Negation in Norwegian: an annotated dataset

Petter Mæhlum[*], Jeremy Barnes[*], Robin Kurtz[†], Lilja Øvrelid[*] and Erik Velldal[*]

[*]University of Oslo, Department of Informatics

[†]National Library of Sweden, KBLab

{pettemae|jeremycb|liljao|erikve}@ifi.uio.no

robin.kurtz@kb.se

Abstract

This paper introduces NoReC$_{neg}$ – the first annotated dataset of negation for Norwegian. Negation cues and their in-sentence scopes have been annotated across more than 11K sentences spanning more than 400 documents for a subset of the Norwegian Review Corpus (NoReC). In addition to providing in-depth discussion of the annotation guidelines, we also present a first set of benchmark results based on a graph-parsing approach.

1 Introduction

This paper introduces a new data set annotating negation for Norwegian. As shown in the example below, the annotations identify both negation *cues* (in bold) and their *scopes* (in brackets) within the sentence:

(1) *Men kanskje* ***ikke*** *[helt troverdig]* .
But maybe not completely credible .
'But maybe not completely credible.'

The underlying corpus is the NoReC$_{fine}$ data set (Øvrelid et al., 2020) – a subset of the Norwegian Review Corpus (NoReC) (Velldal et al., 2018) annotated for fine-grained sentiment, comprising professional reviews from a range of different domains. The new data set introduced here, named NoReC$_{neg}$, is the first data set of negation for Norwegian. We also present experimental results for negation resolution based on a graph-parsing approach shown to yield state-of-the-art results for other languages. All the resources described in the paper – the data set, the annotation guidelines, the models and the associated code – are made publicly available.[1]

The rest of the paper is structured as follows. We start by reviewing related work on negation for other languages in Section 2, with regards to both annotation and modeling. In Section 3 we detail our annotation guidelines, the annotation procedure and further present an analysis of inter-annotator agreement. In Section 4 we then summarize the statistics of the final annotated data set, before presenting the first benchmark results for negation resolution in Section 5. Before concluding, we finally provide a discussion of future work in Section 6.

[1] https://github.com/ltgoslo/norec_neg

2 Related Work

Below we discuss related work on negation, starting with datasets before moving on to modeling.

2.1 Datasets

While NoReC$_{neg}$ is the first dataset annotated for negation for Norwegian, there are a number of existing negation datasets for a range of other languages, such as Chinese (Zou et al., 2016), Dutch (Afzal et al., 2014), English (Pyysalo et al., 2007; Vincze et al., 2008; Morante and Daelemans, 2012; Councill et al., 2010; Konstantinova et al., 2012), German (Cotik et al., 2016), Spanish (Jiménez-Zafra et al., 2018; Diaz et al., 2017), Swedish (Dalianis and Velupillai, 2010; Skeppstedt, 2011), Italian (Altuna et al., 2017), and Japanese (Matsuyoshi et al., 2014). Jiménez-Zafra et al. (2020) provide a thorough survey of existing negation datasets. A large proportion of negation corpora are based on data from the biomedical or clinical domain (Vincze et al., 2008; Dalianis and Velupillai, 2010; Cotik et al., 2016; Diaz et al., 2017). We will here focus on the corpora that are most relevant to the current annotation effort: the SFU Corpus and the ConanDoyle-neg corpus. The SFU corpus also annotates review data, hence is similar to our work in terms of text type, whereas ConanDoyle-neg is one of the most widely used datasets in the field.

The English (Konstantinova et al., 2012) and

Spanish (Jiménez-Zafra et al., 2018) parts of the SFU Review Corpus contain reviews from eight domains (books, cars, computers, cookware, hotels, movies, music, phones) which have been annotated for sentiment at document-level, as well as negation and speculation at sentence-level. The annotation scheme for negation is based primarily on the guidelines developed for the biomedical BioScope corpus (Vincze et al., 2008), which largely employ syntactic criteria for the determination of scope, choosing the maximal syntactic unit that contains the negated content. Unlike BioScope, however, negation cues are not included within the scope in SFU. The corpus does not annotate affixal cues, e.g. *im-* in *impossible*.

The English ConanDoyle-neg corpus contains Sherlock Holmes stories manually annotated for negation cues, scopes, and events (Morante and Daelemans, 2012) and was employed in the 2012 *SEM shared task on negation detection (Morante and Blanco, 2012). The annotation scheme is also based on the scheme employed for the BioScope corpus (Vincze et al., 2008), but with important modifications. In ConanDoyle-neg, the cue is not included in the scope, and it annotates a wide range of cue types, i.e., both sub-token (affixal), single token and multi-token negation cues. Scopes may furthermore be discontinuous, often an effect of the requirement to include the subject within the negation scope. This is in contrast to the annotation scheme found in the SFU corpus, where subjects are not included in the negation scope. Note that the NegPar corpus contains a re-annotated version of the ConanDoyle-neg corpus, which fixes known bugs and also adds Chinese data (Liu et al., 2018).

2.2 Modeling

Traditional approaches to the task of negation detection have typically employed a wide range of hand-crafted features, and often linguistically informed, derived from constituency parsing (Read et al., 2012; Packard et al., 2014), dependency parsing (Lapponi et al., 2012), or Minimal Recursion Semantics structures created by an HPSG parser (Packard et al., 2014). Scope resolution in particular has often been approached as a sequence labeling task, as pioneered by Morante and Daelemans (2009) and later done in several other works (Lapponi et al., 2012; White, 2012; Enger et al., 2017). More recently, neural approaches have been successfully applied to the task. Qian et al. (2016) propose a CNN model for negation scope detection on the abstracts section of the BioScope corpus, which operates over syntactic paths between the cue and candidate tokens. Fancellu et al. (2016) present and compare two neural architectures for the task of negation scope detection on the ConanDoyle-neg corpus: a simple feed-forward network and a bidirectional LSTM. Note that these more recent neural systems disregard the task of cue detection altogether (Fancellu et al., 2016; Qian et al., 2016; Fancellu et al., 2017), relying instead on gold cues and focusing solely on the task of scope detection.

Finally, Kurtz et al. (2020) cast negation resolution as a graph parsing problem and perform full negation resolution using a dependency graph parser (Dozat and Manning, 2018) to jointly predict cues and scopes. The neural model uses a BiLSTM to create token-level representations, and then includes two feed-forward networks to create head- and dependent-specific token representations. Finally, each possible head-dependent combination is scored using a bilinear model. Despite the conceptual simplicity, this model achieves state-of-the-art results. As such, we use this model to evaluate our annotations and include further details in Section 5.

3 Annotations

In the following section we present our negation annotation effort in more detail, including the underlying source of the data. The guidelines we have developed for the annotation of negation cues and scopes in Norwegian are mainly adapted from ConanDoyle-neg (Morante and Daelemans, 2009), NegPar (Liu et al., 2018), and the Spanish SFU corpus (Jiménez-Zafra et al., 2018), modified to suit Norwegian, and with simplifications that will be discussed below. Note that while the complete set of guidelines is distributed with the corpus, we provide a brief overview below together with examples, also discussing inter-annotator agreement.

3.1 The underlying corpus

The negation annotations described below are added to the existing NoReC$_{fine}$ data set[2] (Øvrelid et al., 2020) – a subset of the Norwegian Review Corpus (NoReC) annotated for fine-grained sentiment. The negation layer of the corpus is named

[2] https://github.com/ltgoslo/norec_fine

NoReC$_{neg}$. The full NoReC corpus (Velldal et al., 2018) contains professional reviews from several Norwegian online news sites, spanning a range of different domains, like music, literature, products, movies, restaurants, and more. While NoReC contains more than 43,000 full-text reviews, the subset annotated in NoReC$_{fine}$, and hence also NoReC$_{neg}$, includes 414 full reviews, comprising 11,346 sentences. Note that there are two official standards for written Norwegian; Bokmål (the majority variant) and Nynorsk. While the data set contains a majority of documents written according to the Bokmål standard, four Nynorsk documents are also included.

3.2 Negation in Norwegian

Since our starting point for guideline development is English, we will here discuss linguistic differences between the expression of negation in the two languages. Generally speaking, Norwegian negation does not differ greatly from English. The main means of negating a proposition is by using adverbs, prepositions and quantifiers. The largest differences between the two are syntactic in nature and concern the placement of adverbials, caused by the fact that Norwegian, unlike English, is a V2-language. One clear difference with practical consequences is that certain Norwegian negation cues inflect for grammatical gender and number, notable examples being *ingen (ingen, inga, intet)* 'no' and *løs (-løs, -løst, -løse)* '-less', as seen in example (2) for the affixally negated (a) *meningsløst* 'meaningless' with the neuter ending, (b) *hensynsløse* 'inconsiderate' with plural inflection, and (c) *smakløs* 'tasteless' with no inflection. This property of Norwegian means that there are likely a larger number of different tokens functioning as cues in Norwegian, as compared to English.

(2) (a) [...] blir ganske menings**løst**
(b) [...] hensyns**løse** regnskog-ødeleggere
(c) [...] men ikke smak**løs** .

The discussion of negation in the Norwegian Reference Grammar (Faarlund et al., 1997) is largely limited to a selected few of the possible cues, e.g., *ikke* 'not', *ingen* 'none, no-one' and related forms, and the preposition *uten* 'without'. Golden et al. (2014) contains a brief comment on lexical negation, where they mention *nektende verb* 'negating verbs'. They also mention negative polarity items under a discussion of separate words and expressions in negations.

3.3 Negation cues

A negation cue is a word or a set of words that serve to signal negation. In our annotation scheme we annotate both single token cues, such as adverbs like *ikke* 'not', *aldri* 'never', prepositions, e.g., *uten* 'without', and quantifiers like *ingen* 'no'. We also annotate multi-word cues, such as *(på) ingen måte*, 'in no way', as well as morphological or affixal negation cues, i.e. affixes such as *u-* 'un-/dis-/non-' and *-løs* '-less'. Example (3) shows the widely used negative adverb *aldri* 'never', which scopes over the whole sentence, including the subject *Jeg* 'I', whereas (4) exemplifies the negative determiner *ingen* 'no' which occurs in two conjoined noun phrase objects, where both negation cues scope over the following noun as well as the preceding subject and main verb.

(3) *[Jeg har]* ***aldri*** *[hørt henne synge bedre fra en scene]*
I have never heard her sing better from a stage
'I have never heard her sing better from a stage'

(4) *[Den stiller]* ***ingen*** *[spørsmål] og [gir]* ***ingen*** *[svar]* .
It asks no questions and gives no answers.
'It asks no questions and gives no answers.'

Multi-word cues Multi-word cues are negation cues that span more than one token. These may further be *discontinuous*, as in the case of *(h)verken ... eller* 'neither ... nor', as seen in example (5). As noted by Morante and Daelemans (2012), multi-word cues tend to be fixed/idiomatic expressions – an observation that is largely true for Norwegian as well. One practical difference between the annotation scheme in Morante and Daelemans (2009) and ours, is that we omit prepositions and particles related to these expressions, as in (6), in favor of creating less variation that might create noise in the data, especially in cases where multiple prepositions are associated with similar cues and the association is less fixed.

(5) *[...]* ***verken*** *[manus]* ***eller*** *[skuespillere trekker oss inn på en engasjerende måte]* .
[...] neither script nor actors pull us in on a engaging method .
'[...] neither script nor actors pull us inside in an engaging way'

(6) *Og mest av alt* ***fraværet*** *av [mer*
And most of all the.absence of more
enn bare et kvarter musikk] .
than just a quarter music .
'And most of all, the absence of more than just 15 minutes of music.'

Affixal cues We annotate both free-standing and affixal negation cues. The affixal cues form a rather closed group of cues, with the prefix *u-* and the suffix *-løs* being the most common. However, our annotations show that there is lexical variation, with less common cues such as *-fri* '-free' and *-tom* '-empty'.

Negation vs. Modality One difficulty in annotating cues is to separate between cases of negation in isolation and cases where negation and modality interact. Cases where modality and negation are inseparable, as in *neppe* 'barely' are not annotated, but cases of negation where the modality can be separated, either by it scoping over the negation, or the negation scoping over it, were annotated as negations.

Lexical negation As mentioned above, the discussion of lexical negation in a Norwegian context is limited. We borrow the term 'lexical negation' from Jiménez-Zafra et al. (2020), who split cues into syntactic, lexical, and morphological/affixal, and use the lexical category to mean words that fall outside the 'syntactic' and more frequent cues, like negative adverbs and determiners. Examples from Norwegian include verbal constructs, e.g., *la være* 'refrain from' or *forsvinne* 'disappear' as in (7), and nouns such as *mangel* 'lack'.

(7) *...[Irritasjonen]* ***forsvant*** *da maten*
...the.irritation disappeared when the.food
kom .
arrived .
'...The irritation disappeared when the food arrived.'

Lexicalization and idioms The words that are used as negation cues might also have other functions, and are in some cases part of fossilized expressions. The annotators were instructed to refrain from annotating affixal cues that no longer signal negation. Lexicalization, in particular, is a challenge when it comes to affixal negation, as it can be difficult even for native speakers to judge whether something should be treated as a negation or not. Some cases are clearer than others, such as *uansett* 'regardless', which stems from *ansett* 'viewed/respected', which it clearly does not negate, on the one hand, and on the other hand *usikker* 'uncertain', whose non-negated form *sikker* 'certain' is also frequent. The absence of the non-negated version of the lemma in the language might be a good indicator of lexicalization, and annotators were instructed to avoid annotating such words.

In addition to lexicalized items, there are also cases where a cue can have more than one meaning. One frequent case is the prefix *u-* with nominal roots, a construction that usually results in nouns meaning bad x, as in *uår* lit. 'un-year', which means 'a bad year', or *uvenn*, lit. 'unfriend', meaning 'enemy'. The annotators were instructed to try and dismantle the word in order to see if the word made sense without the negative prefix, in which case it would indicate that it is not completely lexicalized. Even so, these are often difficult judgements for the annotators to make. Furthermore, nominalizations of negated adjectives, such as *uttrykksløshet* 'expressionlessness' and *umenneskelighet* 'inhumanity' were not to be annotated.

Table 1 presents the ten most common cues found in the corpus, where we find both affixal and single token cues. We see that variation in the data is further caused by spelling differences. The adverb *ikke* 'not' can also be used affixally, often, but not always, with a hyphen, as in *ikke-produksjonsklart* 'not-production-ready'. The variation is also due in part to the two language varieties present in the dataset, e.g in the case of Bokmål *ikke* 'not' and Nynorsk *ikkje* 'not'.

3.4 Negation scopes

The scope of a negation is the part of a sentence that has its truth value inverted by the presence of a negation cue. In our annotation scheme, cues are never part of the scope. Subjects are included in the scope if the negation scopes over the main verb, which usually means that the whole proposition is negated, and if the subject or object of a sentence is negated by a determiner or similar, the whole sentence is in the scope, apart from certain fixed elements discussed below. Phrase linking conjunctions are not included. Furthermore, scopes tend be discontinuous. In many cases this is simply due to the the fact that in most sentences, the subject precedes the negation cue, while the predicate follows it.

Cue	Trans.	Frequency	Amb. Rate
ikke	not	1,364	3
u-	un-/dis-/non-	514	83
uten	without	190	0
ingen	none/nobody	134	0
-løs	-less	123	5
aldri	never	95	6
mangle	lack	43	14
ingenting	nothing	23	0
ikkje	not	23	0
verken	neither	21	30

Table 1: List of the 10 most common cues found in the corpus, their translation to English, their frequency as a cue, as well as their ambiguity rate (Amb. Rate), which is defined as $1-$ (the frequency as a cue / the absolute frequency) $\times 100$.

Implicit scope The scope of a cue can be implicit, meaning it is understood from the context. In practice the scope is often expressed in a sentence before or after the cue itself. This is in particular the case with the interjection *nei* 'no', which usually refers back to the proposition it negates. Since our annotation does not span across sentence boundaries, the scope is annotated as implicit in these types of cases.

Subordinate clauses If the negation cue modifies a verb in a subordinate clause, the whole subordinate clause, except the initial subjunction, is part of the scope, see (8) below.

(8) *Det føles derfor som et pluss at*
It feels therefore like a plus that
[plata] ***ikke*** *[er særlig lang]* .
the.record not is especially long .
'It therefore feels like a bonus that the record is not especially long.'

Modifying subjects and objects If a cue, typically a determiner, modifies the subject or the object of a sentence, the whole clause that contains that subject or object is part of the scope, as in (9) below. Note that certain elements, such as subjunctions, conjunctions and sentence adverbs might still not be included.

(9) *[Her viser Selbekk]* ***ingen*** *[nåde]* .
Here shows Selbekk no mercy .
'Here, Selbekk shows no mercy.'

Cue as subject or object In cases where the subject or object are also neagtion cues, the cue is not included in the scope, see (10).

(10) *Og* ***ingen*** *[er hardere enn Regan]* .
And nobody is tougher than Reagan .
'And nobody is tougher than Reagan.'

Exception items The annotation of exception items, such as *untatt* 'except' and *bortsett (fra)* 'except (for)' depends on whether they are within the scope of a negation cue or not. When the item is not within the scope of another cue, it incurs a negation, as in (11). This closely follows the annotation found in Morante and Daelemans (2012) and Liu et al. (2018).

(11) *Sportsseter - som gir god støtte*
Sports-seats - which give good support
unntatt *[lårstøtten for høyvokste*
except the.thigh-support for high-grown
personer] .
people
'Sport seats - which give good support, except for the thigh support for tall people'

When exception items are found within the scope of another negation cue, however, they remove the elements they scope over from the scope of the other negation.

Sentential adverbs and adverbs scoping over negation Two types of adverbs pose certain challenges: sentential adverbs and adverbs that indicate modality. Sentential adverbs such as *heldigvis* 'fortunately' as in (12) are not part of the propositional value of a sentence, but rather function to comment on it (Faarlund et al., 1997). Therefore they are usually outside the scope of the negation, as is shown by (12):

(12) *Heldigvis [skjer dette] nesten* ***aldri*** .
Fortunately happens this almost never .
'Fortunately, this almost never happens.'

Modal adverbs such as *kanskje* 'maybe' can occur both within and outside of the scope of a negation cue, and in these cases the annotators were asked to paraphrase in order to pinpoint the placement of these adverbs.

Negation raising Negation raising is the phenomenon where a negator is "raised" further up in a syntactic tree, which in the case of Norwegian means further towards the beginning of a sentence. What characterizes these types of constructions is

that the negation is adjacent to the verb in the main sentence, even though the negation only scopes over a subsequent subordinate clause. This happens frequently in Norwegian, as in English, with mental state verbs like *mene* 'think', *tro* 'believe', as in (13).

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

Expletive subjects In Norwegian, as in other Scandinavian languages, there are several types of linguistic constructions that involve an expletive subject. A commonly used mechanism in these languages is extraposition, where a clausal argument is postposed, and a formal, semantically void subject *det* 'it' or *der* 'there' functions as the syntactic subject, as in (14). Here we do not treat the expletive subject as the subject of the negated proposition, instead only the extraposed subordinate clause is in scope of the negation. Since *det* 'it' is ambiguous in the sense that it can, in fact, also be referential, the annotators have to assess referentiality during annotation.

(14) *Det [er]* ***aldri*** *[kjedelig å se gode*
It is never boring to see good
replikker fremført i vakre
lines performed in beautiful
omgivelser] .
surroundings.
'It is never boring to see good lines performed in beautiful surroundings.'

Negation in conditional, interrogative, and imperative sentences In the annotation scheme of Morante and Daelemans (2012), they do not annotate negation in non-factual sentences, i.e., conditional, interrogative and imperative sentences. We have chosen to include all negation regardless of its factuality. We believe that negation has implications beyond asserting the factuality of a proposition, and it can be useful for sentiment analysis, among other tasks. For instance, in example (15), the negation is under the scope of the conditional *hvis* 'if', but is still marked, even though it is not a factual proposition.

(15) *Hvis [folk]* ***ikke*** *[hadde snakket til*
if people not had talked to
hverandre i det hele tatt] [...]
each.other in the whole taken [...]
'If people had not talked to each other at all [...]'

Negative polarity items (NPIs) NPIs are lexical entities that are used together with negation cues, and which usually render the sentence ungrammatical should the negation cues be removed without further change. In our annotation scheme, they are contained within the scope of the negation cue. In Norwegian, the negative adverb *ikke* 'not' in combination with the determiner *noe/noen* 'some/any' is a common negative polarity item. However, the most common type of NPIs are adverbs such as *i det hele tatt* 'at all', as in (16), that serve to strengthen the negation.

(16) *[Han kan]* ***ikke*** *[synge i det hele*
He can not sing in the whole
tatt] .
taken .
'He cannot sing at all.'

Foreign language citations The annotated texts frequently contain titles of various products, such as 'Never Run Away'. These cases of foreign language negation cues are not annotated.

Negation cues not indicating negation It is not uncommon for negation cues to be part of expressions that do not indicate negation in combination, e.g., certain fixed expressions such as *hvis ikke* 'otherwise'. Other borderline cases such as the focus marker *ikke bare* 'not only' and the expression *ingen tvil* 'no doubt', were included after discussion, as they are analyzed as introducing a negated reading.

Affixal scope The scope of affixal items is annotated in a slightly different way compared to other cues. If an affixally negated adjective is the predicate, then the whole sentence is included within its scope. If it is part of a noun phrase, then only that noun phrase is inside the scope. Additional adjectives or adverbs in the sentence fall outside the scope, as in (17).

(17) *Passasjerene er for oss* ***u****[kjente]* ,
the.passengers are for us unknown ,
anonyme [fjes] .
anonymous faces .
'The passengers are unknown faces to us.'

3.5 Annotation Procedure

The annotation was performed by several hired student research assistants with a background in linguistics and with Norwegian as their native language. All 414 documents in the original dataset,

comprising 11,346 sentences, were annotated independently by two annotators in parallel. The doubly annotated documents were then adjudicated by a third annotator after a final round of discussions concerning difficult cases. Annotators had the possibility to discuss any potential problems during both the annotation and adjudication period, but were encouraged to follow the guidelines as strictly as possible. The annotation and adjudication were both performed using the web-based annotation tool Brat (Stenetorp et al., 2012).

3.6 Inter-annotator agreement

We have measured the inter-annotator agreement over the full (doubly annotated) dataset in terms of both F_1 and κ scores for cues, full scopes, and scope tokens. The scores show that annotators agree to a very high degree on the identification of cues (0.995 F_1, 0.841 κ). When it comes to negation scopes, the agreement is lower when measured towards full and exact spans (0.632 F_1, 0.34 κ), but quite high when measured on the token-level (0.912 F_1, 0.803 κ).

Due to the adjudication phase of the annotation process, we also have insight into the sources of disagreements between the annotators. As noted above, agreement between annotators is generally high when it comes to cue detection, but surprising disagreements can be seen. These are most likely due to the guidelines being improved as the annotations continued to uncover new challenges. There seems to be a clear tendency for annotators to disagree on less common cues, such as verbs and nouns that indicate negation, as opposed to the more often discussed adverbs and determiners. The annotators rarely agreed on less frequent lexical items such as *forsvinne* 'disappear' and *takke nei til* 'say no to'. However, the disagreements also reflect discussions concerning the inclusion or omission of prepositions, in addition to cue span errors. Annotators generally agree on the more frequent cues. The prefix *u-* 'un-/dis-/non-', seems to have a disproportionately large disagreement score, but discussions among the annotators indicate that this is likely due to prefixes being more difficult to detect when annotating than isolated whole-word tokens. Disagreement is also found regarding modal elements, such as *knapt* 'barely' (almost not) and *for...til* 'too...to' (cannot be).

4 Corpus statistics

Table 2 summarizes the statistics for the final annotated data set. Of the 11,346 sentences in the corpus, we see that just above 20% of them are negated. Out of the negated sentences, 13% contain multiple instances of negation. While, as expected, the number of tokens in a *cue* averages to 1, the average length of *scopes* is close to 7 (with a maximum observed length of 53). Note, however, that a small number of cues (1.4%) also have empty ('null') scopes. We report both any kind of discontinuous scopes (disc.) and true discontinuous scopes (true disc.), where the latter does not count scopes which are only discontinuous because of an intervening cue. While discontinuous scopes are very frequent (70% of scopes), truly discontinuous scopes are much fewer (21%). We see that affixal negation is quite widespread in NoReC$_{neg}$, comprising almost 25% of the cues. Moreover, just above 11% are multi-word cues. While most cues are not particularly ambiguous, e.g., *ikke* 'not', *uten* 'without', others, such as *u-* 'un-/dis-', *mangle* 'lack' or *verken* 'neither' can have rather high rates of ambiguity (meaning that they can occur with both negated and non-negated readings).

5 Experiments

5.1 Modeling approach

In order to benchmark the dataset, we use the semantic graph parsing approach to negation detection proposed by Kurtz et al. (2020), see Section 2. Besides the baseline graph representation originally proposed (***point-to-root***), where all elements of the scope have arcs that point to the cue, we propose several variants. For ***head-first***, we set the first token of the cue as the root node, and similarly set the first token in the scope as the head of the span. All elements within the span have arcs that point to the head, and heads have arcs that point to the root. ***head-final*** is similar, but instead sets the final tokens of spans as the heads. There can be several roots per sequence and not all tokens are connected. Finally, we enrich the dependency labels to distinguish edges that are internal to a holder/target/expression span from those that are external and perform experiments by adding an 'in label' to non-head nodes within the graph, which we call ***+inlabel***.

	Sentences		Cues						Scopes					
	#	neg.	#	avg.	max	disc.	mult.	affixal	#	avg.	max	disc.	true disc.	null
train	8,543	1,768	2,025	1	3	19	228	508	1,995	6.9	44	1,403	423	30
dev	1,531	301	342	1	2	0	39	88	339	7.1	53	236	85	3
test	1,272	263	305	1	2	2	37	69	301	6.5	27	203	58	4
total	11,346	2,332	2,672	1	3	21	304	665	2,635	6.9	53	1,842	566	37

Table 2: Statistics of the dataset – per split and in total – including total number of sentences (#), number of sentences that contain negation (neg.), as well as the number (#) of cues and scopes, along with their average and maximum lengths in tokens. Additionally, we include the number of discontinuous cues and scopes (disc.) as well as true discontinuous (true disc.) for scopes which we discuss in Section 4. Finally, we detail the number of sentences that have multiple cues (mult.), the number of affixal cues, and the number of cues that have no scope (null).

5.2 Results

The negation parser is evaluated using the metrics from the *SEM 2012 shared task (Morante and Blanco, 2012): cue-level F_1 (CUE), scope token F_1 over individual tokens (ST), and the full negation F_1 (FN) metric. In contrast to the *SEM 2012 shared task we do not annotate negated events, meaning that FN only requires an exact match of the negation's cue(s) and, if present, all its scope tokens. We run each experiment five times with different random seeds and report an averaged F_1 score and its standard deviation in Table 3.

The simplest graph representation *point-to-root* generally performs best, most visibly in FN F_1 (66.8). We attribute the variation in performance to a loss of information in the *head-first* and *head-final* variants, making it impossible to retrieve the correct governing negation cue for partially overlapping scopes, thus lowering the score.

In order to see whether these performance differences are statistically significant, we perform bootstrap significance testing (Berg-Kirkpatrick et al., 2012) resampling the test set 10^6 times while setting the significance threshold to $p = 0.05$. Comparing *point-to-root* to *head-first* and *head-final* shows that while the differences seem substantial they are not statistically significant.

A manual error analysis on *point-to-root* shows that the model tends not to predict infrequent cues, e.g., *null* 'zero', *istedenfor* 'instead-of', *savnet* 'missing', while it overpredicts frequent cues, e.g., *ikke* 'not', *ingen* 'no', as well as overgeneralizing the affixal negation *u-* 'un-/dis-/non-' to other words that begin with 'u', but are not negated, e.g., *utfrika* 'freaked-out', *unnagjort* 'finished'. The model also tends to predict slightly shorter scopes (an average of 6.5 tokens for predicted scopes versus 6.7 for gold scopes), while the most common scope-related errors derive from discontinuous scopes, where the model fails on 75.4%. These errors are often due to inversions with the expletive 'det', which is not considered in scope. Although rare (4 examples in test), multi-word cues are also challenging, and the graph model only correctly predicted one of the four. Finally, affixal cues can pose a challenge as well, with the model failing on 67.1% of the sentences containing affixal negation.

	CUE	ST	FN
point-to-root	93.4 (0.5)	83.6 (0.7)	66.8 (0.8)
head-first	92.7 (0.3)	81.9 (1.4)	65.5 (0.6)
+inlabel	92.7 (0.7)	81.8 (1.0)	65.0 (2.2)
head-final	92.7 (0.6)	82.7 (1.8)	64.8 (3.1)
+inlabel	93.1 (0.3)	82.2 (1.5)	65.8 (0.8)

Table 3: Results of our negation parser using the various graph representations. The results are averaged over 5 runs, additionally reporting standard deviation.

6 Future work

As mentioned previously, the underlying corpus NoReC$_{fine}$ is annotated for fine-grained sentiment, including opinion holders, targets, sentiment expressions, and positive/negative polarity. The fact that negation is among the most important compositional phenomena that can affect sentiment in terms of shifting polarity values motivated the choice of this particular dataset for adding the negation annotations. In future work we plan to

further investigate the co-dependencies between negation and sentiment, both through analyzing the existing annotations and through joint modeling.

7 Summary

This paper has introduced the first annotated dataset of negation for Norwegian, NoReC$_{neg}$, where negation cues and their corresponding in-sentence scopes have been annotated across more than 11K sentences spanning more than 400 documents; a subset of the Norwegian Review Corpus (NoReC). In addition to providing in-depth discussion of the annotation guidelines, we have also presented a first set of benchmark results based on a graph-parsing approach.

Acknowledgements

This work has been carried out as part of the SANT project (Sentiment Analysis for Norwegian Text), funded by the Research Council of Norway (grant number 270908). We also want to express our gratitude to the annotators: Anders Næss Evensen, Helen Ørn Gjerdrum, Petter Mæhlum, Lilja Charlotte Storset, Carina Thanh-Tam Truong, and Alexandra Wittemann.

References

Zubair Afzal, Ewoud Pons, Ning Kang, Miriam CJM Sturkenboom, Martijn J Schuemie, and Jan A Kors. 2014. ContextD: an algorithm to identify contextual properties of medical terms in a Dutch clinical corpus. *BMC bioinformatics*, 15(1):1–12.

Begoña Altuna, Anne-Lyse Minard, and Manuela Speranza. 2017. The scope and focus of negation: A complete annotation framework for Italian. In *Proceedings of the Workshop Computational Semantics Beyond Events and Roles*, pages 34–42, Valencia, Spain.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An Empirical Investigation of Statistical Significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea.

Viviana Cotik, Roland Roller, Feiyu Xu, Hans Uszkoreit, Klemens Budde, and Danilo Schmidt. 2016. Negation detection in clinical reports written in german. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)*, pages 115–124.

Isaac Councill, Ryan McDonald, and Leonid Velikovich. 2010. What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 51–59, Uppsala, Sweden.

Hercules Dalianis and Sumithra Velupillai. 2010. How Certain are Clinical Assessments? Annotating Swedish Clinical Text for (Un) certainties, Speculations and Negations. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.

Noa P Cruz Diaz, Roser Morante, Manuel J Mana López, Jacinto Mata Vázquez, and Carlos L Parra Calderón. 2017. Annotating negation in Spanish clinical texts. In *Proceedings of the workshop computational semantics beyond events and roles*, pages 53–58, Valencia, Spain.

Timothy Dozat 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)*, pages 484–490, Melbourne, Australia.

Martine Enger, Erik Velldal, and Lilja Øvrelid. 2017. An open-source tool for negation detection: a maximum-margin approach. In *Proceedings of the EACL workshop on Computational Semantics Beyond Events and Roles (SemBEaR)*, pages 64–69, Valencia, Spain.

Jan Terje Faarlund, Svein Lie, and Kjell Ivar Vannebo. 1997. *Norsk referansegrammatikk*. Universitetsforlaget, Oslo, Norway.

Federico Fancellu, 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*, pages 495–504, Berlin, Germany.

Federico Fancellu, Adam Lopez, Bonnie Webber, and Hangfeng He. 2017. Detecting negation scope is easy, except when it isn't. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 58–63, Valencia, Spain.

Anne Golden, Kirsti Mac Donald, and Else Ryen. 2014. *Norsk som fremmedspråk: Grammatikk*, 4 edition. Universitetsforlaget, Oslo, Norway.

Salud María Jiménez-Zafra, Roser Morante, María Teresa Martín-Valdivia, and L Alfonso Ureña-López. 2020. Corpora annotated with negation: An overview. *Computational Linguistics*, 46(1):1–52.

Salud María Jiménez-Zafra, Mariona Taulé, M Teresa Martín-Valdivia, L Alfonso Urena-López, and M Antónia Martí. 2018. SFU Review SP-NEG: a Spanish corpus annotated with negation for sentiment analysis. a typology of negation patterns. *Language Resources and Evaluation*, 52(2):533–569.

Natalia Konstantinova, 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 8th International Conference on Language Resources and Evaluation*, pages 3190–3195, Istanbul, Turkey.

Robin Kurtz, 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*, pages 14–24, Online.

Emanuele Lapponi, Erik Velldal, Lilja Øvrelid, and Jonathon Read. 2012. UiO2: Sequence-labeling negation using dependency features. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 319–327, Montreal, Canada.

Qianchu Liu, Federico Fancellu, and Bonnie Webber. 2018. NegPar: A parallel corpus annotated for negation. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 3464–3472, Miyazaki, Japan.

Suguru Matsuyoshi, Ryo Otsuki, and Fumiyo Fukumoto. 2014. Annotating the Focus of Negation in Japanese Text. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 1743–1750, Reykjavik, Iceland.

Roser Morante and Eduardo Blanco. 2012. *SEM 2012 shared task: Resolving the scope and focus of negation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 265–274, Montréal, Canada.

Roser Morante and Walter Daelemans. 2009. A metalearning approach to processing the scope of negation. In *Proceedings of the 13th Conference on Computational Natural Language Learning*, Boulder, USA.

Roser Morante and Walter Daelemans. 2012. ConanDoyle-neg: Annotation of negation cues and their scope in Conan Doyle stories. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey.

Lilja Øvrelid, Petter Mæhlum, Jeremy Barnes, and Erik Velldal. 2020. A fine-grained sentiment dataset for Norwegian. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5025–5033, Marseille, France.

Woodley Packard, 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*, Baltimore, USA.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2007. Bioinfer: a corpus for information extraction in the biomedical domain. *BMC bioinformatics*, 8(1):1–24.

Z. Qian, P. Li, Q. Zhu, G. Zhou, Z. Luo, and W. Luo. 2016. Speculation and negation scope detection via convolutional neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, USA.

Jonathon Read, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2012. UiO1: Constituent-based discriminative ranking for negation resolution. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, Montreal, Canada.

Maria Skeppstedt. 2011. Negation detection in Swedish clinical text: An adaption of NegEx to Swedish. *Journal of Biomedical Semantics*, 2 Suppl 3.

Pontus Stenetorp, 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*, pages 102–107, Avignon, France.

Erik Velldal, Lilja Øvrelid, Eivind Alexander Bergem, Cathrine Stadsnes, Samia Touileb, and Fredrik Jørgensen. 2018. NoReC: The Norwegian Review Corpus. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference*, pages 4186–4191, Miyazaki, Japan.

V. Vincze, G. Szarvas, R. Farkas, G. Móra, and J. Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, Suppl 11.

James Paul White. 2012. UWashington: Negation resolution using machine learning methods. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, Montreal, Canada.

Bowei Zou, Guodong Zhou, and Qiaoming Zhu. 2016. Research on chinese negation and speculation: corpus annotation and identification. *Frontiers of Computer Science*, 10(6):1039–1051.