Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 534–542
Marseille, 20-25 June 2022

Negation Detection in Dutch Spoken Human-Computer Conversations

Tom Sweers, Iris Hendrickx and Helmer Strik
Centre for Language Studies, Centre for Language and Speech Technology
Radboud University, Erasmusplein 1, 6500 HD, Nijmegen, The Netherlands
tom.sweers@hotmail.com, iris.hendrickx@ru.nl, helmer.strik@ru.nl

Abstract

Proper recognition and interpretation of negation signals in text or communication is crucial for any form of full natural language understanding. It is also essential for computational approaches to natural language processing. In this study we focus on negation detection in Dutch spoken human-computer conversations. Since there exists no Dutch (dialogue) corpus annotated for negation we have annotated a Dutch corpus sample to evaluate our method for automatic negation detection. We use transfer learning and trained NegBERT (an existing BERT implementation used for negation detection) on English data with multilingual BERT to detect negation in Dutch dialogues. Our results show that adding in-domain training material improves the results. We show that we can detect both negation cues and scope in Dutch dialogues with high precision and recall. We provide a detailed error analysis and discuss the effects of cross-lingual and cross-domain transfer learning on automatic negation detection.

Keywords: negation detection, corpus annotation, transfer learning

1. Introduction

Automatic negation detection is a crucial building block within the dialogue manager of a conversational agent as the one developed in the BLISS project[1] (van Waterschoot et al., 2020). The dialogue manager needs to perform proper negation handling both from a communicative perspective and for content extraction. For example, when a user while talking with an agent says 'i dont understand what you mean' as response to a question from the agent, this should be recognized as a misunderstanding and not as an answer to the question.

Various studies have shown the effectiveness of automated negation detection and scope resolution in English in text mining applications (e.g. (Li and Lu, 2018; Gautam et al., 2018; Chen, 2019; Khandelwal and Sawant, 2020)). This paper presents the first study on the applicability of negation cue detection and scope resolution in Dutch spoken dialogues. Since there exists no spoken Dutch labeled negation corpus, a small Dutch dialogue corpus was annotated with negation cues (negation expressions) and their scopes. This corpus sample is publicly available for research. To counter the relatively small size of the corpus, transfer learning was used by training models on English corpora from different domains. Therefore this study also investigates the applicability of cross-lingual and cross-domain transfer learning in negation cue detection and scope resolution.

We discuss negation from an NLP perspective. In Section 2 we first discuss related work on models for automatic negation detection and corpora annotated with negation in Section 3. Section 4 presents the annotation guidelines and the Dutch spoken corpus sample annotated with negation. We present the models, baselines and experimental setup in Section 5, the results in Section 6 and a thorough error analysis and reflection on our results in Section 7, finally ending with a conclusion.

2. Background

The first NLP approaches to automatic negation detection in texts were rule based. The most influential of these is NegEx (Chapman et al., 2001), a simple regular expression based algorithm that detects both negation cues and scopes. NegEx has been used as a basis for many other rule based approaches to negation handling. The study by Chapman demonstrated one particular weakness in NegEx as it had difficulty finding the correct negation scope when there were many intervening tokens between the affected part of the sentence and the negation cue word.

In the medical text mining domain negation detection has received substantial attention (Huang and Lowe, 2007; **?**; Goldin and Chapman, 2003; Rokach et al., 2008; Morante et al., 2008; Agarwal and Yu, 2010) as in biomedical automatic information extraction it is crucial to determine whether a certain relation does hold or does not hold. Many of these studies are based upon the BioScope corpus (Vincze et al., 2008) that consists of English biomedical scientific texts manually labeled with negation cues and scope. We also found a few studies on negation detection for Dutch within the medical domain. Fivez (2019) created a rule based baseline system and **?**) presented a rule based ConTextD algorithm which has been adapted from the English ConText algorithm (Harkema et al., 2009). They also composed and annotated a Dutch corpus of clinical texts called EMC Dutch and evaluated their algorithm on this corpus. They reported an F-score of 0.87 to 0.93 for negation detection and were satisfied with their results.

[1] `http://bliss.ruhosting.nl`

Current approaches use deep learning architectures for creating automatic negation detection models. Here we discuss only studies that are closely related to our own approach.

?) looked into the cross-lingual negation scope resolution applicability of neural networks (BiLSTM specifically). The authors improved their BiLSTM model by adding the encoding of a dependency tree to the model, which they called D-LSTM. They also experimented with a Graph Convolutional network (GCN), and used an ensemble to combine these two models in different ways. They trained and evaluated their models on an English and a Chinese corpus, while also training the models on the English corpus and evaluating them on the Chinese corpus. For both the English and the Chinese corpus, **?**) report that the ensemble of BiLSTM and D-LSTM outperforms the other models (and thus the state of the art). They show that all cross-lingual models 'approach' the monolingual Chinese results, and thus conclude that it is possible to build a cross-lingual model of negation.

Khandelwal and Sawant (2020) created a model using BERT to detect negation cues following previous work on an attention-based neural approach by Chen (2019). Khandelwal and Sawant (2020) evaluated NegBert on several benchmark data sets such as the previously mentioned the BioScope corpus and reported that their model outperforms the state of the art by a significant margin on all data sets.

Closest to our work is the study by Gautam et al. (2018) who used LSTM models for negation detection in tutorial dialogues (DT-Neg). Instead of written English dialogues, we use Dutch spoken human-computer dialogues and we applied a transfer learning method to overcome the lack of labeling training material.

3. Corpora

An overview of all corpora annotated for negation can be found in the work by Jiménez-Zafra et al. (2020). The authors of this paper have extensively researched all available negation corpora and their respective annotation guidelines in several languages, including English and Dutch.

The negation annotated corpus used most often outside of the biomedical text mining domain is the ConanDoyle-neg corpus (Morante and Daelemans, 2012). It consists of sentences from two stories written by Sherlock Holmes author Arthur Conan Doyle. This corpus was also used in the benchmark *SEM 2012 competition task on negation detection (Morante and Blanco, 2012). The only corpus annotated with negation that consists of (written) dialogues is DeepTutor-Negation (Banjade and Rus, 2016). This corpus consists of annotated dialogues from high-school students conversing with a tutoring dialogue system (DeepTutor). We used both the ConanDoyle-neg and DeepTutor-Negation in our experiments as training material.

We manually labeled the negation cues and scopes in dialogue samples from two spoken Dutch corpora: BLISS (van Waterschoot et al., 2020) and JASMIN (Cucchiarini et al., 2006). The BLISS corpus consists of 55 Dutch spoken conversations between people and a computer about daily life activities and their well-being. The JASMIN corpus contains 95 hours of manually transcribed speech by children, eldery and not-natives. We selected a subsample of conversations with native Dutch elderly people talking about their daily life with a computer[2]. Most utterances are part of question-answer pairs where the questions are posed by a computer and answered by a human. We sampled question-answer pairs that contained at least one negation cue.

Note that the Dutch corpus samples contain spoken human-computer interactions and these conversations tend to contain more confusion and misunderstanding (and thus negation cues) than human-human interactions.

4. Corpus Annotation

In this section we present a brief overview of the guidelines that were used for the negation annotation in the Dutch spoken dialogue sample. We based our annotation guidelines on the already extensive guidelines developed for English (Morante et al., 2011). We annotated both negation cues and the scope of the negation, which is that part of the sentence that describes the event or state being negated by the negation cue. We made some additions to the guidelines to handle typical spoken phenomena such as repetitions, interjections, unintelligible or aborted words, as will be detailed in section 4.2.

4.1. Cues and Scopes

The Dutch data consists of manually transcribed spoken dialogue utterances from human-computer interactions. A negation cue is a word or a sequence of words that expresses negation such as *niet* (En: 'not') or *zonder* (En: 'without'). Multi-word expressions can also function as negation cues such as 'geen sprake van' (En: 'absolutely not'). Typical negation cue words such as 'not' do not always function as negation markers. Such false negation cues most often are cases where the cue word is part of a multi-word expression like *kan het niet helpen* (En: 'beyond my control').

The scope of a negation cue is exactly that part of the sentence which describes the event or state being negated by the negation cue. We follow the approach taken by Morante and Daelemans (2012) and exclude the negation cue itself, sentence final punctuation and discourse modifiers from the scope. The scope can be discontinuous. When the negation cue directly negates

[2] JASMIN dialogues are so called wizard-of-oz conversations in which people thought they were talking to a computer, but in reality the computer was operated by a human.

a verb, the full clause of the verb is the scope for negation as shown in example 1.

(1) [Ik praat] **niet** [vandaag].
En:'I talk not today'

4.2. Adaptations for speech

The most noteworthy addition to the annotation protocol relates to cases where the scope of the negation spans multiple utterances. The spoken dialogues in our sample have the typical question-answer structure and very often a negation cue in a user answer also affects the previous question posed by the computer.

We also had to include rules for annotating typical speech phenomena. Speech interjections like 'uh', are not considered part of the scope. People often misspeak and only utter a part of a word (aborted) as shown in Example 2.

(2) 'k heb ge*a geen voorkeur.
En: I have n* no preference.

Besides aborted words there are also incomprehensible words that could not be understood by the manual transcriber. These are coded as 'ggg' (JASMIN corpus) or 'xxx' (BLISS corpus). We do not consider aborted words or unintelligible speech codes as part of the negation scope as we can simply not be certain of their meaning or purpose in the utterance.

There can also be cases of ungrammatical sentences or spelling mistakes. These are not unique for speech, but do occur with high frequency in our corpus sample. In the extreme case we cannot extract the scope from the transcribed utterance because it is incomprehensible due to unknown, aborted or unrecognized words or due to grammatical or spelling errors, we do not annotate the cue and only mark the utterance as incomprehensible.

Another phenomenon that is frequent in the spoken dialogues is the usage of repetition of negation markers. These turned out to be difficult to label consistently as one can view the repeated negation marker as a separate cue, or as an element expressing emphasis. We show a difficult example in 3 where it is unclear if the negation cues in the answer (A) 'nee nee nee' are referring to the posed question (Q), the phrase 'ik kom nooit nergens' (En:'I never go nowhere') or 'durf ik echt niet hoor' (En: 'I really do not dare'). For this example we annotated the latter option as it was the closest plausible reference.

(3) Q: Kunt u een bezienswaardigheid noemen ?
Denk bijvoorbeeld aan iets dat u daar specifiek wilt gaan zien .
A:mmm ik kom nooit nergens . durf ik echt niet hoor . nee nee nee .
En:Q: Can you recall a place of interest? Think about something that you would like to visit there. A: mmm I never go nowhere. I really do not dare. no no no.

	JB		CD		DT	
	QA	Cues	Sen	Cues	QA	Cues
Trai	250	439	848	984	1088	1088
Val	400	772	144	173		
Test	80	144	-	-	-	-
Total	730	1355	992	1157	1088	1088

Table 1: The size of the annotated corpora.

4.3. Data preparation

As the two Dutch corpora were stored in different file formats, we needed to convert and unify their format for manual annotation.We used an XML-based annotation format, FoLiA[3] (Format for Linguistic Annotation) (Van Gompel and Reynaert, 2013) for the annotated corpus sample.

Before the manual annotation process the cues were automatically annotated. This was done using a list of cues, composed from earlier research (Haeseryn, 1997; Afzal et al., 2014; Fivez, 2019), merged into a regular expression. The annotators were instructed to manually merge separate cues into one multi-word expression cue when deemed necessary, or to remove cues in case of false negation cues. They were also told to still read the entire text, to annotate cues that were not annotated by the regular expression. Affixal cues (like 'ir' in 'irrelevant') have not been labeled by the annotators due to technical limitations of the annotation software. In total 730 question answer pairs were annotated. To compute inter-annotator agreement, 50 pairs were annotated by two annotators, leading to an inter-annotator agreement F-1 score of around 0.94 for the cue and 0.91 for scope labeling.

The focus of this study is cross-lingual negation detection in a corpus of relatively small size. In the Dutch corpus JASMIN-BLISS-Neg (JB) only question answer pairs with cues have been included. The corpus is available for research purposes here: https://github.com/LanguageMachines/JASMIN-BLISS-Negation. The size of the corpus and distribution between training, validation and test sets is shown in Table 1. The Dutch data has a total of 730 question answer pairs, with 1207 cues which span over 1355 words. The average length of the scopes over all data is about 4.8 words. We also list the size of the English corpora, ConanDoyle-neg (CD) and DeepTutor-neg (DT) in Table 1. Note that CD contains sentences that have no negation, while DT and JB only contain question-answer pairs with negation.

[3] https://proycon.github.io/folia/

5. Method

The model used as a basis for the automatic negation detection is NegBERT (Khandelwal and Sawant, 2020). We replaced regular BERT in NegBERT with multilingual BERT (mBERT) (Devlin et al., 2019). We follow up on earlier work where mBERT was used in NegBert to perform cross-lingual negation detection from English to French and Spanish (Shaitarova et al., 2020).

A standard two-stage approach was adopted, using the output of a cue detection model as the input for the scope resolution model. The training input of the model consists of sequences of words and the corresponding labels. We ran a small grid search experiment to determine the best learning rate (3 rates: 1e-6, 1e-5, 2e-5) and batch size (8 and 16) for our experiments. The BERT tokenizer was used to tokenize the sentence, which requires some extra processing as the BERT tokenizer slightly differs from the token representations in the existing corpora. We aligned the manual annotation labels of cues and scopes with the tokens as generated by the BERT tokenizer and added padding labels for the end-of-sentence elements. The cue detection part of the model uses mBERT as a classifier to classify each token as cue or not a cue. The scope resolution part of the model takes the sentence and the cue labels as input when predicting, and classifies each token as part of the scope or not.

5.1. Experimental Setup

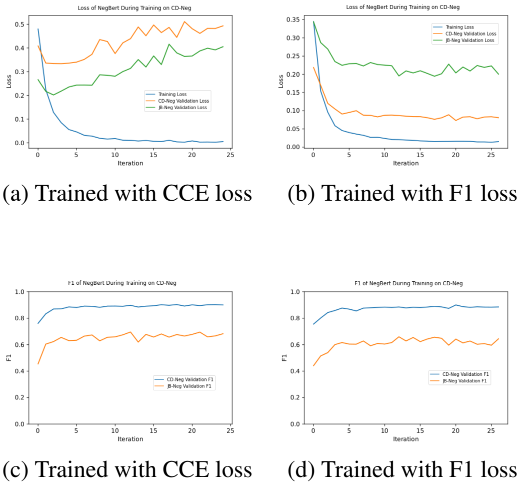
Three different corpora were used in our experiments: ConanDoyle-neg (CD), DeepTutor-neg (DT) and the annotated JASMIN-BLISS-neg (JB) corpora. To evaluate the trained models the precision, recall and F1 scores at the token level were calculated, in line with earlier work (Morante and Blanco, 2012). For the scopes we measured F1 at the token and scope level. At the token level, we count each token as part of the scope or not, while at the much stricter scope level, we only count a true positive when all tokens in the scope were actually labeled as inside the scope. We report in detail on the results on the validation set as we also performed an error analysis on the validation set to gain insight into the differences in performance we observed.

Using NegBert with mBERT and combinations of corpora as training data, the following experiments were conducted. The two parts of the negation detection algorithm, cue detection and scope resolution were trained and evaluated separately. For cue detection different setups were tested and evaluated on the JASMIN-BLISS-neg corpus. First mBERT was fine-tuned on the following combinations of corpora: [1] only ConanDoyle-neg (CD), [2] ConanDoyle-neg and DeepTutor-neg (CD,DT), and [3] ConanDoyle-neg, DeepTutor-neg and a small (not included in the evaluation data) part of JASMIN-BLISS-neg (CD,DT,JB).

During the course of this study, we noticed a problem that required extra experiments. The model seemed to overfit while using the loss function categorical cross-entropy, which was used in the original implementation of NegBert (Khandelwal and Sawant, 2020). We tried to solve this problem by using a different loss function, namely the F1 loss. This loss was chosen since it models the real error better than the categorical cross-entropy loss. We compared the two methods on the scope resolution task when training on ConanDoyle-Neg.

We compared our model to a rule based baseline model. The baseline for cue detection is a simple regular expression that searches for the occurrence of a list of cues. The simplest option for the baseline for scope resolution is to mark all the words in the utterance in which the cue is found. This is called the *utterance baseline*.

The Smart scope baseline rule marks all words in an utterance except when the cue is the only word in the utterance. In that case it includes all the words in the question preceding the utterance. This should improve the performance in scenarios where the user answers just with ‘no’ to a question.



(a) Trained with CCE loss

(b) Trained with F1 loss

(c) Trained with CCE loss

(d) Trained with F1 loss

Figure 1: Graphs for the model training results. Top row reports on loss scores, while bottom row reports F-scores.

6. Results

6.1. Model Training

First we present the results of the scope resolution experiment where two different loss methods were compared. This model was trained on the ConanDoyle-Neg corpus. Figure 1 shows the loss over time on the English training data (blue lines in the graph), the English validation data (orange lines), and the Dutch validation data (green lines) on the first row of images. Below we show the F-scores on the English (blue lines) and Dutch validation data (orange lines). The graph on the top-left (1a) shows the results when the Cross Categorical Loss

Model	Cues		
	Prec.	Rec.	F1
Baseline	0.98	0.99	0.99
CD	0.96	0.77	0.85
CD,DT	0.89	0.82	0.85
CD,DT,JB	0.97	0.99	0.98

Table 2: The results of the cue experiments on validation data.

was used. The graph shows that the loss on the validation data (English and Dutch) increases over epochs, while the training loss decreases, indicating that this model is over-fitting. The top-right graph (1b) shows the results when the F1 loss was used. Here the loss on the validation data slowly decreases, slower, but similarly to the training loss. It is clear that the model trained with the categorical cross entropy loss overfits, while the model trained with F1 loss does not. We used the model trained with F1 loss in the experiments on the validation and test set.

6.2. Cue Detection

We report results on the validation data as we used them for detailed error analysis. The results on the test data are very similar to the scores obtained on the validation set.

Table 2 shows the results for the cue experiments on the validation data. Recall that the validation data contains 400 question answer pairs. The baseline performs very well with a precision of 0.98, a recall of 0.99 and thus an F1 of 0.99. These high results are due to the fact that the rule based baseline is the same method as the method used to automatically annotate the cues. The baseline does not score 100% as some multi-word cues were missing and there were a few false negation cues. The models trained on the English training sets, ConanDoyle-Neg and DeepTutor-Neg (CD,DT) miss some of the cues, which results in a lower recall scores compared to the baseline and the model (CD, DT, JB) that was trained on both Dutch and English material. Note that the trained model does not outperform the baseline. We also see a drop in precision for the CD,DT model, this is is due to a specific type of erroneous cue labels (false positives) that we will discuss in detail in section 7.2.

6.3. Scope Resolution

We report the experiments of negation scope prediction. Note that for these results we used the gold-standard cues as starting points for the scope resolution. Table 3 shows the results of the scope experiments on the validation data.

The models of all three experiments outperform the baseline model substantially with regard to the scope level recall and F1. The model trained on all three corpora performed best.

Model	Scopes Tokens			Scope Level		
	Prec.	Rec.	F1	Prec.	Rec.	F1
Utt Base	0.56	0.59	0.57	0.96	0.13	0.23
Smart Base	0.44	0.76	0.55	0.98	0.16	0.28
CD	0.76	0.49	0.6	0.99	0.24	0.39
CD,DT	0.75	0.64	0.69	0.99	0.33	0.49
CD,DT,JB	0.88	0.83	0.85	1.00	0.57	0.73

Table 3: Results of the scope resolution experiments on the validation dataset (using gold-standard cues) with utterance and smart baseline, and the three training models.

Model	Scope Tokens			Scope Level		
	Prec.	Rec.	F1	Prec.	Rec.	F1
CD	0.73	0.45	0.56	0.82	0.18	0.29
CD,DT	0.71	0.57	0.63	0.79	0.29	0.42
CD,DT,JB	0.77	0.81	0.79	0.87	0.59	0.70

Table 4: The results of full negation detection on the test set.

6.4. Full Negation Detection

In this section we present the scores for full negation detection, chaining the models for cue detection and scope resolution. This implies that errors made in the cue detection step propagate to errors in the scope resolution as happens in any practical application of the module. Table 4 reports scores on combined automatic cue and scope detection on the test set. As the cue detection is performed with very high recall and precision scores, the effect of error propagation on the overall full negation detection task causes only a very marginal drop in the score compared to scope resolution based on ‘perfect’ negation cues.

7. Discussion

We performed an error analysis on the validation data to inspect the typical errors made by the different models. We divide our observations into cross-lingual errors and cross-domain errors.

7.1. Cross-Lingual Transfer Learning Errors

We first take a closer look at the false positives in the cue detection model. Table 5 shows the most frequent cases of false positive cues produced by the different models. Only the words *niet, nee* and *geen* are actual potential negation cues in Dutch. The most remarkable

Word	Freq	Base	CD	CD,DT	CD,DT,JB
’t	80	0	6	50	0
niet	454	5	5	4	5
wel	109	0	5	3	0
moeilijk	7	0	4	0	0
’n	10	0	0	3	0
nee	149	10	0	3	10
geen	101	2	2	2	2

Table 5: The most interesting cases of **false positive** word cue counts on the validation set for baseline (base) and the three different training models.

error is the fact that the model trained on ConanDoyle-Neg and DT-Neg erroneously labels “’t” 50 times as a cue. This token is short for the Dutch determiner ‘het’ (En: ‘it’). We suspect that this error is caused by a resemblance to the English cue ‘n’t’ (short for ‘not’). That this happens more frequently with the second model probably has two reasons. First, since ‘n’t’ is a more informal writing form, it has a higher relative frequency in the DT-Neg dialogues than in the ConanDoyle-Neg stories. Second, with the inclusion of DT-Neg there is an absolute higher frequency of the ‘’t’ available in the training material.

Table 7 shows the occurrence of each cue in the JASMIN-BLISS-Neg validation data in the second column and the number of true positives of each cue detection experiment in the following columns. Often occurring cues such as *niet*, *geen* and *niets* have been predicted correctly more than 90% of the time. These cues correspond to the English cues *not* (*n’t*), *no* and *nothing*, respectively. The cue *niks* (En:‘nothing’) was not recognized by the English training models, but was predicted correctly by the third model. The negation cue *nee* stands out as it has been predicted correctly only once out of the 139 ‘nee’ cues by the first model while the second and third model perform much better. This is likely due to the fact that *nee* is a negation cue most often used as an answer to a yes or no question, or to deny an earlier statement made by someone else. These questions are more likely to happen in dialogue than in other textual genres. Such dialogues can be also present in fiction. Interestingly it only occurs in two fold in the ConanDoyle-Neg corpus as *‘no, no’*.

As the second and third model both include dialogues in training material, these models are trained on examples of the marker ‘no’ in a question-answer setting.

7.2. Cross-Domain Transfer Learning Errors

The model of the first experiment (training on ConanDoyle-Neg) obtained a substantially higher precision than recall score on the validation data. One possible explanation for this is that in CD the scope has always been limited to the sentence in which the cue occurs. The Dutch dialogues from JB often contain negation cues where the scope is the previous question, in fact out of 687 cues in the validation set 120 cues have tokens in the scope that are not in the same utterance as the cue. An example can be found in Example 4.

(4) Q: [Wilt u gebruik maken van een reisgids] ?
A: **nee** . **nee** .
(En: Q: ‘ Would you like to use a travelguide ?
A: no . no . ’)

Table 6 shows the distribution of the most occurring false positives in the results of the three different models in scope detection. Some of these false positives are high frequent words (*ik, dat, heb* (En: I, that, have)), that were erroneously predicted due to various individual sentence-specific reasons. One error that occurs relatively often in the first two models is the inclusion of common speech patterns such as *uh*, or unrecognized words or background noise in the scope. Another interesting speech pattern that often causes false positives is *hoor*, which is an interjection used to emphasize a statement (including negations). Example 5 shows a rather complex, partly ungrammatical user answer. The model incorrectly included the interjection as part of the negation scope of cue ‘niet’ (not) in the last utterance:

(5) Gold: Dat klopt niet . U krijgt nog 1 kans . klopt ’t ... ggg . krijg de nog een kans . nou ja . [’k zou ’t] anders **niet** [weten] hoor .
Prediction: Dat klopt niet . U krijgt nog 1 kans . klopt ’t ... ggg . krijg de nog een kans . nou ja . [’k zou ’t anders] **niet** [weten hoor] .
(En: That is not right . You get another 1 chance . is ’t right ... ggg. get the yet a chance . well yes . I would not know otherwise .)

Other often occurring false positives are negation modifiers such as *echt* (En:‘really’). As these are relatively rare modifiers, we expect that a larger training set with Dutch examples could solve the problem of these false positives.

One interesting error that can occur in all three models is a case of negative concord in the corpus. In Example 6 the annotator deemed the negation *ben nooit niks kwijt geworden* (En: ‘have never nothing been lost’) as negative concord, meaning that the two negators signal a single negation and not a positive. The first model did interpret this as a double negation where it would result in a single positive, where *niks* is in the scope of *nooit.*

(6) Q: Bent u van plan waardevolle spullen mee te nemen die dag ?
A: nou [ben] nog **nooit niks** [kwijt geworden] zal wel meevallen ik heb u niet verstaan u moet harder praten ggg oh mij mooi . n kunt u harder praten ja . ja foutmelding .
(En: ‘are you planning to carry valuables on that day ? A: well have never nothing been lost will

Words	CD-Neg		CD-Neg,DT-Neg		CD-Neg,DT-Neg,JB-Neg		Frequency
	FP	Predicted Positives	FP	Predicted Positives	FP	Predicted Positives	
meer	62	64	65	67	1	2	171
uh	25	25	34	34	7	7	367
ik	20	237	30	334	18	331	1444
nee	19	19	23	23	0	0	281
echt	14	14	14	15	12	13	41
dat	13	69	20	105	15	149	800
helemaal	12	12	24	25	10	10	86
ggg	11	11	17	17	3	3	191
heb	11	94	8	104	2	103	373
hoor	9	9	12	13	6	7	25

Table 6: The distribution of the most occurring (in the first model) false positives in scope detection, and shows these false positives for all three models.

Cues	Freq	Base	CD	CD,DT	CD,DT,JB
niet	449	449	449	416	449
nee	139	139	1	81	138
geen	99	99	99	97	99
niks	31	31	0	0	31
nooit	22	22	22	21	22
niets	8	8	7	6	7
nergens	6	6	0	0	6
zonder	4	4	4	3	4
niemand	3	3	3	3	3
sprake	1	0	0	0	0
van	1	0	0	0	0
neen	1	0	0	0	1

Table 7: Cues and their frequency in the validation set followed by **true positive** cue counts for baseline (base) and the three different training models.

> be okay i did not understand you need to talk louder ggg oh me nice . can you talk louder yes. yes error .')

7.3. Reflection on Results

In this study we observed that solely for the cue detection, a simple regular expression as baseline works just as well as training a supervised learning model. We had hoped that an supervised method would have the advantages of picking up new unseen cues and recognizing the difference between false and real negation cues. Our results show that indeed a new cue word ('neen') was correctly picked up by the supervised model trained on combined English and Dutch material. The model did not do a better job on recognizing false negation cues than the baseline. These false cues remain difficult to detect and we want to remark that also for the human annotators false cues are hard to recognize consistently.

For detecting the scopes of negation supervised models largely outperformed the baseline. This is in line with the findings in earlier research in cross-lingual transfer learning for negation detection (Shaitarova et al., 2020), where ConanDoyle-Neg was used to train a model evaluated on Spanish and French corpora.

Regarding the question of the effect of cross-domain transfer learning, and including dialogue based data, something more has to be said about the difference in performance of the two models trained solely on English data. In scope resolution the model trained on data that included the English dialogues performed substantially better.Cues that typically occur in the domain of spoken dialogues such as *nee* ('no' answer to a question) and *niks* (informal spoken form of 'nothing') were not detected by the English fiction trained models. The effect of including and excluding data in the same language as the evaluation data was also shown. The results clearly show the limitation of zero-shot cross-lingual transfer learning, and show the possibilities of cross-lingual transfer learning with minimal training data in the target language. The analysis showed the limitations of cross-lingual transfer learning as well.

As for the corpus and its accompanying guideline that were created for this study, it can be said that adopting English guidelines to create a guideline for Dutch, two Germanic languages, is a relatively straightforward task. The harder part was adapting the guidelines for spoken language and conversations. The rules for scopes had to be adapted such that the entire question answer pair could be included and rules had to be added to exclude speech interjections and unrecognized words.

7.4. Further Research

The results of this study are promising, since it appears that negation detection is possible for spoken dialogues in Dutch without actually training on spoken dialogues in Dutch. Training a model only on English examples leads to a reasonable performance with a tradeoff of rather high precision (around 80% prec at the scope level, Table 4, but low recall meaning that not all cues are picked up, but those that are recognized are also mostly correct. Even better results were achieved when a small set of Dutch spoken dialogue examples was added.

This research was motivated by the desire to increase understanding of using negation detection in conversational agents. Current performance is sufficient to investigate in future steps how one could integrate a negation recognition module based on NegBert within the dialogue manager of a conversational agent.

Another issue that requires more research is the earlier discussed best method for modeling repetition of negation cues. We now treated them as separate cue elements, but another possible solution is to model repeated cues as a different sub class entirely. A more thorough linguistic analysis is required to determine how to handle and interpret repeated cues. In the current study we also excluded affixal negation, and a future work should include affixal negation detection as well.

More research is also required into the size of the minimal training data set. In this study we used a fixed size of minimal training data sets (250 dialogue pairs). The effect of the size of the minimal training data set requires a new study. Furthermore, the dialogues in the current sample are rather similar as the number of repeated questions lead to similar answers. So far we also only worked with manually transcribed speech and not raw automatically recognized speech (ASR output). Larger and more varied training material will improve the robustness of the negation detection model.

8. Conclusion

This study researched the applicability of negation detection for Dutch spoken conversations, specifically by means of cross-lingual and cross-domain transfer learning. To do this we adapted existing annotation guidelines for English negation cues and scopes to a guideline for negation annotation in Dutch spoken conversations. Based on these guidelines a Dutch corpus with spoken human computer conversations was annotated.

We performed cross-domain and cross-lingual transfer learning experiments by first training a model on English fiction. Then we trained a model on English fiction and English written dialogue. Finally we added a small set of the created Dutch corpus to the aforementioned English data and trained a third model. We found that the model including Dutch data performed better than the other models in both cue detection and scope resolution. The model using both English fiction and dialogue outperformed the model only trained on English fiction by a smaller margin in scope resolution. These two models had a similar F1-score in cue detection, with the latter having a higher precision while the former has a higher recall.

To get more insight into the applicability of both cross-domain and cross-lingual negation detection, we conducted an extensive error analysis. This showed that the lower precision in the model trained on data including English dialogues was mostly due to a specific error and is thus not generalizable. Combined with the higher scores in scope resolution for this model (compared to the model only trained on English fiction) we must conclude that cross-domain learning has a slight but substantial impact on performance for negation detection on Dutch spoken conversations. The analysis showed that this is mostly due to cases where the negation scope was extended to the previous question while the cue is in the answer. The English dialogues also contained such examples.

By adapting existing guidelines for negation detection in English to Dutch spoken conversations, this study contributed new methods and insights that can be further adapted to spoken conversations in other languages. This study also produced a corpus annotated according to these guidelines, which is the first Dutch and first spoken dialogues corpus to be annotated with negation cues and scopes. This corpus can be used in future studies, for example in research that applies negation detection to a conversational agent. This study is also the first one to show the applicability of negation detection in Dutch spoken dialogues.

9. Acknowledgments

This work is part of the research programme Data2-Person with project no. 628.011.029, which is (partially) financed by the Dutch Research Council (NWO). We thank Katarina Laken, Zjoske Koks and Maarten van Gompel for their help in the annotation process and we thank Roser Morante for her valuable reflections on negation annotation.

Afzal, Z., Pons, E., Kang, N., Sturkenboom, M. C., Schuemie, M. J., and Kors, J. A. (2014). ContextD: an algorithm to identify contextual properties of medical terms in a Dutch clinical corpus. *BMC Bioinformatics*, 15(1):373.

Agarwal, S. and Yu, H. (2010). Biomedical negation scope detection with conditional random fields. *Journal of the American Medical Informatics Association*, 17(6):696–701, 11.

Banjade, R. and Rus, V. (2016). DT-neg: Tutorial dialogues annotated for negation scope and focus in context. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3768–3771, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., and Buchanan, B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310.

Chen, L. (2019). Attention-based deep learning system for negation and assertion detection in clinical notes. *International Journal of Artificial Intelligence & Applications*, 10:1–9, 01.

Cucchiarini, C., Van hamme, H., van Herwijnen, O., and Smits, F. (2006). JASMIN-CGN: Extension of the spoken Dutch corpus with speech of elderly people, children and non-natives in the human-machine interaction modality. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*, pages 4171–4186. Association for Computational Linguistics.

Fivez, P. (2019). Negation detection of concepts in Dutch clinical text. *GitHub repository*, November. original-date: 2019-01-15T01:39:05Z.

Gautam, D., Maharjan, N., Banjade, R., Tamang, L. J., and Rus, V. (2018). Long short term memory based models for negation handling in tutorial dialogues. In *Proceedings of the Thirty-First International FLAIRS Conference*.

Goldin, I. M. and Chapman, W. W. (2003). Learning to detect negation with 'not' in medical texts. In *In Workshop at the 26th ACM SIGIR Conference*.

Walter Haeseryn, editor. (1997). *Algemene Nederlandse spraakkunst*. M. Nijhoff ; Wolters Plantyn, Groningen: Deurne, 2., geheel herziene druk edition.

Harkema, H., Dowling, J. N., Thornblade, T., and Chapman, W. W. (2009). ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics*, 42(5):839–851, October.

Huang, Y. and Lowe, H. J. (2007). A Novel Hybrid Approach to Automated Negation Detection in Clinical Radiology Reports. *Journal of the American Medical Informatics Association*, 14(3):304–311, 05.

Jiménez-Zafra, S. M., Morante, R., Teresa Martín-Valdivia, M., and Ureña-López, L. A. (2020). Corpora annotated with negation: An overview. *Computational Linguistics*, 46(1):1–52.

Khandelwal, A. and Sawant, S. (2020). NegBERT: A transfer learning approach for negation detection and scope resolution. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC '20)*, pages 5739–5748, Marseille, France. European Language Resources Association (ELRA).

Li, H. and Lu, W. (2018). Learning with structured representations for negation scope extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 533–539, Melbourne, Australia, July. Association for Computational Linguistics.

Morante, R. and Blanco, E. (2012). *SEM 2012 shared task: Resolving the scope and focus of negation. In **SEM 2012: Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 265–274, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Morante, R. and Daelemans, W. (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, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Morante, R., Liekens, A., and Daelemans, W. (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, Honolulu, Hawaii, October. Association for Computational Linguistics.

Morante, R., Schrauwen, S., and Daelemans, W. (2011). Annotation of negation cues and their scope Guidelines v1.0, Technical Report Series CTR-003. Technical report, CLiPS - University of Antwerp, Belgium, May.

Rokach, L., Romano, R., and Maimon, O. (2008). Negation recognition in medical narrative reports. *Information Retrieval*, 11:499–538, 12.

Shaitarova, A., Furrer, L., and Rinaldi, F. (2020). Cross-lingual transfer-learning approach to negation scope resolution. In *Swiss Text Analytics Conference & Conference on Natural Language Processing 2020*, CEUR Workshop Proceedings. CEUR-WS, June.

Van Gompel, M. and Reynaert, M. (2013). Folia: A practical xml format for linguistic annotation - a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3:63–81, 12.

van Waterschoot, J., Hendrickx, I., Khan, A., Klabbers, E., de Korte, M., Strik, H., Cucchiarini, C., and Theune, M. (2020). BLISS: An agent for collecting spoken dialogue data about health and well-being. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 449–458, Marseille, France, May. European Language Resources Association.

Vincze, V., Szarvas, G., Farkas, R., Móra, G., and Csirik, J. (2008). The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(11):S9, Nov.