)	68.21 (1.60)
M_{100}	**91.87** (0.46)	**88.66** (1.47)	**74.07** (1.74)

Table 6.9: Results of models after further fine-tuning on $\text{NorMed}_{\text{neg}}$ with different training set sizes. The models are evaluated against the adjusted gold standard of the held-out test set. The metrics from the 2012 *SEM shared task are used. We report the average across 5 runs.

	Original+RE		
	CUE	ST	FN
M	84.19 (0.84)	82.94 (0.65)	58.57 (2.19)
M_{10}	86.11 (1.53)	83.23 (0.66)	64.54 (2.18)
M_{40}	88.42 (1.44)	84.86 (0.57)	68.21 (1.60)
M_{100}	**91.87** (0.46)	**88.66** (1.47)	**74.07** (1.74)

Table 6.10: Results of models after further fine-tuning on $\text{NorMed}_{\text{neg}}$ with different training set sizes. The models are evaluated against the original gold standard of the held-out test set after extraction of negation affixes from the predictions. The metrics from the 2012 *SEM shared task are used. We report the average across 5 runs.

Chapter 7

Conclusion

This chapter summarizes the work conducted as part of this thesis. Through a presentation of our findings, we attempt to answer the research questions defined in chapter 1. Furthermore, we share our perspective on the contributions of this research, and finally, we provide suggestions for future work.

7.1 Summary

This part is a summary of all preceding chapters except the introduction and background. We describe our main findings and how these contribute to answering our research questions, which we repeat below:

RQ1: Can we achieve state-of-the-art results for negation resolution in Norwegian with a neural sequence labeling system?

RQ2: How applicable are the $\text{NoReC}_{\text{neg}}$ resources to new projects and new domains?

RQ3: Can a negation resolution system fine-tuned on review articles be ported into the medical domain without a loss of performance?

RQ4: How are the results in the medical domain affected by further fine-tuning of the aforementioned system on medical text?

Chapter 3

In chapter 3, we conducted experiments with a neural sequence-labeling system for negation resolution in Norwegian text based on NegBERT (Khandelwal and Sawant, 2020). The system consists of a separate cue detection and scope resolution model, each utilizing transformer-based embeddings with a classification layer on top. All models were fine-tuned and evaluated on the Norwegian review dataset $\text{NoReC}_{\text{neg}}$ (Mæhlum et al., 2021), and two rounds of experiments were performed. In the first round, we experimented with variations of the scope resolution model and found that it benefits from training on both negated and non-negated sentences

as opposed to only negated ones, and from receiving information on the word form of the negation cue. The second round consisted of language model experiments, and our findings indicate that the language model NB-BERT-large (Kummervold et al., 2021) leads to the best negation resolution performance. Using the evaluation metrics from the *SEM 2012 Shared Task (Morante and Blanco, 2012), it scores 93.73 for cue-level F_1, 87.57 for scope tokens F_1 and 73.29 for full negation F_1 on the held-out test set. We compare these results to Mæhlum et al. (2021) and establish that the answer to **RQ1** is yes; state-of-the-art results for Norwegian negation resolution can be achieved using a sequence-labeling approach.

Chapter 4

Chapter 4 is a review of the $\text{NoReC}_{\text{neg}}$ dataset with respect to annotation guidelines and practice. We found multiple cases of inconsistent annotations that we attributed to underspecified guidelines, and we identified several cases where annotation practice deviates from the guidelines. This partly answers **RQ2**: With future projects in mind, the ambiguity of parts of the guidelines is not ideal. Inconsistencies and annotation errors indicate that there is some room for improvement in the dataset as well.

Chapter 5

In chapter 5, we described the process of reannotating *The Norwegian GastroSurgery Biomedical Negation Corpus* (Sadhukhan, 2021) based on a slightly modified and extended version of the $\text{NoReC}_{\text{neg}}$ guidelines. We introduced this dataset as $\text{NorMed}_{\text{neg}}$. A high degree of inter-annotator agreement was measured for the annotation task. We found that the vocabulary of $\text{NorMed}_{\text{neg}}$ is highly influenced by words belonging to the medical and clinical domains, and that the distribution of negation in $\text{NorMed}_{\text{neg}}$ is characterized by a large proportion of affixal negation, including many negations triggered by affixes derived from Latin and Greek, which are not present in $\text{NoReC}_{\text{neg}}$. In the annotation phase, these patterns triggered an extension of the original guidelines with a few assumptions covering cases observed in the biomedical data. Specifically, this was related to the mentioned Graeco-Latin negation affixes and to affixal negation in nominalizations. This provides another part of the answer to **RQ2**; the $\text{NoReC}_{\text{neg}}$ guidelines are applicable to the domain of biomedical articles, but a few domain-related refinements are necessary. Most of our additional assumptions, however, were not triggered by the change in domain, but rather by the originally insufficient specification of certain guidelines.

Chapter 6

Chapter 6 documents the transfer of models trained on review articles ($\text{NoReC}_{\text{neg}}$) into the medical domain ($\text{NorMed}_{\text{neg}}$). Our results show that the system performs notably poorer on $\text{NorMed}_{\text{neg}}$ than on data from the

training domain. As expected, it fails to detect a large number of negation patterns in NorMed$_{neg}$ that are rare or non-existing in NoReC$_{neg}$. This provides an answer to **RQ3**: No, porting the system fine-tuned on review articles directly into the medical domain leads to a drop in performance.

Through further fine-tuning of the models on approximately 1200 sentences from NorMed$_{neg}$, we answered **RQ4**: An additional round of fine-tuning on target-domain data improved performance on the NorMed$_{neg}$ test set drastically. A positive effect on performance was seen with smaller amounts of data as well. From this we infer that an improved performance can be obtained in the medical domain with limited annotation effort.

Another part of this chapter dealt with the role of the NoReC$_{neg}$ annotation scheme in a clinical or medical context. We elaborated on some of the differences between this annotation scheme and annotations specifically targeting clinical text. Scopes annotated according to NoReC$_{neg}$ contain many terms that are not *actually* described as absent, but also a few cases of these *pertinent negatives* (Chapman et al., 2001) that were not caught by the clinically oriented annotation by Sadhukhan (2021). This touches upon **RQ2** again; if the goal is to recognize pertinent negatives only, the NoReC$_{neg}$ annotation scheme cannot be used directly without some form of post-processing. We proposed syntactic analysis and annotation of negation factuality as methods to adapt it to a clinical setting.

7.2 Contributions

Among the contributions of this thesis is the first neural sequence-labeling system for negation resolution in Norwegian, trained on the NoReC$_{neg}$ dataset (Mæhlum et al., 2021). We also hope that our review of the NoReC$_{neg}$ annotation guidelines will offer some insightful observations to future efforts to annotate negation for Norwegian and possibly other languages. Furthermore, we present the NorMed$_{neg}$ dataset, a Norwegian biomedical dataset with negation cue and scope annotations. This may be a useful resource in other research targeting negation in medical text. Our findings regarding negation types and distribution in biomedical research literature may also be of use in such projects. For our mentioned negation resolution system, we provide results from evaluation on NoReC$_{neg}$ and NorMed$_{neg}$, thus measuring the impact on model performance of a change in domain. We also regard as an interesting contribution our tentative modeling results, which indicate that an additional round of fine-tuning on biomedical text leads to improved negation resolution in this domain, with the size of the effect depending on the amount of data. The code and data splits used to create our models can be accessed on GitHub.[1] NorMed$_{neg}$ is available in a separate repository.[2]

[1] https://github.com/marieef/master-thesis_code
[2] https://github.com/marieef/NorMed_neg/

7.3 Future work

Due to limited time, we were not able to explore all parts of this thesis with an equally high level of detail. This applies to our negation modeling experiments. We leave the opportunity to conduct more extensive studies on the optimization of hyperparameter values and training settings for future research. We consider it likely that this can lead to improved results.

A review of our lists of patterns used for the extraction of affixes in evaluation could also be performed. As far as we can see, it is not clear from Mæhlum et al. (2021) how this was handled by them.

Furthermore, we identify some possible next steps for research in the field of negation resolution in Norwegian. One such step could be to train new instances of existing systems like ours whenever a new Norwegian or multilingual language model is made available. The rapid development of new language models pre-trained on increasing amounts of data should be taken advantage of in negation modeling. Even at the time of finishing this thesis, new language models from the NorLM initiative responsible for the NorBERT models have been made available (Samuel et al., 2023). In future work, these should be tested on negation resolution, on $\text{NoReC}_{\text{neg}}$ (Mæhlum et al., 2021) as well as on the new $\text{NorMed}_{\text{neg}}$ dataset.

There is also potential in further experiments with more advanced model architectures, such as graph-based approaches (Kurtz et al., 2020; Mæhlum et al., 2021). We imagine that such methods could be used in combination with contextualized embeddings from transformer-based language models. In addition, recent results indicate that negation detection can benefit from pre-training language models on augmented data and with a negation-masking objective (Truong et al., 2022).

As for the data, based on our review of the $\text{NoReC}_{\text{neg}}$ annotation guidelines and practice, we would suggest a revision of the guidelines in order to disambiguate the cases we perceived as confusing. We believe this would make the annotation scheme easier to apply to future negation annotation efforts. It would also provide an opportunity to reannotate $\text{NoReC}_{\text{neg}}$ and thus remove inconsistencies and annotation errors, which in turn might lead to even better modeling results.

In the medical and clinical subfield, we emphasize the annotation of more data as a priority. As we know, access to clinical data is strictly regulated. In the case of biomedical negation research, however, data access should be less of an obstacle. Tidsskriftet for Den Norske Legeforening[3] *(eng: The Journal of the Norwegian Medical Association)* is a source of medical research papers in Norwegian, covering the whole range of medical specialties. Since the contents of $\text{NorMed}_{\text{neg}}$ originate from the domain of gastrointestinal surgery only, one interesting task would be to annotate negation in texts belonging to other medical fields. Another important contribution would be a more fine-grained analysis of the amount of annotated data required to achieve satisfactory results for negation resolution. We would also like to mention that another round of

[3]Webpage: https://tidsskriftet.no/

quality assurance of $\text{NorMed}_{\text{neg}}$ could be beneficial. We believe the number of annotation errors is low, but the results from chapter 6 indicate that there are a few missing annotations.

If clinical text is accessed and annotated, it could be used for evaluation of our models, both the system fine-tuned on review articles only, and the system with an additional round of fine-tuning on biomedical text. Further fine-tuning on clinical data would be an essential task as well. We also highlight the need for a more detailed evaluation of the applicability of the $\text{NoReC}_{\text{neg}}$ annotation scheme in clinical negation resolution, which requires medical expertise. If considered unsuitable, we would advise the adaptation of the scheme to a clinical context and the training of a negation resolution system based on this.

As a final remark, experiments with multilingual BERT (Devlin et al., 2019) have shown promising results for clinical negation scope resolution for other languages *without* fine-tuning on clinical data (Hartmann and Søgaard, 2021). An evaluation of this with respect to Norwegian clinical text would be useful.

Bibliography

Agarwal, Shashank and Hong Yu (2010). "Biomedical negation scope detection with conditional random fields." In: *Journal of the American Medical Informatics Association* 17.6, pp. 696–701. DOI: 10.1136/jamia.2010.003228. URL: https://doi.org/10.1136/jamia.2010.003228.

Allvin, Helen, Elin Carlsson, Hercules Dalianis, Riitta Danielsson-Ojala, Vidas Daudaravičius, Martin Hassel, Dimitrios Kokkinakis, Heljä Lundgrén-Laine, Gunnar H Nilsson, Øystein Nytrø, Sanna Salanterä, Maria Skeppstedt, Hanna Suominen, and Sumithra Velupillai (2011). "Characteristics of Finnish and Swedish intensive care nursing narratives: a comparative analysis to support the development of clinical language technologies." In: *Journal of Biomedical Semantics* 2 (Suppl 3). DOI: 10.1186/2041-1480-2-S3-S1. URL: https://doi.org/10.1186/2041-1480-2-S3-S1.

Budrionis, Andrius, Hercules Dalianis, Kassaye Yitbarek Yigzaw, Alexandra Makhlysheva, and Taridzo Chomutare (2018). "Negation detection in Norwegian medical text : Porting a Swedish NegEx to Norwegian. Work in progress." In: Compilation of abstracts in The Seventh Swedish Language Technology Conference (SLTC-2018), pp. 74–77. URL: https://www.diva-portal.org/smash/get/diva2:1358014/FULLTEXT01.pdf.

Chapman, Wendy W., Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan (2001). "A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries." In: *Journal of Biomedical Informatics* 34.5, pp. 301–310. DOI: 10.1006/jbin.2001.1029. URL: https://doi.org/10.1006/jbin.2001.1029.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2020). "Unsupervised Cross-lingual Representation Learning at Scale." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. URL: https://aclanthology.org/2020.acl-main.747.

Dalianis, Hercules (2018). *Clinical Text Mining*. Cham: Springer International Publishing. DOI: 10.1007/978-3-319-78503-5. URL: http://link.springer.com/10.1007/978-3-319-78503-5 (visited on 04/09/2023).

Dalianis, Hercules and Maria Skeppstedt (2010). "Creating and evaluating a consensus for negated and speculative words in a Swedish clinical corpus." In: *Proceedings of the Workshop on Negation and Speculation in*

Natural Language Processing, pp. 5–13. URL: https://aclanthology.org/W10-3102.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

Dozat, Timothy and Christopher D. Manning (2018). "Simpler but More Accurate Semantic Dependency Parsing." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 484–490. DOI: 10.18653/v1/P18-2077. URL: https://aclanthology.org/P18-2077.

Elazhary, Hanan (2017). "NegMiner: An Automated Tool for Mining Negations from Electronic Narrative Medical Documents." In: *International Journal of Intelligent Systems and Applications* 9.4, pp. 14–22. DOI: 10.5815/ijisa.2017.04.02. URL: https://doi.org/10.5815/ijisa.2017.04.02.

Fancellu, Federico, Adam Lopez, and Bonnie Webber (2016). "Neural Networks For Negation Scope Detection." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 495–504. DOI: 10.18653/v1/P16-1047. URL: https://aclanthology.org/P16-1047.

fin (n.d.). In: *Bokmålsordboka*. Språkrådet (eng: The Language Council of Norway) and Universitetet i Bergen (eng: University of Bergen). URL: https://ordbokene.no/bm/14683 (visited on 05/07/2023).

forhindre (n.d.). In: *Bokmålsordboka*. Språkrådet (eng: The Language Council of Norway) and Universitetet i Bergen (eng: University of Bergen). URL: https://ordbokene.no/bm/16336 (visited on 02/20/2023).

Grigonyte, Gintarė, Maria Kvist, Sumithra Velupillai, and Mats Wirén (2014). "Improving Readability of Swedish Electronic Health Records through Lexical Simplification: First Results." In: *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pp. 74–83. DOI: 10.3115/v1/W14-1209. URL: https://aclanthology.org/W14-1209.

Hagemann, Kristin (2020). *determinativ (grammatikk)*. In: *Store Norske Leksikon (eng: The Great Norwegian Encyclopedia) at snl.no*. URL: https://snl.no/determinativ_-_grammatikk (visited on 02/16/2023).

Hagemann, Kristin (2023). *kopula*. In: *Store Norske Leksikon (eng: The Great Norwegian Encyclopedia) at snl.no*. URL: https://snl.no/kopula (visited on 02/17/2023).

Harkema, Henk, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman (2009). "ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports." In: *Journal of Biomedical Informatics* 42.5, pp. 839–851. DOI: 10.1016/j.jbi.2009.05.002. URL: https://www.sciencedirect.com/science/article/pii/S1532046409000744.

Hartmann, Mareike and Anders Søgaard (2021). "Multilingual Negation Scope Resolution for Clinical Text." In: *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pp. 7–18. URL: https://aclanthology.org/2021.louhi-1.2.

Helseforskningsloven (2008). *Lov om medisinsk og helsefaglig forskning (LOV-2008-06-20-44)*. URL: https://lovdata.no/dokument/NL/lov/2008-06-20-44.

Helwe, Chadi, Simon Coumes, Chloé Clavel, and Fabian Suchanek (2022). "TINA: Textual Inference with Negation Augmentation." In: *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4086–4099. URL: https://aclanthology.org/2022.findings-emnlp.301.

hindre (n.d.). In: *Bokmålsordboka*. Språkrådet (eng: The Language Council of Norway) and Universitetet i Bergen (eng: University of Bergen). URL: https://ordbokene.no/bm/23615 (visited on 02/20/2023).

Hossain, Md Mosharaf, Antonios Anastasopoulos, Eduardo Blanco, and Alexis Palmer (2020). "It's not a Non-Issue: Negation as a Source of Error in Machine Translation." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3869–3885. DOI: 10.18653/v1/2020.findings-emnlp.345. URL: https://aclanthology.org/2020.findings-emnlp.345.

Hosseini, Arian, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordoni, and Aaron Courville (2021). "Understanding by Understanding Not: Modeling Negation in Language Models." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1301–1312. DOI: 10.18653/v1/2021.naacl-main.102. URL: https://aclanthology.org/2021.naacl-main.102.

Isenius, Niklas (2012). "Abbreviation detection in Swedish medical records. The development of SCAN, a Swedish clinical abbreviation normalizer." Master's thesis. Department of Computer and Systems Sciences, Stockholm University. URL: https://daisy.dsv.su.se/fil/visa?id=77070.

Isenius, Niklas, Sumithra Velupillai, and Maria Kvist (2012). "Initial results in the development of SCAN: a Swedish Clinical Abbreviation Normalizer." In: *The CLEF 2012 Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis*. URL: https://ceur-ws.org/Vol-1178/CLEF2012wn-CLEFeHealth-IseniusEt2012.pdf.

Jiménez-Zafra, Salud María, Mariona Taulé, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López, and M. Antónia Martí (2018). "SFU ReviewSP-NEG: a Spanish corpus annotated with negation for sentiment analysis. A typology of negation patterns." In: *Language Resources and Evaluation* 52.2, pp. 533–569. DOI: 10.1007/s10579-017-9391-x. URL: https://doi.org/10.1007/s10579-017-9391-x.

Jiménez-Zafra, Salud Marıa, Roser Morante, Marıa Teresa Martın-Valdivia, and L. Alfonso Ureña-López (2020). "Corpora Annotated with Negation: An Overview." In: *Computational Linguistics* 46.1, pp. 1–52. DOI: 10.1162/coli_a_00371. URL: https://aclanthology.org/2020.cl-1.5.

Khandelwal, Aditya and Benita Kathleen Britto (2020). "Multitask Learning of Negation and Speculation using Transformers." In: *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pp. 79–87. DOI: 10.18653/v1/2020.louhi-1.9. URL: https://aclanthology.org/2020.louhi-1.9.

Khandelwal, Aditya and Suraj Sawant (2020). "NegBERT: A Transfer Learning Approach for Negation Detection and Scope Resolution." In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 5739–5748. URL: https://aclanthology.org/2020.lrec-1.704.

Konstantinova, Natalia, Sheila C.M. de Sousa, Noa P. Cruz, Manuel J. Maña, Maite Taboada, and Ruslan Mitkov (2012). "A review corpus annotated for negation, speculation and their scope." In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 3190–3195. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/533_Paper.pdf.

Kummervold, Per, Freddy Wetjen, and Javier de la Rosa (2022). "The Norwegian Colossal Corpus: A Text Corpus for Training Large Norwegian Language Models." In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 3852–3860. URL: https://aclanthology.org/2022.lrec-1.410.

Kummervold, Per E, Javier De la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld (2021). "Operationalizing a National Digital Library: The Case for a Norwegian Transformer Model." In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pp. 20–29. URL: https://aclanthology.org/2021.nodalida-main.3.

Kurtz, Robin, Stephan Oepen, and Marco Kuhlmann (2020). "End-to-End Negation Resolution as Graph Parsing." In: *Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies*, pp. 14–24. DOI: 10.18653/v1/2020.iwpt-1.3. URL: https://aclanthology.org/2020.iwpt-1.3.

Kutuzov, Andrey, Jeremy Barnes, Erik Velldal, Lilja Øvrelid, and Stephan Oepen (2021). "Large-Scale Contextualised Language Modelling for Norwegian." In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pp. 30–40. URL: https://aclanthology.org/2021.nodalida-main.4.

Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer (2016). "Neural Architectures for Named Entity Recognition." In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 260–270. DOI: 10.18653/v1/N16-1030. URL: https://aclanthology.org/N16-1030.

Lapponi, Emanuele, Erik Velldal, Lilja Øvrelid, and Jonathon Read (2012). "UiO 2: Sequence-labeling Negation Using Dependency Features." In: **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pp. 319–327. URL: https://aclanthology.org/S12-1042.

Lazib, Lydia, Bing Qin, Yanyan Zhao, Weinan Zhang, and Ting Liu (2018). "A syntactic path-based hybrid neural network for negation scope detection." In: *Frontiers of Computer Science* 14, pp. 84–94. DOI: 10.1007/s11704-018-7368-6. URL: https://link.springer.com/article/10.1007/s11704-018-7368-6.

Li, Junhui, Guodong Zhou, Hongling Wang, and Qiaoming Zhu (2010). "Learning the Scope of Negation via Shallow Semantic Parsing." In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 671–679. URL: https://aclanthology.org/C10-1076.

Li, Xin, Lidong Bing, Wenxuan Zhang, and Wai Lam (2019). "Exploiting BERT for End-to-End Aspect-based Sentiment Analysis." In: *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pp. 34–41. DOI: 10.18653/v1/D19-5505. URL: https://aclanthology.org/D19-5505.

Lison, Pierre, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid (2021). "Anonymisation Models for Text Data: State of the art, Challenges and Future Directions." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4188–4203. DOI: 10.18653/v1/2021.acl-long.323. URL: https://aclanthology.org/2021.acl-long.323.

Liu, Hongfang, Yves A. Lussier, and Carol Friedman (2001). "A study of abbreviations in the UMLS." In: *AMIA Annual Symposium Proceedings*, pp. 393–397. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243414/.

Liu, Qianchu, Federico Fancellu, and Bonnie Webber (2018). "NegPar: A parallel corpus annotated for negation." In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. URL: https://aclanthology.org/L18-1547.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. Version: 1. arXiv: 1907.11692 [cs.CL].

Loshchilov, Ilya and Frank Hutter (2019). "Decoupled Weight Decay Regularization." In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

Lövestam, Elin, Sumithra Velupillai, and Maria Kvist (2014). "Abbreviations in Swedish Clinical Text – use by three professions." In: *e-Health–For Continuity of Care*. Vol. 205. Studies in Health Technology and Informatics. IOS Press, pp. 720–724. DOI: 10.3233/978-1-61499-432-9-720. URL: https://doi.org/10.3233/978-1-61499-432-9-720.

Luo, Huaishao, Lei Ji, Tianrui Li, Daxin Jiang, and Nan Duan (2020). "GRACE: Gradient Harmonized and Cascaded Labeling for Aspect-based Sentiment Analysis." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 54–64. DOI: 10.18653/v1/2020.findings-emnlp.6. URL: https://aclanthology.org/2020.findings-emnlp.6.

Mæhlum, Petter, Jeremy Barnes, Robin Kurtz, Lilja Øvrelid, and Erik Velldal (2021). "Negation in Norwegian: an annotated dataset." In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pp. 299–308. URL: https://aclanthology.org/2021.nodalida-main.30.

Mehrabi, Saeed, Anand Krishnan, Sunghwan Sohn, Alexandra M. Roch, Heidi Schmidt, Joe Kesterson, Chris Beesley, Paul Dexter, C. Max Schmidt, Hongfang Liu, and Mathew Palakal (2015). "DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx." In: *Journal of Biomedical Informatics* 54, pp. 213–219. DOI: 10.1016/j.jbi.2015.02.010. URL: https://doi.org/10.1016/j.jbi.2015.02.010.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013a). *Efficient Estimation of Word Representations in Vector Space*. Version: 3. arXiv: 1301.3781 [cs.CL].

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean (2013b). *Distributed Representations of Words and Phrases and their Compositionality*. Version: 1. arXiv: 1310.4546 [cs.CL].

Morante, Roser and Eduardo Blanco (2012). "*SEM 2012 Shared Task: Resolving the Scope and Focus of Negation." In: **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pp. 265–274. URL: https://aclanthology.org/S12-1035.

Morante, Roser and Eduardo Blanco (2021). "Recent advances in processing negation." In: *Natural Language Engineering* 27.2, pp. 121–130. DOI: 10.1017/S1351324920000534. URL: https://doi.org/10.1017/S1351324920000534.

Morante, Roser and Walter Daelemans (2009). "A Metalearning Approach to Processing the Scope of Negation." In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pp. 21–29. URL: https://aclanthology.org/W09-1105.

Morante, Roser and Walter Daelemans (2012). "ConanDoyle-neg: Annotation of negation cues and their scope in Conan Doyle stories." In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 1563–1568. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/221_Paper.pdf.

Morante, Roser, Sarah Schrauwen, and Walter Daelemans (2011). "Annotation of negation cues and their scope: Guidelines v1." In: *Computational linguistics and psycholinguistics technical report series, CTRS-003*, pp. 1–42. URL: https://medialibrary.uantwerpen.be/oldcontent/container2712/files/ctrs3.pdf.

Morante, Roser and Caroline Sporleder (2012). "Modality and Negation: An Introduction to the Special Issue." In: *Computational Linguistics* 38.2, pp. 223–260. DOI: 10.1162/COLI_a_00095. URL: https://aclanthology.org/J12-2001.

Mosbach, Marius, Maksym Andriushchenko, and Dietrich Klakow (2021). "On the Stability of Fine-tuning BERT: Misconceptions, Explanations,

and Strong Baselines." In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=nzpLWnVAyah.

Nasjonalbiblioteket (n.d.). *Models*. URL: https://ai.nb.no/models/ (visited on 05/11/2023).

Nasjonalbiblioteket AI lab (2021). *Pretrained Models*. URL: https://github.com/NBAiLab/notram#pretrained-models (visited on 05/11/2023).

Névéol, Aurélie, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum (2018). "Clinical natural language processing in languages other than English: opportunities and challenges." In: *Journal of Biomedical Semantics* 9. DOI: 10.1186/s13326-018-0179-8. URL: https://doi.org/10.1186/s13326-018-0179-8.

Nielsen, Dan Saattrup (2023). *NLU Benchmark*. URL: https://scandeval.github.io/nlu-benchmark/ (visited on 05/11/2023).

Nizamuddin, Uddin and Hercules Dalianis (2014). "Detection of spelling errors in Swedish clinical text." In: *1st Nordic workshop on evaluation of spellchecking and proofing tools (NorWEST2014), SLTC 2014*. URL: https://divvun.no/sv/events/workshops/NorWEST2014/abstracts/Uddin_Dalianis.pdf.

Nordic Language Processing Laboratory (2023). *Vectors/norlm/norbert*. URL: http://wiki.nlpl.eu/Vectors/norlm/norbert (visited on 05/11/2023).

Nordquist, Richard (2020). *Definition and Examples of Postmodifiers in English Grammar*. URL: https://www.thoughtco.com/postmodifier-grammar-1691519 (visited on 05/07/2023).

Olsson, May (2011). "Vem begriper patientjournalen?" Bachelor's thesis. Linnaeus University. URL: https://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-13058.

Packard, Woodley, Emily M. Bender, Jonathon Read, Stephan Oepen, and Rebecca Dridan (2014). "Simple Negation Scope Resolution through Deep Parsing: A Semantic Solution to a Semantic Problem." In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 69–78. DOI: 10.3115/v1/P14-1007. URL: https://aclanthology.org/P14-1007.

Pilan, Ildiko, Pål H. Brekke, and Lilja Øvrelid (2020). "Building a Norwegian Lexical Resource for Medical Entity Recognition." In: *Proceedings of the LREC 2020 Workshop on Multilingual Biomedical Text Processing (MultilingualBIO 2020)*, pp. 9–14. URL: https://aclanthology.org/2020.multilingualbio-1.2.

Prabhakaran, Vinodkumar, Owen Rambow, and Mona Diab (2010). "Automatic Committed Belief Tagging." In: *Coling 2010: Posters*, pp. 1014–1022. URL: https://aclanthology.org/C10-2117.

Qian, Zhong, Peifeng Li, Qiaoming Zhu, Guodong Zhou, Zhunchen Luo, and Wei Luo (2016). "Speculation and Negation Scope Detection via Convolutional Neural Networks." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 815–825. DOI: 10.18653/v1/D16-1078. URL: https://aclanthology.org/D16-1078.

Read, Jonathon, Erik Velldal, Lilja Øvrelid, and Stephan Oepen (2012). "UiO1: Constituent-Based Discriminative Ranking for Negation Resolution." In: **SEM 2012: The First Joint Conference on Lexical and Com-*

putational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), pp. 310–318. URL: https://aclanthology.org/S12-1041.

Reitan, Johan, Jørgen Faret, Björn Gambäck, and Lars Bungum (2015). "Negation Scope Detection for Twitter Sentiment Analysis." In: *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 99–108. DOI: 10.18653/v1/W15-2914. URL: https://aclanthology.org/W15-2914.

Ren, Yafeng, Hao Fei, and Qiong Peng (2018). "Detecting the Scope of Negation and Speculation in Biomedical Texts by Using Recursive Neural Network." In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 739–742. DOI: 10.1109/BIBM.2018.8621261. URL: https://ieeexplore.ieee.org/document/8621261.

Sadhukhan, Debarati (2021). "Building and evaluating the NegEx negation detection system for Norwegian biomedical text." Master's thesis. Department of Computer and Systems Sciences, Stockholm University. URL: https://daisy.dsv.su.se/fil/visa?id=233579.

Samuel, David, Andrey Kutuzov, Samia Touileb, Erik Velldal, Lilja Øvrelid, Egil Rønningstad, Elina Sigdel, and Anna Sergeevna Palatkina (2023). "NorBench – A Benchmark for Norwegian Language Models." In: *The 24th Nordic Conference on Computational Linguistics*. URL: https://openreview.net/forum?id=WgxNONkAbz.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016). "Neural Machine Translation of Rare Words with Subword Units." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725. DOI: 10.18653/v1/P16-1162. URL: https://aclanthology.org/P16-1162.

Skeppstedt, Maria (2011). "Negation detection in Swedish clinical text: An adaption of NegEx to Swedish." In: *Journal of Biomedical Semantics* 2 (Suppl 3). DOI: 10.1186/2041-1480-2-S3-S3. URL: https://doi.org/10.1186/2041-1480-2-S3-S3.

Skeppstedt, Maria, Maria Kvist, and Hercules Dalianis (2012). "Rule-based Entity Recognition and Coverage of SNOMED CT in Swedish Clinical Text." In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 1250–1257. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/521_Paper.pdf.

Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii (2012). "brat: a Web-based Tool for NLP-Assisted Text Annotation." In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 102–107. URL: https://aclanthology.org/E12-2021.

Tanushi, Hideyuki, Hercules Dalianis, Martin Duneld, Maria Kvist, Maria Skeppstedt, and Sumithra Velupillai (2013). "Negation Scope Delimitation in Clinical Text Using Three Approaches: NegEx, PyConTextNLP and SynNeg." In: *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pp. 387–397. URL: https://aclanthology.org/W13-5635.

Truong, Thinh, Timothy Baldwin, Trevor Cohn, and Karin Verspoor (2022). "Improving negation detection with negation-focused pre-training." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4188–4193. DOI: 10.18653/v1/2022.naacl-main.309. URL: https://aclanthology.org/2022.naacl-main.309.

Velldal, Erik, Lilja Øvrelid, Eivind Alexander Bergem, Cathrine Stadsnes, Samia Touileb, and Fredrik Jørgensen (2018). "NoReC: The Norwegian Review Corpus." In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. URL: https://aclanthology.org/L18-1661.

Velldal, Erik, Lilja Øvrelid, Jonathon Read, and Stephan Oepen (2012). "Speculation and Negation: Rules, Rankers, and the Role of Syntax." In: *Computational Linguistics* 38.2, pp. 369–410. DOI: 10.1162/COLI_a_00126. URL: https://aclanthology.org/J12-2005.

Vincze, Veronika, György Szarvas, Richárd Farkas, György Móra, and János Csirik (2008). "The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes." In: *BMC Bioinformatics* 9 (Suppl 11). DOI: 10.1186/1471-2105-9-S11-S9. URL: https://doi.org/10.1186/1471-2105-9-S11-S9.

Wiegand, Michael, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, and Andrés Montoyo (2010). "A survey on the role of negation in sentiment analysis." In: *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pp. 60–68. URL: https://aclanthology.org/W10-3111.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean (2016). *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. Version: 2. arXiv: 1609.08144 [cs.CL].

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel (2021). "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498. DOI: 10.18653/v1/2021.naacl-main.41. URL: https://aclanthology.org/2021.naacl-main.41.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le (2019). "XLNet: Generalized Autoregressive Pretraining for Language Understanding." In: *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pp. 5753–5763. URL: https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html.