Automatic Negation and Speculation Detection in Veterinary Clinical Text

Katharine Cheng **Timothy Baldwin** **Karin Verspoor**
School of Computing and Information Systems
The University of Melbourne
Australia
katharinec@student.unimelb.edu.au
tb@ldwin.net
karin.verspoor@unimelb.edu.au

Abstract

The automatic detection of negation and speculation in clinical notes is vital when searching for genuine instances of a given phenomenon. This paper describes a new corpus of negation and speculation data, in the veterinary clinical note domain, and describes a series of experiments whereby we port a CRF-based method across from the BioScope corpus to this novel domain.

1 Introduction

Negation and speculation are common in clinical texts, yet pose a challenge for natural language processing of these texts. Negation indicates the absence or opposite of something, and is defined within the previously released BioScope corpus (a collection of biomedical and clinical documents annotated for the task of negation/speculation detection) to be the "implication of the non-existence of something" (Szarvas et al., 2008). For example, the statement *no abnormalities were found in the patient* indicates the absence of abnormalities in the patient. Speculation is used to indicate uncertainty or the possibility of something, and is defined within BioScope to be statements of "the possible existence of something". For example, *there is possible bacterial infection* indicates that an infection might be present, without any certainty that it is. Both are commonly used in clinical texts as a means of ruling out diagnostic possibilities and hypothesising.

This paper will discuss a method for detecting negation and speculation over clinical records from the Veterinary Companion Animal Surveillance System (VetCompass) project.[1] The VetCompass project is a database of veterinary clinical records for tracking animal health. The database may be used for research on the effects and usage of a particular drug, or the prevalence and distribution of a disease. Such studies are typically performed by querying for terms relevant to a drug or disease of interest, and analysing the retrieved clinical records. However, results identified using keyword matching are often speculative or negated mentions rather than true occurrences. By automatically detecting negation and speculation, we aim to suppress these results, and provide a higher-utility set of documents to the user.

The task of negation/speculation detection is often defined in terms of two subtasks: (1) signal (or cue) detection; and (2) scope detection. Negation/speculation signal (or cue) detection involves determining which words in a sentence indicate that a negation/speculation is occurring. Negation/speculation scope detection involves determining which words in a sentence the negation/speculation applies to, under the constraints that: (a) the cue word is contained within the span of the scope; and (b) the span is contiguous. Consider two examples from the clinical notes subset of the BioScope corpus:

(1) The lungs are well expanded, but $[\![_{\text{NEG}}$ not hyperinflated $_{\text{NEG}}]\!]$.

(2) Mild thoracic curvature, $[\![_{\text{SPEC}}$ possibly positional $_{\text{SPEC}}]\!]$.

The cues here for negation and speculation are *not* and *possibly*, respectively, and the words inside the brackets are within the scope of the cues.

We apply this task formulation to the veterinary clinical notes of VetCompass. The VetCompass records (which mainly consists of notes from veterinary general practitioners) have a few important differences from the radiology clinical notes of the publicly available BioScope corpus. First, radiology notes are often shared between clinicians treating the same patient, and as such are generally written to be accessible to others. In notes from veterinary general practitioners, it is often the case that a single clinician treats the patient,

[1] http://www.rvc.ac.uk/vetcompass

Katharine Cheng, Timothy Baldwin and Karin Verspoor. 2017. Automatic Negation and Speculation Detection in Veterinary Clinical Text. In *Proceedings of Australasian Language Technology Association Workshop*, pages 70–78.

meaning that clinical notes are largely for personal consumption, and thus are highly idiosyncratic in nature. Second, while radiology clinical notes are often professionally transcribed from an oral account by the clinician, in the veterinary general practice context, notes are authored directly by the clinician as text. Inevitably, this is done under time pressure, meaning that the text is often ungrammatical and lacks punctuation. Examples (3) and (4) exemplify negation and speculation in VetCompass:

(3) Mm - moist ⟦NEG no skin tent NEG⟧

(4) Adv ⟦SPEC poss bacterial infection SPEC⟧, adv ⟦SPEC can be allergy in origin SPEC⟧

Such differences in usage between veterinary clinicians and other medical professionals such as radiologists are a major focus of this work, in adapting the annotation framework from BioScope to this new domain.

This paper attempts to address the following research questions: (1) Can the task of negation/speculation detection be applied to veterinary clinical records? (2) Are models trained over the human clinical records of the BioScope corpus applicable to veterinary clinical notes?

This paper describes the process of annotating negation and speculation in veterinary clinical records. We then demonstrate that the task of negation and speculation detection can be successfully applied to veterinary clinical notes using a simple conditional random field (CRF) model. We additionally show that models trained on a related out-of-domain corpus such as the BioScope have utility over veterinary clinical records, in particular for negation detection.

2 Literature Review

2.1 Previous Work in Negation and Speculation Detection

Most work on negation and speculation detection has focused on biomedical documents such as biological research papers and clinical notes, with the latter being most relevant to this research.

Early approaches to negation detection were primarily rule-based. One of the best-known systems for negation detection is NegEx (Chapman et al., 2001), which is based on regular expressions containing a negation cue term (such as *no* or *not*). Another rule-based negation detection system is NegFinder (Mutalik et al., 2001).

More recently, machine learning approaches have become popular. Morante et al. (2008) proposed a machine learning approach that consists of two phases: (1) classification of whether each token in a sentence is a negation cue, and (2) classification of whether each token is part of the negation scope of a given cue. Both phases used a memory-based classifier using features such as the the wordform of the token, part-of-speech (POS) tag, and chunk tags of the token and neighbouring tokens. The approach was also applied to speculation detection (Morante and Daelemans, 2009a), and incorporated into a meta-learning approach to the second phase of negation scope detection (Morante and Daelemans, 2009b). Other approaches that use machine learning include the work of Agarwal and Yu (2010a,b) that uses conditional random fields (CRFs) to detect negation and speculation, and Cruz Díaz et al. (2012) who experimented with the use of decision trees and support vector machines.

Most work on negation and speculation detection has focused on a specific corpus and domain, with some exceptions. Wu et al. (2014) investigated the generalisability of different negation detection methods over different domains, and found that performance often suffers without in-domain training data. Miller et al. (2017) also investigated the use of different unsupervised domain adaptation algorithms for negation detection in the clinical domain and found that such algorithms only achieved marginal increase in performance compared to systems that use in-domain training data.

2.2 Veterinary NLP

We are only aware of a few papers that have applied natural language processing in the veterinary domain. Ding and Riloff (2015) conducted work on detecting mentions of medication usage in a discussion forum for veterinarians, and categorizing the usage of the medication. A classifier determines whether each word is part of a medication mention using features such as the POS tags and neighbouring words The output of the medication mention detector is used by another classifier to determine its usage category such as whether the clinician prescribed the medication or changed it.

Text classification is a task that had been previously applied to veterinary clinical records. Anholt et al. (2014) performed classification of a collection of veterinary medical records to identify

cases of enteric syndrome. Lam et al. (2007) used clinical records of racing horses to categorise their reason for retirement. Duz et al. (2017) used classification to identify cases of certain conditions and drug use in clinical records from equine veterinary practices. In each of these studies, a dictionary was compiled to identify and detect phrases that indicate a certain category.

2.3 BioScope Corpus

Currently, there is no publicly available corpus for training models over veterinary clinical notes. However, the BioScope corpus (Szarvas et al., 2008) provides a relevant dataset from which to train out-of-domain models. It is a publicly available collection of biomedical documents that have been annotated for both negation and speculation, in the form of cue words and their scope (see Section 1). BioScope consists of three subcollections: clinical radiology notes, biological papers, and abstracts of biological papers from the GENIA corpus (Collier et al., 1999).

3 VetCompass Corpus

The VetCompass project is a collection of clinical records of veterinary consultations from several participating practices, to support analysis of animal health trends (McGreevy et al., 2017). To conduct these studies, clinicians use an information retrieval (IR) front-end to retrieve clinical records related to their particular information need, based on Boolean searches. A major bottleneck for the naive IR setup of returning all matching documents is the prevalence of term occurrences in negated or speculative contexts, which dominate the results for many queries. This is the primary motivation for this research: to improve the quality of the search results by filtering out document matches where the component term only occurs in negated or speculative context. The major challenge here is that the language used in the veterinary clinical notes of VetCompass differs from that used in related publicly available datasets such as the BioScope radiology clinical notes.

3.1 Discussion of VetCompass Corpus

The corpus used in this work was constructed from a random sample of 1 million clinical records from VetCompass UK.[2] VetCompass clinical records contain a wide variety of text. Many records contain free text describing the clinician's observations, hypotheses, and descriptions of treatments and future actions. However, there are also records that contain only billing information, document the weight of the patient, or are reminders to perform certain actions like sending an invoice to the owner of the patient.

Compared to the BioScope radiology clinical notes, VetCompass clinical notes are much more informal, possibly due to the fact that they are largely "notes to self" (see Section 1). As such, ad hoc abbreviations and shortening of terms as shown in Examples (3) and (4) are very common, and informal speculative expressions such as *feels like* and *looks like* are prevalent:

(5) Skin ⟦$_{\text{NEG}}$ not quite so erythematous $_{\text{NEG}}$⟧ but some scurf and ⟦$_{\text{SPEC}}$ looks like superficial pyoderma $_{\text{SPEC}}$⟧.

(6) ⟦$_{\text{SPEC}}$ Feels like lipoma $_{\text{SPEC}}$⟧, but ⟦$_{\text{NEG}}$ cannot confirm without lab tests $_{\text{NEG}}$⟧.

(7) Adv ⟦$_{\text{SPEC}}$ sounds like colitis $_{\text{SPEC}}$⟧ so disp emds btu adv o if no improvement resee and bring in sample.

There are certain negation and speculation cue terms that appear only in the VetCompass corpus such as:

(8) Examination: v lively, ⟦$_{\text{NEG}}$ nad on oral exam $_{\text{NEG}}$⟧ and ghc all fine.

(9) Assessment: ⟦$_{\text{SPEC}}$ gastritis?? $_{\text{SPEC}}$⟧

The term *nad* is often used in place of *no acute distress* or *no abnormalities detected*, and is an instance of negation. Question marks were often used as speculative cue terms such as in Example (9). The use of domain-specific cue terms presents a challenge for applying models that were trained on a corpus like BioScope clinical notes.

Misspellings, grammatical errors and lack of punctuation are also common in the text of the veterinary general practice clinical notes, e.g.:

(10) But depressed last 4 days and srop preds 2.5mg abruptly 4 days ago and sneezing ⟦$_{\text{NEG}}$ wuithout nasla discharge $_{\text{NEG}}$⟧ 2 days too.

(11) Gave deepest sympathies; ⟦$_{\text{SPEC}}$ unsure of cause $_{\text{SPEC}}$⟧ ⟦$_{\text{SPEC}}$ poss underlying condition causing gut stasis or non-specific abdominal pain symptoms or acute embolus this morning $_{\text{SPEC}}$⟧

In Example (10), the negation cue *without* is misspelled. In Example (11), punctuation is missing, making it hard to clearly separate the different statements in the sentence, and suggesting that pure parser-based approaches will struggle over this data.

[2] http://www.rvc.ac.uk/VetCOMPASS

In terms of annotation, while some abbreviations, shorthands, misspellings, and punctuation errors are easy to interpret, others are more difficult to understand:

(12) - other poss: renal diz (given that had low sg + proteinuria, ˆBUN/ˆPhosp BUT N - creat)/liver diz (given hepatomegally on rads + ˆALP, Bile acids, Cholest, ? low sod/K+ ratio - could be related to kids or addisonian crisis BUT no hx of pu/pd

Symbols like ˆ require domain expertise to interpret. The appearance of terms like *poss* indicates that the sentence contains speculation but the irregular use of punctuation makes determining the correct boundaries of the speculation scope difficult. In fact, the absence of certain punctuation marks such as full stops can make it difficult for sentence tokenizers to work correctly.

In the VetCompass corpus, a single statement of speculation is sometimes expressed using multiple speculation cue terms, e.g.:

(13) History- o concerned swollen lower lip, ⟦$_{\text{SPEC}}$ thinks poss stung $_{\text{SPEC}}$⟧, been there 2d

Here, the clinician is reporting that the owner of the patient (shortened to *o*) speculated that the patient was stung, as indicated by two cue terms, *thinks* and *poss*, presumably to indicate their lack of confidence in the statement. Such instances of "double hedging" are very rare in BioScope, presenting an extra point of differentiation.

3.2 Annotation Guidelines

Here, we outline the annotation guidelines for the VetCompass corpus, which borrow heavily from the BioScope annotation guidelines. As per the BioScope annotation guidelines, sentences from VetCompass are annotated for speculation if they express uncertainty or speculation, and annotated for negation if they express the non-existence of something. The min-max strategy of BioScope annotation is also followed (Szarvas et al., 2008). Negation/speculation **cues** are annotated such that the minimal unit that expresses negation/speculation by itself is marked. **Scopes** are then annotated relative to cue words, to have maximal size or the largest syntactic unit possible. Below, we detail important deviations from the BioScope annotation guidelines, which are motivated in part by the usage of the negation/speculation detection system in an information retrieval context.

3.2.1 Annotation of Cues

The VetCompass annotation guidelines use the same set of cue words as BioScope, with the addition of *NAD* (a negation cue — see above), question marks (which are potentially speculation cues — see above), and shortened and misspelled variants of cue words (like *poss* for *possible*).

As with BioScope, not all occurrences of a negation or speculation keyword indicate negation or speculation. For instance, occurrences of negation or speculation keywords in descriptions of proposed actions are generally not annotated for negation or speculation. Examples of such cases are:

(14) Advised to not give last onsior due to d+.

(15) Suggested FNA if increase in size

In Example (14), *not* is not annotated as a negation cue since the sentence is stating a recommendation rather than expressing the absence or opposite of anything. In Example (15), *suggested* is not annotated since it is being used in the sense of proposing an action rather than hypothesising. These examples are also not annotated because of the utility they might provide for a clinician. If a clinician was researching *FNA*, the document containing Example (15) would be potentially useful for understanding situations where such a procedure was proposed. However, actions that were performed in the past that contains negation or speculation would be annotated such as *cannot* in Example (6) which is clearly expressing the opposite of the ability to perform that action.

Conditionals are another situation where negation or speculation keywords may not always be annotated as cues. If a negation or speculation keyword appears in the clause expressing the condition (clause containing the *if*), then they should not be annotated as cues as demonstrated in the following examples:

(16) Adv if O not wanting to consider euthanasia then need to get a veterinary behaviourist involved ASAP

(17) Stop treatment immediately if vomiting or diarrhoea occurs

Here, there is not clear negation or speculation, but rather the lack of something in the conditional (e.g. *consider euthanasia*) or consequent (e.g. *treatment*). While these two sentences may be annotated under the BioScope annotation guidelines, we chose not to do this for the VetCompass clinical records because of the utility they might provide for a clinician. Even if a certain term is negated

inside of a conditional, there is usually other information in the clinical record that provides instructions about what to do in non-negated circumstances which is useful for a clinician. In the case of a term being speculated inside of a conditional, the consequences of the term occurring is certain even if the condition had not occurred.

3.2.2 Annotation of Scopes

In many cases, negation and speculation scopes start at cue terms and end at the end of the clause or sentence. However, punctuation is often omitted, meaning that boundaries of clauses and sentences can be unclear. The annotator must use their own judgement and interpretation of the sentence in order to create a suitable annotation. The following example demonstrates a sentence where an annotator must interpret the sentence to understand where the clause boundaries are:

(18) Abdo palpation ok ⟦NEG no pain NEG⟧ ⟦SPEC poss a little bloated SPEC⟧ ⟦NEG no fluid NEG⟧ thrill abdominally temp normal has had ongoing GI issues occasional use of steroids.

Unlike the BioScope annotations, VetCompass clinical records were not annotated to contain nested speculation scopes, i.e. speculation scopes are never contained within other speculation scopes. This decision was motivated by the expected retrieval usage of the negation/speculation system: such information does not provide additional information to help filter out negated or speculated mentions of certain terms from search results. An example of the implication of this guideline is shown in the following sentence that is annotated with one negation scope and one speculation scope:

(19) ⟦NEG No obvious mass NEG⟧, ⟦SPEC suspect poss trichobezoars? SPEC⟧

The above sentence would have been annotated as three nested speculation scopes under the BioScope annotation guidelines. However, using the VetCompass annotation guidelines, only a single speculation scope will be annotated, containing three separate speculations cues. If a user had wanted to search for documents with *trichobezoars*, this sentence will not be retrieved regardless of whether the nested structure is annotated or not. However, nested negation scopes in VetCompass are annotated. Moreover, speculation scopes that are nested within a negation scope and vice versa are also annotated.

	κ	F1-score
Negation Cue	0.80	80.3
Speculation Cue	0.65	65.5
Negation Scope	0.73	54.8
Speculation Scope	0.73	63.3

Table 1: Inter-annotator agreemenent rates

3.3 Annotation Process

1041 records were randomly selected for annotation. These were divided into a training set, development set and test set, comprising 624, 208 and 209 records, respectively. The data was single-annotated by the first author using the BRAT annotation tool (Stenetorp et al., 2012), in consultation with the other authors in instances of doubt.

100 records (containing 586 sentences) from the test set were selected and annotated by one of the other authors, following the guidelines in Section 3.1. The agreement between the two annotators was calculated using Cohen's kappa (κ) and F1-score (obtained by treating the annotations made by the main annotator as the gold-standard). We measure the amount that the two annotators agreed that a particular token is a negation/speculation cue or scope. The inter-annotator agreement is described in Table 1.

The κ values in Table 1 demonstrate a reasonable amount of agreement between the two annotators. However, there is still some subjectivity, particularly for the speculation cues.

There are several reasons for the discrepancy in annotations between the two annotators: (1) the limited experience in linguistics and text analysis on the part of the main annotator of VetCompass; (2) the lack of pre-training for annotating the VetCompass corpus for the other annotator, beyond receiving the annotation guidelines; and (3) the different levels of familiarity with the datasets of BioScope and VetCompass.

3.4 Preparation of corpus

Sentence tokenization was performed to prepare the corpus for usage, based on the findings of Read et al. (2012). The output of the sentence tokenizer was converted into the BRAT annotation format so that the output could be manually corrected if needed. However, the correction was not a systematic process. A sentence tokenization output was corrected only if it was clearly incorrect from a quick inspection during the annotation process. Most corrections only occurred when nega-

Total documents	1041
Total sentences	6582
Total words	50222
Avg. sentence length	10.06±8.38
% negated documents	41.59
% negated sentences	11.21
Avg. neg. span length	3.69±1.97
% speculated documents	20.65
% speculated sentences	5.15
Avg. spec. span length	5.60±3.57

Table 2: VetCompass NegSpec Corpus Statistics

tion/speculation scopes had the potential to cross sentence boundaries or in clear instances where correct sentence boundaries were not added. Only about 10% of the corpus underwent correction for sentence tokenization.

3.5 Summary of Corpus

Table 2 provides details of the annotated corpus. In general, large variations in sentence length can be observed: some sentences are as short as two words (e.g. reporting the patient weight), while others contain long detailed descriptions of the consultation.

The annotated VetCompass corpus contains a slightly lower proportion of negated sentences compared to those in the BioScope clinical notes (where 13.55% of the sentences were annotated as negated), and a much lower proportion of speculative sentences (compared to 13.39% in BioScope).

4 Methodology

4.1 Model Description

To evaluate whether the task of negation and speculation detection can be applied to the veterinary clinical notes of VetCompass, a simple linear-chain conditional random field (CRF: Lafferty et al. (2001)) model was trained, in the form of a re-implementation of the negation and speculation detection methods proposed by Agarwal and Yu (2010a,b).

The negation detection system consists of two parts: a cue detection system, and a scope detection system. The cue detection system is a CRF that classifies whether or not a given token is a negation cue. A CRF was used for cue detection to be able to model contexts in which cues appear in both negation and non-negation contexts, and to model multiword cues. The scope detection system is also a CRF, and classifies whether or not a token in a sentence is part of a negation scope. The negation cue CRF uses only the words of the sentence as features. For the negation scope CRF, both the words of the sentence and the POS tags were used. When POS tags are used, the words that are part of a negation cue (that were detected by the negation cue CRF model) were either retained or replaced with a special CUE tag. The speculation detection system has a similar setup, except the system classifies a token as being the inside or outside of a speculation cue or scope.

The cue detection system is based on the following features: the target word, and the two words to the left and right of the target word. The scope detection system determines if a token is inside or outside a negation or speculation signal using either the words and POS tags of the token, five tokens to the left and right.

Our experiments are based on the corpus described in Section 3.4. The size of the context window for the CRF model was selected based on preliminary experiments with the development set. The parameter that achieved the best F-score over that set was chosen. NLTK[3] was used to tokenise the sentence and obtain the POS tags. As our CRF learner, we used CRF++ v0.58.[4] In our experiments, CRF models were either trained on BioScope clinical dataset, VetCompass, or both.

4.2 Baselines

We used NegEx system and LingScope as baselines. LingScope is a Java implementation of the CRF models developed by Agarwal and Yu (2010a,b). It contains models that were pretrained using the BioScope clinical data. Though our CRF model and LingScope were based on the same paper, LingScope differs from our models through the use of a different CRF implementation (using the CRF model provided by the Abner tool (Settles, 2005)), the size of context window used for the classification, and the POS tagger (the Stanford POS tagger).

We used a Python implementation of NegEx.[5] This version of NegEx detects negation scopes to be between a trigger term/phrase identified by NegEx and either a conjunction, start or end of a sentence (which can be longer than the limit of five tokens in the original version of NegEx by Chapman et al. (2001)).

[3] http://www.nltk.org/

[4] https://taku910.github.io/crfpp/

[5] https://code.google.com/archive/p/negex/

4.3 Experimental Setup

Our experiments were based on the fixed split of the corpus described in Section 3.3. We evaluate both the cue detection and scope detection system using precision ($\mathcal{P}$), recall ($\mathcal{R}$) and micro-average F-score ($\mathcal{F}$). Evaluation was performed on a token-level based on whether it is inside or outside of any negation/speculation cue or scope.

We experimented with using different training data to determine whether models trained on out-of-domain data such as BioScope clinical data are suitable for veterinary clinical notes. Since BioScope clinical dataset is much larger than VetCompass, we also experimented with oversampling of instances from the VetCompass training data when both corpora were used for training (at oversampling rates of 1, 2 and 5). When an oversampling rate of 2 is used, we use two duplicates of each VetCompass training record during the training process, and similarly for oversampling rate of 5.

5 Results

Results for negation cue detection and negation scope detection are presented in Table 3 and Table 4, respectively. Results for speculation cue detection and speculation scope detection are presented in Table 5 and Table 6, respectively.

When trained only on BioScope clinical data, the CRF systems (for both cue detection and scope detection) performed worse than their respective baselines. The model only outperforms the baselines when VetCompass training data is used. For negation cue detection and scope detection, incorporating both BioScope clinical data with VetCompass records as training instances helps improves the F-scores for most cases. Further marginal improvements can be achieved with oversampling of the VetCompass training instances as well in most cases.

However, for speculation cue and scope detection, the inclusion of BioScope clinical data with VetCompass training data helps improve the recall but reduces the precision, leading to only marginal improvements in F-scores. Oversampling VetCompass helps to improve the precision, recall and F-score slightly, but the precision is still lower than when the BioScope clinical data was not included in the training set. In both speculation cue detection and scope detection results, the recall is consistently much lower than the precision. The results achieved for speculation detection are also much lower than those achieved for negation detection.

	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
NegEx	73.2	73.2	73.2
LingScope	89.1	71.1	79.1
CRF (VC)	89.3	78.5	83.6
CRF (BIO)	75.2	63.1	68.6
CRF (BIO + VC)	90.2	80.5	85.1
CRF (BIO + VC×2)	89.4	85.2	87.3
CRF (BIO + VC×5)	89.5	85.9	87.7

Table 3: Results for Negation Cue Detection Training data used for CRF models are either BioScope (BIO) and VetCompass (VC) or both

System	Training Set	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
NegEx	—	56.3	75.4	64.4
LingScope (word)	—	79.3	52.4	63.1
LingScope (POS; keep cue)	—	66.8	64.4	65.6
LingScope (POS; replace cue)	—	65.9	62.6	64.2
CRF (word)	VC	87.9	64.4	74.4
	BIO	70.3	57.8	63.4
	BIO + VC	86.8	68.1	76.3
	BIO + VC×2	87.4	68.0	76.5
	BIO + VC×5	88.1	68.3	77.0
CRF (POS; keep cue)	VC	86.6	68.0	76.1
	BIO	78.2	51.1	61.8
	BIO + VC	84.8	71.3	77.5
	BIO + VC×2	85.1	74.3	79.3
	BIO + VC×5	85.5	73.3	79.0
CRF (POS; replace cue)	VC	81.5	67.0	73.6
	BIO	63.6	55.7	59.4
	BIO + VC	82.2	70.7	76.0
	BIO + VC×2	82.4	74.4	78.2
	BIO + VC×5	82.1	73.9	77.8

Table 4: Results for Negation Scope Detection Training data used for CRF models are either BioScope (BIO) and VetCompass (VC) or both

	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
LingScope	43.3	27.6	33.7
CRF (VC)	88.7	44.8	59.5
CRF (BIO)	19.7	24.8	21.9
CRF (BIO + VC)	76.5	49.5	60.1
CRF (BIO + VC×2)	79.7	52.4	63.2
CRF (BIO + VC×5)	81.4	54.3	65.1

Table 5: Results for Speculation Cue Detection Training data used for CRF models are either BioScope (BIO) and VetCompass (VC) or both

5.1 Error Analysis

Unsurprisingly, when the cue detection system does not incorporate VetCompass data, cues that appear only in VetCompass records were usually not detected. For negation, these cues in-

System	Training Set	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
LingScope (word)	—	27.4	27.4	27.4
LingScope (POS; keep cue)	—	35.9	28.9	32.0
LingScope (POS; replace cue)	—	40.6	28.2	33.3
CRF (word)	VC	91.4	27.8	42.6
	BIO	28.1	28.0	28.1
	BIO + VC	78.1	33.5	46.9
	BIO + VC×2	79.9	33.9	47.6
	BIO + VC×5	81.7	35.2	49.2
CRF (POS; keep cue)	VC	80.7	30.2	43.9
	BIO	30.0	18.8	23.1
	BIO + VC	67.7	34.4	45.6
	BIO + VC×2	71.5	37.4	49.1
	BIO + VC×5	73.3	40.9	52.5
CRF (POS; replace cue)	VC	84.2	33.9	48.4
	BIO	27.6	25.8	26.7
	BIO + VC	71.5	40.0	51.3
	BIO + VC×2	72.2	40.9	52.2
	BIO + VC×5	75.0	43.3	54.9

Table 6: Results for Speculation Scope Detection. Training data used for CRF models are either BioScope (BIO) and VetCompass (VC) or both

clude *NAD*, *unable*, and contractions such as *doesn't*. For speculation, these cues include question marks, *poss* and *think*. In speculation cue detection, it was particularly important to have in-domain training data as there are more domain-specific speculation cues.

However, even with VetCompass training data, the cue detection systems (particularly speculation cue detection) still have difficulty detecting all of the cues. Some of this was caused by cue words being misspelled (e.g. *doestn* instead of *doesn't*) or a variant not seen in the training data (such as *susp* for *suspect*). A useful feature could be to use word or string similarity to known cue terms to overcome this issue. Author or patient metadata could also be useful, since some of this is consistent across consultations for a given individual. Such data could be used as additional features for a classifier or by having separate models for different authors/patients.

However, even cues where the form appears in the training data are still sometimes not detected by our system, particularly for speculation cues. This may be because the system was not able to generalise from the limited training data. There was also a greater variety of speculation cues than negation cues. This observation, combined with the smaller proportion of sentences that were speculative, means that there were less training instances for each possible speculation cue.

Both negation and speculation cues also have false-positives that resulted from identifying negation-like or speculation-like terms, such as *not bad*. The speculation cue detection system also often did not detect speculation cues that contained negation-like terms such as *not sure*, while the negation cue detection system incorrectly classifies the *not* in this example as a negation cue.

The errors in cue detection create further errors in the associated scope detection system. However, even with correctly detected cues, the scope detection system still has problems with recall. In most of these cases, the system does not correctly determine one token at the start or end of the scope as being part of it. If the scope is very long, the system will often only detect the first few tokens as being part of the scope and miss the remaining tokens. Scopes where the cues are question marks are also often smaller than the reference annotation, as the system usually only includes the token directly to the left or right of the question mark as part of the speculation scope.

6 Conclusions and Further Work

This paper describes the annotation of a new dataset for negation and speculation detection over veterinary clinical notes. We reimplemented a simple CRF approach for detecting negation and speculation cues and scope, and trained the model over VetCompass training data, BioScope, or both. Our results demonstrated that while datasets such as the BioScope clinical corpus have utility, in-domain training data is often necessary to attain reasonable performance levels, particularly for speculation detection.

Further work will focus on improving the recall of negation and speculation detection systems for veterinary clinical notes. Improving the recall is important for the IR use case that the system will be deployed in. We will also focus on expanding the features used for classification, and experiment with different classifiers. Another focus could be on learning features that are particular to the different authors of notes, and using these to improve negation and speculation detection.

References

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

Shashank Agarwal and Hong Yu. 2010b. Detecting hedge cues and their scope in biomedical text with conditional random fields. *Journal of Biomedical Informatics* 43(6):953–961.

R Michele Anholt, John Berezowski, Iqbal Jamal, Carl Ribble, and Craig Stephen. 2014. Mining free-text medical records for companion animal enteric syndrome surveillance. *Preventive Veterinary Medicine* 113(4):417–422.

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.

Nigel Collier, Hyun Seok Park, Norihiro Ogata, Yuka Tateishi, Chikashi Nobata, Tomoko Ohta, Tateshi Sekimizu, Hisao Imai, Katsutoshi Ibushi, and Junichi Tsujii. 1999. The genia project: corpus-based knowledge acquisition and information extraction from genome research papers. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 271–272.

Noa P Cruz Díaz, Manuel J Maña López, Jacinto Mata Vázquez, and Victoria Pachón Álvarez. 2012. A machine-learning approach to negation and speculation detection in clinical texts. *Journal of the Association for Information Science and Technology* 63(7):1398–1410.

Haibo Ding and Ellen Riloff. 2015. Extracting information about medication use from veterinary discussions. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*. pages 1452–1458.

Marco Duz, John F Marshall, and Tim Parkin. 2017. Validation of an improved computer-assisted technique for mining free-text electronic medical records. *JMIR Medical Informatics* 5(2):e17.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*. pages 282–289.

K Lam, Tim Parkin, Christopher Riggs, and Kenton Morgan. 2007. Use of free text clinical records in identifying syndromes and analysing health data. *Veterinary Record* 161(16):547–51.

Paul McGreevy, Peter Thomson, Navneet Dhand, David Raubenheimer, Sophie Masters, Caroline Mansfield, Tim Baldwin, Ricardo Soares Magalhaes, Jacquie Rand, Peter Hill, Anne Peaston, James Gilkerson, Martin Combs, Shane Raidal, Peter Irwin, Peter Irons, Richard Squires, David Brodbelt, and Jeremy Hammond. 2017. VetCompass Australia: Big data and real-time surveillance for veterinary science. *Animals* 7(10).

Timothy Miller, Steven Bethard, Hadi Amiri, and Guergana Savova. 2017. Unsupervised domain adaptation for clinical negation detection. *BioNLP 2017* pages 165–170.

Roser Morante and Walter Daelemans. 2009a. Learning the scope of hedge cues in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*. Association for Computational Linguistics, pages 28–36.

Roser Morante and Walter Daelemans. 2009b. A metalearning approach to processing the scope of negation. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pages 21–29.

Roser Morante, Anthony Liekens, and Walter Daelemans. 2008. Learning the scope of negation in biomedical texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 715–724.

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.

Jonathon Read, Rebecca Dridan, Stephan Oepen, and Lars Jørgen Solberg. 2012. Sentence boundary detection: A long solved problem? *COLING (Posters)* 12:985–994.

Burr Settles. 2005. Abner: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics* 21(14):3191–3192.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 102–107.

György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. 2008. The bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*. Association for Computational Linguistics, pages 38–45.

Stephen Wu, Timothy Miller, James Masanz, Matt Coarr, Scott Halgrim, David Carrell, and Cheryl Clark. 2014. Negation's not solved: generalizability versus optimizability in clinical natural language processing. *PloS one* 9(11):e112774.