Multilingual Negation Scope Resolution for Clinical Text

Mareike Hartmann
Department of Computer Science
University of Copenhagen
Denmark
hartmann@di.ku.dk

Anders Søgaard
Department of Computer Science
University of Copenhagen
Denmark
soegaard@di.ku.dk

Abstract

Negation scope resolution is key to high-quality information extraction from clinical texts, but so far, efforts to make encoders used for information extraction negation-aware have been limited to English. We present a universal approach to multilingual negation scope resolution, that overcomes the lack of training data by relying on disparate resources in different languages and domains. We evaluate two approaches to learn from these resources, training on combined data and training in a multi-task learning setup. Our experiments show that zero-shot scope resolution in clinical text is possible, and that combining available resources can improve performance in most cases.

1 Introduction

Information extraction (IE) from clinical text, such as electronic health records (EHR) or clinical trial narratives, is a promising application of machine learning that can potentially benefit many areas of the health sector (Dalianis, 2018). IE systems are applied to facilitate administrative tasks by assigning medical codes (Stanfill et al., 2010), to extract phenotype information about patients (Gehrmann et al., 2018), and to improve patient care by monitoring healthcare associated infections (Proux et al., 2011) and adverse drug events (ADE) (Luo et al., 2017). Results are promising, but the majority of work focuses on clinical text in English (Névéol et al., 2018). This puts at a disadvantage patients in countries where health narratives are recorded in languages other than English. This gap can be overcome by multilingual IE systems that are applicable to text in multiple languages. Such multilingual systems can, on the one hand, lead towards improving healthcare within non-English speaking countries. On the other hand, they allow to gather information across countries, which is particularly interesting for rare diseases with few cases per country, and can increase the statistical power of an analysis (Jensen et al., 2012; Névéol et al., 2018). With this paper, we contribute to improving IE from clinical text in languages other than English, which is a step towards improving healthcare for all.

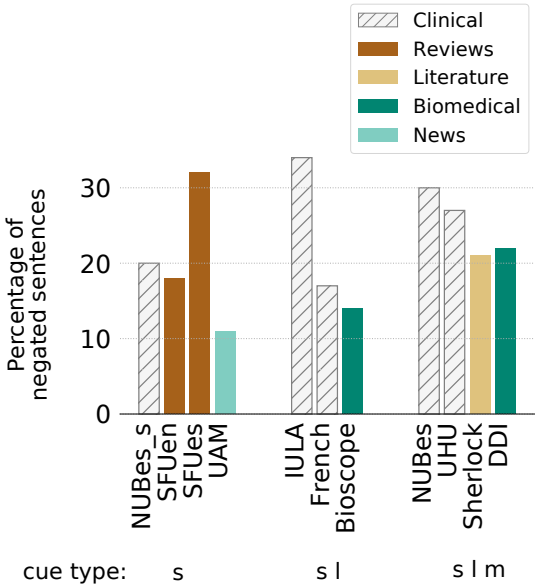

Figure 1: Percentage of negated sentences in clinical and non-clinical text. Datasets are grouped according to the type of negation cue they annotate (syntactic (s), lexical (l), morphological (m), see further details in). Numbers for datasets not listed in Section 4 (UAM (Sandoval and Salazar, 2013) and UHU (Diaz et al., 2017)) are taken from (Jiménez-Zafra et al., 2020). NUBes_s counts only syntactic cues in NUBES.

Negation is a phenomenon that has received considerable attention in IE models for clinical and biomedical text, as there is large interest in identifying negated concepts (Mutalik et al., 2001; Chapman et al., 2001; Mowery et al., 2012), e.g. negated medical events (Nawaz et al., 2013) or negated drug-drug interactions (Bokharaeian et al., 2016). Absence of symptoms or the fact that chemical re-

Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis, pages 7–18
April 19, 2021.

actions are not observable are crucial knowledge in the clinical and biomedical domain (Elkin et al., 2005; Krallinger, 2010). This is reflected in Figure 1, which shows the percentage of negated sentences in clinical compared to non-clinical text. Explicitly integrating negation information into machine learning (ML) models can improve performance for relation extraction (Chowdhury and Lavelli, 2013) and more general NLP tasks such as sentiment analysis (Barnes et al., 2020) and machine translation (Fancellu and Webber, 2014), which become increasingly popular for the clinical domain (Denecke and Deng, 2015). Even though pre-trained language models tailored to the biomedical and clinical domains (Lee et al., 2020; Peng et al., 2019; Alsentzer et al., 2019) now produce state-of-the-art results for several downstream tasks, such language models do not sufficiently capture the semantics of negation (Ettinger, 2020; Kassner and Schütze, 2020).

One way of accessing negation information is to explicitly detect all negated words in a sentence, a task which is referred to as *negation scope resolution* (Morante et al., 2008). The task has recently been successfully approached by fine-tuning a pre-trained language model on labeled target data (Sergeeva et al., 2019; Khandelwal and Sawant, 2020). For the clinical domain however, we cannot rely on multilingual labeled target data to be available. This is partly due to data privacy issues, which lead to few publicly available datasets in the clinical domain (Chapman et al., 2011; Velupillai et al., 2018) [1], and partly due to the fact that other languages are underrepresented compared to English in clinical NLP (Névéol et al., 2018).[2] In summary, when building a multilingual negation scope resolution system for clinical text, we are facing a lack of training data. Our approach is hence to use the available data as best as possible to build a negation resolution model that works on data in multiple languages.

Negation scope resolution has also been considered for non-clinical text (Morante and Sporleder, 2012; Morante and Blanco, 2012), and several datasets spanning various domains and including a small amount of non-English languages are available (see Section 4). In this work, we investigate if and how these disparate data sources can serve as training resource to resolve negation scope in multiple languages in the clinical domain in a zero-shot setup. To enable transfer across languages we rely on multilingual BERT (mBERT) (Devlin et al., 2019), a multilingual pre-trained language encoder that has proven capable of zero-shot cross-lingual transfer in other tasks[3] (Wu and Dredze, 2019), and has recently been applied for zero-shot transfer to French clinical text (Shaitarova et al., 2020). One challenge arising from the available scope resolution datasets is that they are annotated according to different annotation schemes (see Section 3.3), raising the question if and how they can be combined as a training resource (Barnes et al., 2020; Jiménez-Zafra et al., 2020).

We explore two strategies to handle disparities between the resources, a simple concatenation of datasets after a partial conversion of annotations, and a multi-task learning (MTL) setup, where each dataset is handled as a different task. The MTL setup also allows us to explore additional auxiliary tasks that can potentially help in resolving negation scope by making use of available resources. Here, we consider a classification task for negated events in the clinical domain.

Contributions We study an approach for zero-shot cross-lingual transfer for negation scope resolution in clinical text, exploiting data from disparate sources by data concatenation, or in an MTL setup.[4] We demonstrate that it is possible to achieve decent performance across Spanish and French clinical texts without scope annotated training data in the target language, or in the target domain.[5] We further explore an auxiliary task that makes use of negated medical event detection data, however finding that they cannot improve performance.

2 Related Work

Recent work used large pre-trained language models, in particular BERT, for negation scope resolution as a sequence labeling task (Sergeeva et al., 2019; Khandelwal and Sawant, 2020) and produced

[1] For example, there are several clinical datasets of EHRs annotated with negation, but they are not publicly available (see Chapter 4.7.3 in Dalianis (2018)).

[2] This trend can also be observed in NLP datasets outside of the clinical domain (Bender, 2011). Also in domains other than the clinical domain, the range of languages in publicly available datasets for negation scope resolution is small (Jiménez-Zafra et al., 2020).

[3] https://sites.research.google/xtreme

[4] We make our code and trained models available at https://github.com/coastalcph/multi_neg_scope

[5] We however rely on a small annotated validation dataset in target language and domain for choosing the best model.

(1) El ojo derecho **no** completa <u>la aducción</u> (procedure) .

(2) **No** existen <u>datos de focalidad neurológica</u> (medical finding) **ni** <u>signos de meningismo</u> (medical finding).

(3) Es **incapaz de** levantarse de la silla **sin** <u>ayuda</u>

Table 1: Example sentences from the IULA corpus. Negation cues are marked in bold, negation scopes are underlined and the overbrace indicates additional annotations of medical concepts.

new state-of-the-art results on the BIOSCOPE and SHERLOCK datasets. Barnes et al. (2020) deviate from the two-step sequence labeling approach and propose a sequence labeling model that detects cues and scope in one step. Kurtz et al. (2020) show that it is beneficial to frame the negation scope resolution task as a graph parsing problem. All the works listed above focus exclusively on English data, and only few works attempt to do multi- or cross-lingual negation scope resolution.

Fancellu et al. (2018) were the first to present a zero-shot approach for negation scope resolution. They find that transfer from English to a Chinese version of the SHERLOCK corpus is possible, using an LSTM and a graph convolutional network (GCN) in combination with static cross-lingual word embeddings. However, both their models rely on PoS-tags and more importantly on dependency parses of the input sentences.[6]

Shaitarova et al. (2020) present zero and few-shot experiments for the FRENCH clinical data and Spanish review data using mBERT fine-tuned on a concatenation of English datasets. Their model is equivalent to our single task model trained on concatenated data ($\mathbf{ST}_{\text{cat}}$ in Section 6). While their work demonstrates the general applicability of the approach, we focus on the applicability to clinical text in multiple languages, by exploring more training resources, and more importantly by evaluating the generalization performance of the approach across three clinical datasets in two languages.

3 Negation Scope Resolution

Negation is a phenomenon in language that changes the truth value of a proposition. From a linguistic perspective, a negation signal or cue, i.e. an expression that indicates negation, is an operator, which has a scope over parts of a sentence.[7]

In example sentence (1) in Table 1 above, the negation cue **no** (denoted in bold) affects the underlined parts of the sentence, which are referred to as the scope of this negation cue. In many clinical IE tasks, the goal is to resolve the negation of medical concepts or entities, i.e. to determine if they are present or absent, for example the procedure of eye movement in (1) and the medical finding of focal signs and signs of meningism in (2). This can be approached as a binary classification task predicting if the medical entity is negated or not (Chapman et al., 2001). Depending on the application, however, one might want to resolve negation of a different set of medical concepts (such as relations), in which case the classifier has to be re-trained on the new set of concepts. Morante et al. (2008) suggest to first find the whole scope of the negation cue [8], and then check if the concepts of interest are contained in the scope or not. With this setup, a system can be re-used across concepts, and does not have to be re-trained for a different downstream task. Also, as negation is inherent in text from any domain, this approach opens the possibility to learn to resolve negation scope in other domains, and transfer this knowledge to clinical text.

3.1 Resolving Negation Scope

The negation scope resolution task is usually solved in two steps: First, negation cues are identified, often using a lexicon. Second, the scope of these negation cues, i.e. the words that are affected by the negation cue, are identified. The second step is approached as a sequence labeling task, where given a negation cue, each word in the sequence is labeled with respect to whether it is affected by this cue (in scope) or not (out of scope). In the second

[6] Obtaining dependency parses for text from several domains and in several languages requires expensive adaptation of parsers, and is not applicable for our setup. Hence, Fancellu et al. (2018) is not included as a baseline in our experiments.

[7] In this work, we only consider intra-sentential negation, i.e. negation that affects words within the sentence containing the negation cue.

[8] This setup is now commonly referred to as negation scope resolution (Morante and Blanco, 2012).

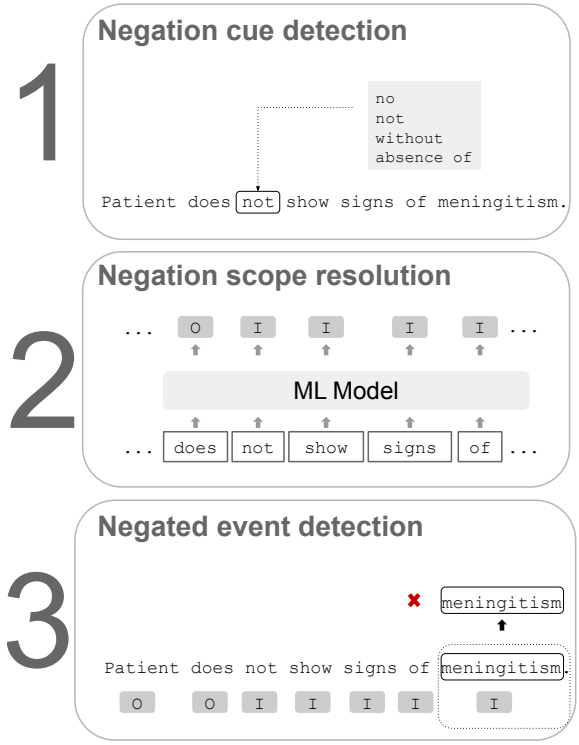

Figure 2: Pipeline for resolving negation in clinical data. First, negation cues are identified using a list of pre-defined negation cues. Second, the scope of the detected negation cue is identified using a machine learning model, which assigns a binary scope label (either `I` for in- or `O` for out-of-scope) to each token in the sequence. Finally, the predicted negation scope can be used to identify negated events. Our work focuses on the second step. As ML model, we use an MTL architecture with a shared pre-trained multilingual encoder.

step, information about the cue is provided as input to the model, and the model handles one cue per sequence, i.e. sentences with multiple negation cues are represented using multiple input sequences with one marked cue each.[9] Figure 2 shows the individual steps of this pipeline approach.

3.2 Cue Detection

Most recent work, including ours, focuses on the second step of the pipeline and resolves negation scope given gold cues, arguing that negation cues in a practical setting can best be identified using a

[9]Barnes et al. (2020) propose a model that does cue detection and scope resolution in one pass, by adding a special cue label for cues in the scope resolution labeling task. Here, the scope is not conditioned on a specific cue, which means with several cues in a sentence, we do not know which cue is associated with a predicted scope. While this setup seems more practical to use with a downstream task in mind, the two stage setup allows to have a dedicated language specific cue detection component, which is particular interesting in a cross-lingual zero-shot setup.

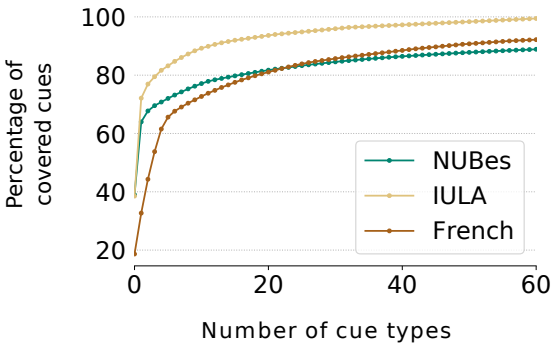

Figure 3: Distribution of negation cues in the datasets. A small number of different cues is responsible for a large amount of negations.

lexicon-based approach. We think this is a reasonable approach, as in the annotation process for many datasets, cue annotation involves a predefined lexicon. Figure 3 shows that a fixed set of negation cues accounts for the majority of negations in the data. Following Jiménez-Zafra et al. (2020), we discriminate between three types of negation cues. Syntactic negation cues such as **not** and **without** form the largest group of cues. Lexical negation cues are words with a meaning that indicates a negation, such as **lack of**. Morphological negation cues are words that contain a morpheme expressing the negation, such as prefix **a-** in **a**symptomatic. We do not include unsupervised experiments for cue detection, as the available datasets annotate different types of cues (see Table 2).

3.3 Difference in Annotation Schemes

In addition to annotating different types of negation cues, the available datasets also follow different annotation schemes for elements to include in the scopes of these cues, e.g. if subjects, cue tokens and punctuation are included in the scope or not. In Table 2, we list annotation features of the datasets involved. In their review on corpora annotated with negation, Jiménez-Zafra et al. (2020) emphasize these differences in annotation and state that *as a consequence, it would not be possible to merge all of them to training a negation processing system.* Problems with differences in annotation are also noted by Barnes et al. (2020), who find that concatenating datasets for training does not lead to improved results. We propose two steps to alleviate this problem: If possible, we convert different annotations to a similar scheme. In particular, we

	Statistics			Annotation					
Dataset	**Domain**	**Lang**	**# Train**	**Type**			**Subj**	**Cue**	**Punc**
IULA	clinical records	es	771	l	s		∼	×	✓
NUBES	clinical records	es	5,297	l	s	m	∼	×	×
FRENCH	clinical records	fr	1,272	l	s		✓	×	×
BIOSCOPE	biomedical publications	en	1,469	l	s		×	✓	×
SHERLOCK	short stories	en	618	l	s	m	✓	×	×
SFUES	reviews	es	2,796		s		∼	✓	×
SFUEN	reviews	en	2,458		s		×	×	×
DDI	biomedical database	en	892	l	s	m	?	✓	✓

Table 2: Dataset statistics and difference in annotation schemes. The test split sizes of the IULA, NUBES, and FRENCH datasets are 173, 1152 , and 272 sentences, respectively. Negation types are either lexical (l), syntactic (s), or morphological (m). ∼ signifies that the subject is included in the scope only in specific cases. Features in the last two columns can easily be converted to a standardized annotation (include cue token, exclude punctuation). Negation type and inclusion/exclusion of subject cannot be accounted for without extra effort.

always include cue tokens in the scope, and exclude punctuation if the punctuation marker denotes the end of a scope. The other differences in annotation, in particular inclusion of subjects and differences in annotated cue type, cannot be equalized easily. Here, we aim at still learning as much as possible from the available data with different annotation schemes by using an MTL model with output layers specific for each dataset, and hence tailored to each annotation scheme.

4 Datasets

In the following, we describe the clinical datasets as well as the training resources used in our experiments. A more detailed overview of annotation guidelines, negation types and negation components can be found in the survey of Jiménez-Zafra et al. (2020), as well as in the original works associated with the datasets. Statistics on the available data can be found in Table 2.

4.1 Clinical Negation Scope Resolution Datasets

IULA The IULA corpus (Marimon et al., 2017) is a collection of Spanish clinical records from *several services of one of the main hospitals in Barcelona (Spain)* (Marimon et al., 2017) and includes text from five sections of the electronic record: physical exploration, evolution, radiology, current process, and comparative explorations. Here, syntactic and lexical negation cues are annotated. Subjects are (almost) always excluded from the scope if they precede the verb. The corpus also contains annotations for four types of medical entities: body structure, substance, clinical finding, and procedure.

NUBes The NUBES corpus (Lopez et al., 2020) is a collection of anonymized Spanish clinical records from a private hospital. The text is extracted from seven sections present in the electronic record: chief complaint, present illness, physical examination, diagnostic tests, surgical history, progress notes, therapeutic recommendations. Here, syntactic, lexical, and morphological negation cues are annotated.

French The FRENCH corpus (Dalloux et al., 2019) is a collection of clinical trial protocols in French which were obtained from the registry of the Gustave Roussy hospital as well as the French National Cancer Institute. The corpus includes text parts about the patient inclusion criteria and the description of the procedure of the trials. Here, syntactic and lexical negation cues are annotated. In this corpus, there is no marked association between a negation cue and its scope, hence sentences with different negation cues have to be processed at once.

4.2 Other Negation Scope Resolution Datasets

Bioscope The BIOSCOPE corpus (Vincze et al., 2008) is a collection of English biomedical and clinical texts and consists of three parts: abstracts

and full papers of biomedical publications, and radiology reports. Unfortunately, we were not able to obtain access to the radiology reports, hence in our experiments we only include the biomedical publications (abstracts and full papers).

Sherlock The SHERLOCK corpus (Morante and Daelemans, 2012) comprises two English Conan Doyle short stories. It was used in the 2012 Shared Task on negation scope resolution and is the most popular benchmark corpus. Here, syntactic, lexical and morphological negation cues are annotated.

DDI The DDI corpus (Bokharaeian et al., 2014) is a collection of texts from the chemical and pharmaceutical database DrugBank, and English Medline abstracts.

SFU$_{en}$ The SFUEN corpus (Konstantinova et al., 2012) is a collection of product reviews in English. The reviews are extracted from the reviewing website *Epinions.com* and cover 8 different categories[10].

SFU$_{es}$ The SFUES corpus (Jiménez-Zafra et al., 2018) is a collection of Spanish reviews from the reviewing website *Ciao.es* also covering 8 different categories.

4.3 Auxiliary Task

One obvious choice of auxiliary task is the detection of negated events, which is a sequence classification task, where the event is gold annotated and replaced by a special token.[11] We use the instances labeled as *present* and *absent* in the English M2C2 assertion dataset (Uzuner et al., 2011) in order to generate a binary sentence classification task.

5 Approach

MTL Setup The easiest way to join the disparate training resources annotated for negation scope resolution is to simply concatenate them (Barnes et al., 2020; Shaitarova et al., 2020). However, the difference between annotations in the negation scope resolution datasets (see Section 3.3) makes it reasonable to use an MTL setup that aims at learning several tasks at once in order to better solve a target task. Here, we treat each training dataset as a separate task. The MTL setup also allows to exploit negation information in datasets associated with tasks other than negation scope resolution, e.g. negated event detection. We use an MTL model with hard parameter sharing (Caruana, 1997; Collobert et al., 2011), i.e. all model parameters except for the task specific output layers are shared between the different tasks. The model is shown in Figure 4. Using multilingual BERT as shared multilingual encoder enables the model to learn from datasets in different languages, and to do zero-shot cross-lingual transfer at inference time. In particular, we run our experiments using the MT-DNN framework[12] (Liu et al., 2019).

Sampling During training, we sample batches from a task i according to $p_i \propto N_i$, where N_i is the size of the dataset associated with task i.We implement annealed sampling following Stickland and Murray (2019) as sampling according to $p_i \propto N_i^{\alpha}$, with $\alpha = 1 - f\frac{e-1}{E-1}$. Here, e refers to the current epoch, E is the maximum number of training epochs and f a fixed annealing value.[13] The idea is that α decreases while training progresses, and hence the difference in dataset size becomes less prevalent later in the training. In the last epoch, batches are sampled from each task with equal probability, regardless of dataset size. For all experiments, we fix the number of total batches sampled in each epoch to the number of distinct batches over all datasets. We experiment with two variants for sampling training instances. We either only train on sentences that contain a negation, or we also include sentences that do not contain any negation. In the latter setting, we downsample the amount of sentences without negation to be equal to the number of negated sentence (which means that in this setting, the models are trained on twice as much data as in the negated-only setting). Both sampling variants are treated as hyperparameters and we select the best setting on the validation data.

Zero-shot Setup We assume no training data for the clinical datasets and test in a zero-shot setup, after converting all annotations according to the steps described in Section 3.3. During training, we use per-token F1-scores on the development split of the BIOSCOPE corpus for model selection. To identify the combination of training datasets that are expected to perform best on the target datasets,

[10] books, cars, computers, cookware, hotels, films, music, and phones

[11] Difference and similarities between the tasks of negated event detection and negation scope resolution are discussed in (Stenetorp et al., 2012)

[12] https://github.com/namisan/mt-dnn

[13] Following Stickland and Murray (2019) we set $f = 0.8$. We set E to the number of training epochs.

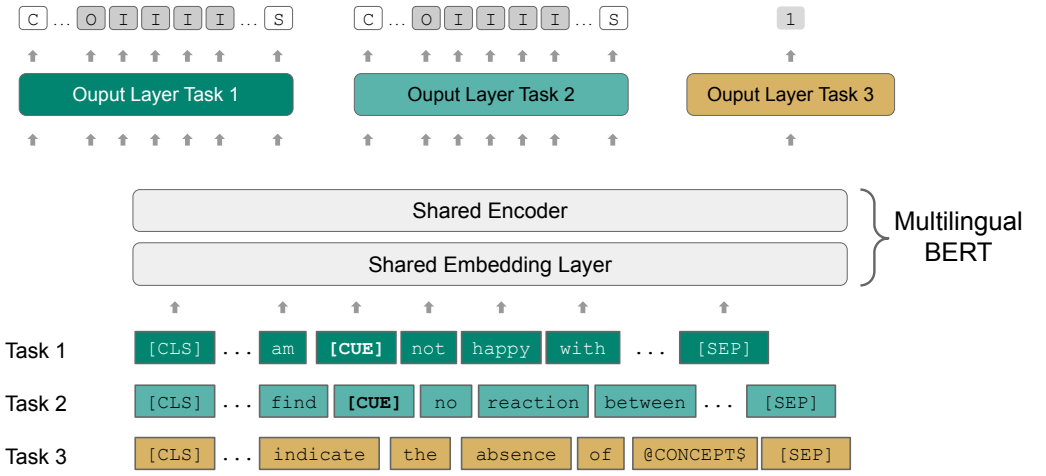

Figure 4: MTL model with a shared multilingual encoder for multilingual negation scope resolution. The multilingual embedding and encoder layers are shared between all the tasks, whereas each task has a task-specific output layer. Task 1 and 2 are sequence labeling tasks for negation scope resolution on product reviews and biomedical text, respectively. Task 3 is a binary sequence classification task on the M2C2 dataset. For the negation scope resolution tasks, information about the cue is added to the sequence in the form of a special token preceding the cue token.

we use F1-scores averaged over the development splits of the three clinical datasets. This reflects the use-case of a user annotating a small set of target sentences, in order to choose among the best pre-trained negation scope resolvers that can be readily applied to the specific target dataset.

5.1 Pretrained Multilingual Encoder

As pre-trained multilingual encoder, we use multilingual BERT, a transformer-based language model pre-trained with a masked language modeling and next sentence prediction objective on Wikipedia text in multiple languages.[14] The encoder can easily be adapted to downstream classification tasks by putting a randomly initialized classification layer on top of the pre-trained encoder.[15] mBERT has shown cross-lingual transfer abilities when fine-tuned on target data as well as in zero-shot setups (Wu and Dredze, 2019; Artetxe et al., 2020).

Following the standard input scheme (Devlin et al., 2019), for all fine-tuning tasks, we add a special `[CLS]` token serving as an aggregated sentence representation to the beginning of each sentence, and a special `[SEP]` token indicating the end of a sequence to the end of the sentence.

Sequence Labeling For the sequence labeling tasks, we assign labels for each sub-token. Negation cue information is provided to the model by inserting a special `[CUE]` token before the negation cue (see Figure 4).[16] Special labels are assigned to the `[CUE]` token, `[CLS]` token, `[SEP]` token and subword tokens that are not the first subword of a word. These are ignored during evaluation, which means for evaluation a word is considered correctly labeled if the first subword token generated from that word receives the correct label.

Sequence Classification For the sequence classification task, we use the `[CLS]` token as aggregated sentence representation and assign a label to the whole sequence by feeding the encoded `[CLS]` token representation through the output classification layer.

[14] We use the `bert-base-multilingual-cased` version.

[15] In our experiments, we fine-tune both the encoder and the classification layer on the downstream task.

[16] Khandelwal and Sawant (2020) found this method to work better than *replacing* the negation cue by the special token. Sergeeva et al. (2019) suggest to add a special cue embedding to the subtoken embedding at the input layer, which we found to perform comparable to the insertion method on the BIOSCOPE abstracts.

	Model	Resources	Test data					
			IULA		NUBES		FRENCH	
			F1	PCS	F1	PCS	F1	PCS
H	Punct	-	91.09	84.38	79.37	68.06	85.32	45.96
SU	mBERT	in-domain	$96.23_{0.23}$	$89.38_{0.63}$	$95.66_{0.22}$	$88.02_{0.78}$	$95.01_{0.28}$	$81.49_{0.77}$
ZS	**ST**	SFUEN	$93.73_{0.12}$	$88.13_{0.63}$	$89.42_{0.50}$	$75.35_{2.02}$	$88.63_{0.11}$	$55.27_{1.53}$
	ST$_{cat}$	BIOSCOPE + SFUEN	$94.21_{0.88}$	$87.29_{1.44}$	$90.24_{0.21}$	$77.08_{1.31}$	$89.15_{0.90}$	$52.34_{4.07}$
	MTL$_{n}$	BIOSCOPE, SFUEN	$94.24_{1.52}$	$86.67_{4.16}$	$89.41_{0.34}$	$75.67_{2.16}$	$87.93_{0.89}$	$47.43_{5.12}$
	MTL$_{n+e}$	SFUEN, M2C2	$93.49_{0.57}$	$85.63_{1.65}$	$89.88_{0.41}$	$74.33_{2.33}$	$87.89_{0.23}$	$53.29_{1.39}$

Table 3: Results for negation scope resolution with gold cues on the clinical datasets (with converted annotations). Scores are averaged over three runs with different random seeds, with standard deviations as subscripts. We report results for heuristics (**H**), supervised (**SU**) and zero-shot (**ZS**) experiments. For ZS, we report the best single task model (**ST**) with training on concatenated data (**ST**$_{cat}$), and multi-task models trained on negation scope resolution tasks (**MTL**$_{n}$) or with and additional event detection task (**MTL**$_{n+e}$). Best performance in the **ZS** setup is marked in gray.

5.2 Evaluation

We report two widely used evaluation metrics for negation scope prediction: Percentage of correct spans (PCS) and F1 over scope tokens (Morante and Blanco, 2012). The latter is the standard F1-score computed on the token level, whereas PCS is computed on the span level and considers a predicted span correct if it exactly matches the gold span. PCS is stricter in general and due to different annotation schemes (with respect to inclusion of subject and type of negation cue), PCS is a challenging criterion in zero-shot setups when models are trained on an annotation scheme that differs from the target annotation scheme.

6 Experiments

We split all datasets into 70% for training, 15% for validation, 15% for testing, and predict negation scope in the test splits of the three clinical datasets IULA, NUBES, and FRENCH. Results are shown in Table 3.[17] We report zero-shot performances (**ZS**) of a single task model trained on a single dataset (**ST**) and on a concatenation of datasets (**ST**$_{cat}$). For the multi-task model, we either train on several negation scope resolution tasks (**MTL**$_{n}$), or add an additional event detection task (**MTL**$_{n+e}$). For each model, we report the best combination of training datasets determined by the best average F1-score across the validation splits of the three clinical datasets. Validation scores for other dataset combinations can be found in our code repository.

As baselines, we report in-domain performance of mBERT , which provides an upper baseline indicating the performance gain if annotated target data was available. We also include a punctuation baseline, as previous work has found that many test sentences are easy to label from cue token to next punctuation mark (Fancellu et al., 2017; Sergeeva et al., 2019). The punctuation baseline labels all tokens following the negation cue until the next punctuation mark[18] as *in scope*. In all experiments, mBERT is fine-tuned with the default hyperparameters for sequence labeling tasks implemented in MT-DNN, which means that we use the adamax optimizer with a learning rate of 5e-5 and a batch size of 8. The maximum sequence length is set to 512.[19] All models are trained for a maximum of 50 epochs, and the best model is selected according to F1-score on the BIOSCOPE validation split.

6.1 Results

Baselines The in-domain experiments show that NUBES and FRENCH are harder to predict than IULA. The performance of the punctuation baseline as measured by F1-score is already high, but

[17] Note that our results are reported on the test sets with converted annotations, and hence not directly comparable to numbers reported on the original data.

[18] We consider as punctuation marks any of the following characters: *.,:;!?()[]*

[19] Sequences longer than the maximum sequence length are truncated, rather than split into shorter sequences. However, this affects none of the sequence in our test sets.

still considerably lower than the performance of the in-domain models.

Zero-shot For the zero-shot experiments, we report the best **ST** model trained on a single dataset across all test sets and metrics, which is the one trained on SFUEN. This is surprising, because neither domain, nor language, nor annotated cue types in SFUEN correspond to the features of the test data.

Combining resources by concatenating datasets improves performance according to F1-score across all datasets. $\mathbf{ST}_{\text{cat}}$ outperforms $\mathbf{MTL}_{\text{n}}$ in all scenarios, except for F1 on IULA where they perform comparably. Hence, even though training datasets are annotated with different annotation schemes that cannot easily be converted to a common scheme, simply concatenating the datasets and training a single-task model seems to be more effective than using a multi-task model.

In addition, we find that adding event detection as an auxiliary task in the $\mathbf{MTL}_{\text{n+e}}$ model cannot improve over the models that only do scope resolution, and in some cases are even detrimental. Overall, we see that it is possible to resolve negation scope in clinical text without labeled training data, and that the a single-task model trained on concatenated data works best.

7 Conclusion

Negation is a frequent phenomenon in clinical text, and resolving its scope can benefit clinical IE tasks. As clinical IE has huge potential to improve healthcare, it should ideally be available regardless of language. Availability in multiple languages, however, is hindered by a lack of annotated data for fine-tuning language models in task-specific data in the target language. In this paper, we show that even without labeled data in the target domain or target language, negation scope resolution in clinical text is possible, by fine-tuning a multilingual language model on available resources from other domains and languages. Even though annotation schemes for negation scope differ, combining available resources for training improves performance in most cases.

Acknowledgements

The authors thank the Lundbeck foundation for funding their research in the context of the BrainDrugs project.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Jeremy Barnes, Erik Velldal, and Lilja Øvrelid. 2020. Improving sentiment analysis with multi-task learning of negation. *Natural Language Engineering*, pages 1–21.

Emily M Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6(3):1–26.

Behrouz Bokharaeian, Alberto Diaz, and Hamidreza Chitsaz. 2016. Enhancing extraction of drug-drug interaction from literature using neutral candidates, negation, and clause dependency. *PloS one*, 11(10).

Behrouz Bokharaeian, Alberto Diaz, Mariana Neves, and Virginia Francisco. 2014. Exploring negation annotations in the drugddi corpus. In *Fourth workshop on building and evaluating resources for health and biomedical text processing (BIOTxtM 2014)*, pages 1–8. Citeseer.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310.

Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'Avolio, Guergana K Savova, and Ozlem Uzuner. 2011. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association*, 18(5):540–543.

Md Faisal Mahbub Chowdhury and Alberto Lavelli. 2013. Exploiting the scope of negations and heterogeneous features for relation extraction: A case study for drug-drug interaction extraction. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 765–771.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch.

Journal of Machine Learning Research, 12:2493–2537.

Hercules Dalianis. 2018. *Clinical text mining: Secondary use of electronic patient records*. Springer Nature.

Clément Dalloux, Natalia Grabar, and Vincent Claveau. 2019. Détection de la négation: corpus français et apprentissage supervisé. *Revue des Sciences et Technologies de l'Information-Série TSI: Technique et Science Informatiques*, pages 1–21.

Kerstin Denecke and Yihan Deng. 2015. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial intelligence in medicine*, 64(1):17–27.

Jacob Devlin, 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)*, pages 4171–4186.

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.

Peter L Elkin, Steven H Brown, Brent A Bauer, Casey S Husser, William Carruth, Larry R Bergstrom, and Dietlind L Wahner-Roedler. 2005. A controlled trial of automated classification of negation from clinical notes. *BMC medical informatics and decision making*, 5(1):13.

Allyson Ettinger. 2020. What BERT is not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Federico Fancellu, Adam Lopez, and Bonnie Webber. 2018. Neural networks for cross-lingual negation scope detection. *arXiv preprint arXiv:1810.02156*.

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: Volume 2, Short Papers*, pages 58–63.

Federico Fancellu and Bonnie Webber. 2014. Applying the semantics of negation to smt through n-best list re-ranking. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 598–606.

Sebastian Gehrmann, Franck Dernoncourt, Yeran Li, Eric T Carlson, Joy T Wu, Jonathan Welt, John Foote Jr, Edward T Moseley, David W Grant, Patrick D Tyler, et al. 2018. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PloS one*, 13(2):e0192360.

Peter B Jensen, Lars J Jensen, and Søren Brunak. 2012. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405.

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 Ureña-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.

Nora Kassner and Hinrich Schütze. 2020. Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818.

Aditya Khandelwal and Suraj Sawant. 2020. NegBERT: A transfer learning approach for negation detection and scope resolution. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5739–5748, Marseille, France. European Language Resources Association.

Natalia Konstantinova, Sheila CM De Sousa, Noa P Cruz Díaz, Manuel J Mana López, Maite Taboada, and Ruslan Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In *Lrec*, pages 3190–3195.

Martin Krallinger. 2010. Importance of negations and experimental qualifiers in biomedical literature. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 46–49.

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. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of ACL*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Salvador Lima Lopez, Naiara Perez, Montse Cuadros, and German Rigau. 2020. Nubes: A corpus of negation and uncertainty in spanish clinical texts. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5772–5781.

Yuan Luo, William K Thompson, Timothy M Herr, Zexian Zeng, Mark A Berendsen, Siddhartha R Jonnalagadda, Matthew B Carson, and Justin Starren. 2017. Natural language processing for ehr-based pharmacovigilance: a structured review. *Drug safety*, 40(11):1075–1089.

Montserrat Marimon, Jorge Vivaldi, and Núria Bel. 2017. Annotation of negation in the IULA Spanish clinical record corpus. In *Proceedings of the Workshop Computational Semantics Beyond Events and Roles*, pages 43–52, Valencia, Spain. Association for Computational Linguistics.

Roser Morante 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)*, pages 265–274.

Roser Morante 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)*, pages 1563–1568.

Roser Morante, Anthony Liekens, and Walter Daelemans. 2008. Learning the scope of negation in biomedical texts. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 715–724.

Roser Morante and Caroline Sporleder. 2012. Modality and negation: An introduction to the special issue. *Computational linguistics*, 38(2):223–260.

Danielle L Mowery, Sumithra Velupillai, and Wendy Chapman. 2012. Medical diagnosis lost in translation–Analysis of uncertainty and negation expressions in English and Swedish clinical texts. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 56–64.

Pradeep G Mutalik, Aniruddha Deshpande, and Prakash M Nadkarni. 2001. Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the umls. *Journal of the American Medical Informatics Association*, 8(6):598–609.

Raheel Nawaz, Paul Thompson, and Sophia Ananiadou. 2013. Negated bio-events: analysis and identification. *BMC bioinformatics*, 14(1):14.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of biomedical semantics*, 9(1):12.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65.

Denys Proux, Caroline Hagège, Quentin Gicquel, Suzanne Pereira, Stefan Darmoni, Frédérique Segond, and Marie Hélène Metzger. 2011. Architecture and systems for monitoring hospital acquired infections inside hospital information workflows. In *Proceedings of the Second Workshop on Biomedical Natural Language Processing*, pages 43–48.

Antonio Moreno Sandoval and Marta Garrote Salazar. 2013. La anotación de la negación en un corpus escrito etiquetado sintácticamente. *Revista Iberoamericana de Linguistica*, 8:45–61.

Elena Sergeeva, Henghui Zhu, Amir Tahmasebi, and Peter Szolovits. 2019. Neural Token Representations and Negation and Speculation Scope Detection in Biomedical and General Domain Text. In *Proceedings of LOUHI 2019*, pages 178–187, Hong Kong. Association for Computational Linguistics.

Anastassia Shaitarova, Lenz Furrer, and Fabio Rinaldi. 2020. Cross-lingual transfer-learning approach to negation scope resolution. In *Swiss Text Analytics Conference & Conference on Natural Language Processing 2020*.

Mary H Stanfill, Margaret Williams, Susan H Fenton, Robert A Jenders, and William R Hersh. 2010. A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association*, 17(6):646–651.

Pontus Stenetorp, Sampo Pyysalo, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Bridging the gap between scope-based and event-based negation/speculation annotations: A bridge not too far. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 47–56, Jeju, Republic of Korea. Association for Computational Linguistics.

Asa Cooper Stickland and Iain Murray. 2019. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning. In *International Conference on Machine Learning*, pages 5986–5995.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Sumithra Velupillai, Hanna Suominen, Maria Liakata, Angus Roberts, Anoop D Shah, Katherine Morley, David Osborn, Joseph Hayes, Robert Stewart, Johnny Downs, et al. 2018. Using clinical natural language processing for health outcomes research: Overview and actionable suggestions for future advances. *Journal of biomedical informatics*, 88:11–19.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, 9(11):1–9.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844.