Ir-Man: An Information Retrieval Framework for Marine Animal Necropsy Analysis

Alexander F. B. Carmichael*
Computing Science and Mathematics
University of Stirling
Stirling, UK
a.f.carmichael@stir.ac.uk

Deepayan Bhowmik
Computing Science and Mathematics
University of Stirling
Stirling, UK
deepayan.bhowmik@stir.ac.uk

Johanna Baily
Institute of Aquaculture
University of Stirling
Stirling, UK
j.l.baily@stir.ac.uk

Andrew Brownlow
Scottish Marine Animal Stranding Scheme
Scotland's Rural College (SRUC)
Inverness, UK
andrew.brownlow@sruc.ac.uk

George J. Gunn
Epidemiology Research Unit
Scotland's Rural College (SRUC)
Inverness, UK
george.gunn@sruc.ac.uk

Aaron Reeves
Epidemiology Research Unit
Scotland's Rural College (SRUC)
Inverness, UK
aaron.reeves@sruc.ac.uk

ABSTRACT

This paper proposes Ir-Man (Information Retrieval for Marine Animal Necropsies), a framework for retrieving discrete information from marine mammal post-mortem reports for statistical analysis. When a marine mammal is reported dead after stranding in Scotland, the carcass is examined by the Scottish Marine Animal Strandings Scheme (SMASS) to establish the circumstances of the animal's death. This involves the creation of a 'post-mortem' (or necropsy) report, which systematically describes the body. These semi-structured reports record lesions (damage or abnormalities to anatomical regions) as well as other observations. Observations embedded within these texts are used to determine cause of death. While a cause of death is recorded separately, many other descriptions may be of pathological and epidemiological significance when aggregated and analysed collectively. As manual extraction of these descriptions is costly, time consuming and at times erroneous, there is a need for an automated information retrieval mechanism which is a non-trivial task given the wide variety of possible descriptions, pathologies and species. The Ir-Man framework consists of a new ontology, a lexicon of observations and anatomical terms and an entity relation engine for information retrieval and statistics generation from a pool of necropsy reports. We demonstrate the effectiveness of our framework by creating a rule-based binary classifier for identifying bottlenose dolphin attacks (BDA) in harbour porpoise gross pathology reports and achieved an accuracy of 83.4%.

CCS CONCEPTS

• **Applied computing** → **Bioinformatics**; • **Information systems** → **Ontologies**; **Information extraction**.

KEYWORDS

Information retrieval, marine animal, necropsy analysis, ontology.

ACM Reference Format:
Alexander F. B. Carmichael, Deepayan Bhowmik, Johanna Baily, Andrew Brownlow, George J. Gunn, and Aaron Reeves. 2020. Ir-Man: An Information Retrieval Framework for Marine Animal Necropsy Analysis. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '20), September 21–24, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3388440.3412417

1 INTRODUCTION

Monitoring and surveillance of wildlife is fundamental for the development of understanding of the factors which impact the well-being of populations, species and ecosystems. These activities are especially difficult when applied to the marine mammal domain, as direct observation of living animals in their environments is often impractical. Observation of dead animals, when they become accessible, provides a critical source of data for our knowledge of these populations, and information gathered from such events is particularly important. When a cetacean or pinniped becomes stranded and dies, and when its carcass is examined by trained investigators, the resulting post-mortem (PM) report provides a snapshot of the animal's condition. Collectively, such data provides a unique insight into the general welfare of marine mammal populations and may reveal problems facing the species' environment as a whole.

Williams *et al.* [23, 24], for example have extensively monitored the levels of toxic polychlorinated biphenyls (PCBs) in harbour porpoises. PCB levels are directly affected by human pollution due to the compound's use in some manufactured goods. Similarly, Nelms *et al.* [19] have analysed the presence of microplastics found in stranded cetacea using PM examinations. These examples show that PM examinations can be used to observe the human impact on marine ecology.

*Also with Epidemiology Research Unit, Scotland's Rural College (SRUC).

BCB '20, September 21–24, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-7964-9/20/09…$15.00
https://doi.org/10.1145/3388440.3412417

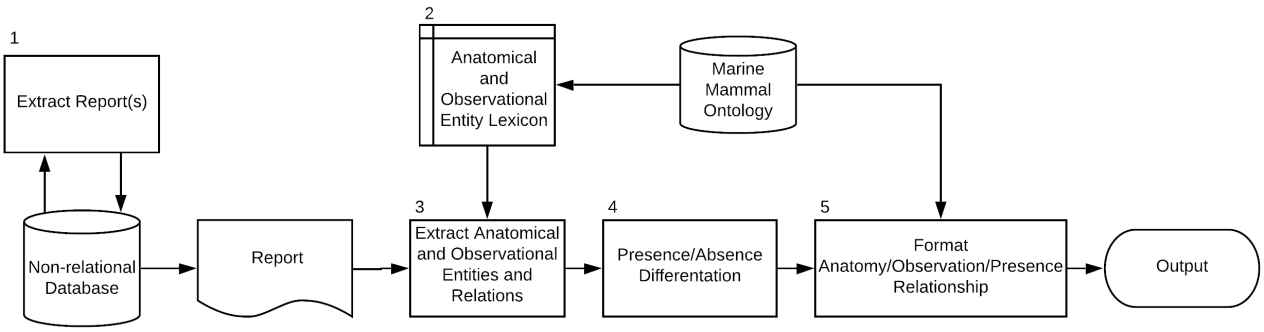

Figure 1: Overview of the Ir-Man framework along with its information retrieval flow diagram.

PM reports are generated in a semi-structured format, with information embedded across multiple free text sections which makes the retrieval of pathological findings a non-trivial problem. Furthermore, there are many cases where multiple indicators are described which relate to multiple distinct pathologies. The ability to effectively confirm or rule out the presence of a pathology based on descriptions of abnormalities would allow for clearer understanding of the problems facing marine mammals. Information retrieval approaches have been applied extensively in human pathology [3, 8, 11, 21, 25, 26], and other animal pathologies [2, 7, 13], however, no work currently exists for marine mammal pathology free text.

To address such gaps we propose Ir-Man (Information Retrieval for Marine Animal Necropsies), a new framework for retrieving discrete information from marine mammal post-mortem (necropsy) reports for statistical analysis. We infer that the hierarchical nature of descriptive terms used in marine mammal PM reports can be represented within an ontology, in which a class of terms represent a specific observation, and subclasses represent distinct varieties of observations. Our approach involves the extraction of a number of key pieces of information: the term used to describe particular observations; whether the context indicates presence or absence; and the anatomical region to which the observation relates. Once extracted, these fields can be used to create a deterministic classifier based on the presence or absence of either general pathological indicators or indicators of pathologies in specific anatomical regions. We also extract a reference that can be used to link and aggregate information across different report types. The overview of Ir-Man framework is depicted in Figure 1.

In building the framework, this paper considers the marine mammal PM reports, particularly the gross pathology reports and describes various framework components for extraction of gross pathology findings. In evaluating the effectiveness of Ir-Man, we experimented with an exemplar use case of identifying bottlenose dolphin attacks on harbour porpoises and reported our findings. Our main contributions are:

- We propose a new framework for information retrieval for marine mammal necropsy analysis using an ontology driven entity relation approach.
- We design three ontologies which contain terms relevant to cetacean gross pathology reports that are based on observations, anatomy, and pathology, respectively.
- We develop a lexicon based entity-relationship engine that can record the presence or absence of observations which can confirm or rule out pathologies.
- We measure the effectiveness of our retrieval approach by creating and evaluating a deterministic classifier for cases of bottlenose dolphin attacks (BDAs) on harbour porpoises.

2 RELATED WORK

Our framework encompasses several information retrieval methods and fields. While current literature covers a relatively large number of information retrieval approaches in the biomedical domain (particularly on human health), only a handful of attempts were made in the veterinary domain. We critically analysed both domains separately in order to position our proposed framework appropriately.

2.1 Information Retrieval in Biomedical Domain

Related approaches for information extraction in the biomedical domain are numerous. Chapman *et al.* [3] proposed NegEx, a tool for determining the presence or absence of clinical findings in discharge summaries. Their approach was used to analyse $76,049$ screening and $17,656$ diagnostic mammography reports. Even though this approach extracts conclusions, rather than observations – which are the focus of our work - the applications are similar. More recently Gao *et al.* [8] extracted a number of features from mammographic reports: mass, calcification, asymmetry and architectural distortion.

Friedlin *et al.* [6] developed Medical Exploratory Data Analysis over Text (MEDAT), a text analytics system for medical domains and demonstrated their system on radiology reports. Comelli *et al.* [4] also applied text mining to radiology reports. They leveraged the entity relationships represented in their radiology ontology, to extract relevant medical terms from mammographic reports. While their approach was exhaustive within the mammographic domain, the texts were in Italian and the application domain is far more specific than that of marine mammal gross pathology.

Gong *et al.* [11] developed a biomedical information retrieval approach for terminologies related to breast cancer. Their approach

involves entity extraction, entity relationship identification, and visualisation. Entities are extracted based on conditional random fields while entity relationships were extracted using co-occurrence statistics. Sudeshna *et al.* [21] aimed to identify symptoms and treatments of heart disease using a machine learning based approach. Based on suggested identified symptoms, texts would be classified into treatments. Zhao *et al.* [26] created CausalTriad, an approach toward the discovery of pseudo causal relationships between entities. They evaluated their approach on HealthBoards message board data and Traditional Chinese Medicine data. Yang *et al.* [25] used an ontology-based text mining approach for the extraction of data from Chinese EMRs. This work focused on the mining of stroke cases. Gero and Ho [9] proposed NamedKeys, a keyphrase extraction approach which they evaluated on PubMed abstracts. They also describe a benchmark dataset for biomedical keyphrase extraction.

While a variety of different clinical text types have been the subject of such research, there is also a wide body of research into automated biomedical literature reviews. Navathe [18] used UMLS (Unified Medical Language System) [1] and a gene ontology to represent biomedical concepts, and an SVM to classify literature from the Centre of Disease Control (CDC) based on relevant keywords. Mala *et al.* [15] researched the use of ontology in semantic medical text mining with WordNet. Gong *et al.* [10] used a dictionary-based approach to extract biomedical concepts from literature. This was done using an algorithm called the Variable-step Window Identification Algorithm (VWIA), matched terms to biomedical entities using POS tagging and organisation based on phrasing. Their technique was applied to 10 Medline abstracts and produced promising results. Mate *et al.* [16] focused on creating a process of extraction, transformation and loading (ETL) of electronic medical records.

Although not used in our approach, it should be noted that emerging *deep learning* has become popular in the biomedical domain with neural network based methods being used to enhance text mining techniques [12, 22].

2.2 Information Retrieval in Veterinary Domain

All of the reports listed above applied information retrieval techniques to biomedical text pertaining to humans. In the veterinary domain, Bollig *et al.* [2] used a machine learning based approaches for extraction of different pathologies from free text. Furrer *et al.* [7] built a text mining tool for veterinary surveillance by linking terms identified in necropsies to existing ontologies. Küker *et al.* [13] later used this tool to analyse pig and cattle necropsies and found that free text necropsy reports are a valuable resource for animal health surveillance.

At present no work exists on information retrieval from marine mammal necropsy reports. Given the importance of PMs in furthering understanding of marine mammals and marine ecology more generally, an information retrieval framework that can aggregate observations for statistical and epidemiological analysis would be especially useful.

3 THE FRAMEWORK

In developing the proposed Ir-Man framework, we consider a number of steps that are involved in the extraction of observations from marine mammal necropsies. Firstly, free text is pulled from necropsy documents, and then individual reports sections (*i.e.*, the gross pathology report section and if applicable, the histopathology and bacteriology report sections) are extracted. Text is divided into sentence tokens before individual words are tagged based on part-of-speech. Entities and the relationships between them are then grouped using a feature-based grammar. Each entity is then checked against our lexicon of anatomical, pathological and observational terms which is generated using our ontology. Presence or absence of a described feature is then established by checking for negation. It is fundamental to record negative occurrences of identifiers (absences) as well as positive occurrences, (presence) as both can be leveraged in a deterministic classification system. The overall retrieval process is outlined in Figure 1 and the pseudo-code in Algorithm 1 along with the description of each framework components below.

Algorithm 1: Information retrieval pipeline. Output of the entity-relationship extraction engine is used to identify observations, attributed anatomies and detect negation.

```
Result: relationships
sentences = sentenceTokenisation(text);
observations;
while not at end of sentences do
    RELChunkedSentence = preprocess(sentence);
    identifyNamedEntities(RELChunkedSentence);
    while not at end of sentences do
        if No Observational Entities then
            break to next relationship;
        end
        if Observational Entity AND No Anatomical Entity
         then
            observations <- 'unattributed' observation;
            break to next relationship
        end
        if Observational Entity and Anatomical Entity then
            observations <- anatomy, observation;
        end
    end
end
while not at end of observations do
    negatedObservation <- mark_negated(observation) ;
    if observation == negatedObservation then
        presence <- True;
    else
        presence <- False;
    end
end
```

3.1 Data Set: Marine Mammal Stranding Reports

The data used in this project was generated using PM reports of cetacea produced by the Scottish Marine Animal Stranding Scheme (SMASS) between 2012 and 2019. When generating these reports, the pathologist records features of the carcass, including condition, morphology, pathological lesions and observations. Information relating to the body's condition usually refers to the level of autolysis

Body condition: Fat

External examination
Body orifices: NAD
Ectoparasites: NAD – None seen
Fins and flukes: NAD – Intact, no rake marks

Integument
Epidermis: Rake marks over left flank/tailstock. Severe scavenger damage at right side of head
Blubber: NAD – Good layer, not jaundiced
Subcutaneous tissue: Bruising over lateral spinous processes and right side head region
Mammary glands: NE

Figure 2: A gross pathology extract from a harbour porpoise necropsy report.

or physical damage to the remains. Morphometric measurements, such as blubber thickness and body length, are also taken. These features all help inform the pathologist of a probable cause of death. PM reports include gross pathology reports, which describe, in detail, the characteristics of the body as a whole, and those of specific anatomies. The final PM report then contains a number of sections: basic information (sex, date, location etc.); morphometric data; a gross pathology; bacteriology and histopathology reports, where applicable; and a conclusion which includes comments, cause of death, and an indicator of confidence in diagnosis.

The material analysed for this paper consists of 193 gross pathology reports on harbour porpoises (*Phocoena phocoena*). This species was chosen for a number of reasons: the relatively high number of reports produced by SMASS on harbour porpoises; easy future integration of other cetacean species due to transferability of the harbour porpoise anatomy; and the prevalence of BDAs listed as the cause of death, which allows us to establish the suitability of our framework for detecting exhibited pathologies. Bottlenose dolphins are known to violently attack harbour porpoises, usually leaving parallel incisions which are referred to as 'rake marks'. It is these rake marks which are used as a primary indicator of a BDA, and as such, the use of the term is relatively consistent, making it a good candidate for evaluating the effectiveness of our approach.

While the language used in these gross pathology reports is specialised, there is some structure to the reports which can be leveraged. An anatomical region of interest will often be used as a heading followed by a free text description. This can be seen in Figure 2. Acronyms such as NAD (no abnormalities detected) and NE (not examined) are also important and distinct. One can rule out some pathological conditions when no abnormality is detected, but not when a region has not been examined.

3.2 Gross Pathology Report Extraction

The SMASS post-mortem reports were stored in Microsoft Word Open XML Format (DOCX) files. We parsed documents and stored fields in a non-relational MongoDB[1] database. Where applicable, specific text fields were extracted by searching for field names which where indicative of a field's presence. An example would the species field, where we used the string "SPECIES:" as the field indicator and the string following it in the line as the field to be extracted (*e.g.*, "*delphinus delphis*"). When a field was left blank, no value was stored in the database. We normalised fields by grouping synonymous terms. For example, the case of the species field this involved pairing the scientifc names (*e.g.*, "*delphinus delphis*") with their corresponding common names (*e.g.*, "short-beaked common dolphin"). Free text sections such as the gross pathology reports were obtained by identifying relevant section headers and extracting the text between them. When the space between section headers consisted only of white-space or short strings such as "Not examined", the section was not extracted. All extracted fields and sections were stored in a local MongoDB database.

3.3 Ontology Development

The framework uses our bespoke ontologies to organise terms, and to provide context that would otherwise be unavailable. While multi-species ontologies such as Uberon [17] do exist, it was decided that a smaller more manageable ontology would be more appropriate for this task. We identified three main branches of relevant terminologies for our purposes. The first is a representation of anatomy, where classes represent different anatomical regions. The second is the pathology ontology which was used to record different conditions which can be represented in PM reports. The third is the observation ontology, which groups terms into classes and sub-classes where children represent an extra degree of specification that may not apply to all within the parent class. For all classes a representative label is stored in the "rdfs:label" annotation, and manually generated synonymous terms are stored in our own "synonym" annotation. Ontologies were developed using Protégé [20] (shown in Figure 3) and stored in RDF/XML format.

3.3.1 Observation Ontology. The observation branch of the ontology makes use of the semantic relationships between terms. When terms are very similar semantically, but one gives a greater degree of specification, the more specific term is considered a child of the other. For example, reports may specify that "fluid" or "brown fluid" is present. Not all fluid is brown fluid, so a parent-child relationship is created between the terms. This allows for distinctions between different types of fluids and their descriptions such as mucoid, protein-rich or amniotic fluids. The ontology was populated manually by producing lists of terms from the reports based frequent unigrams, bigrams, and trigrams, as well as collocations using pointwise mutual information (PMI). Previously established anatomical and disease related terminologies were filtered to accelerate the process. The structure of the observational ontology is shown in Figure 4.

3.3.2 Pathology Ontology. The pathology ontology (shown in Figure 5) is used to represent different conditions and the semantic relationships between them. This was initially created using the diseases or conditions listed as a cause of death within the SMASS database. These were also mined from the reports using known target strings

[1] https://www.mongodb.com/

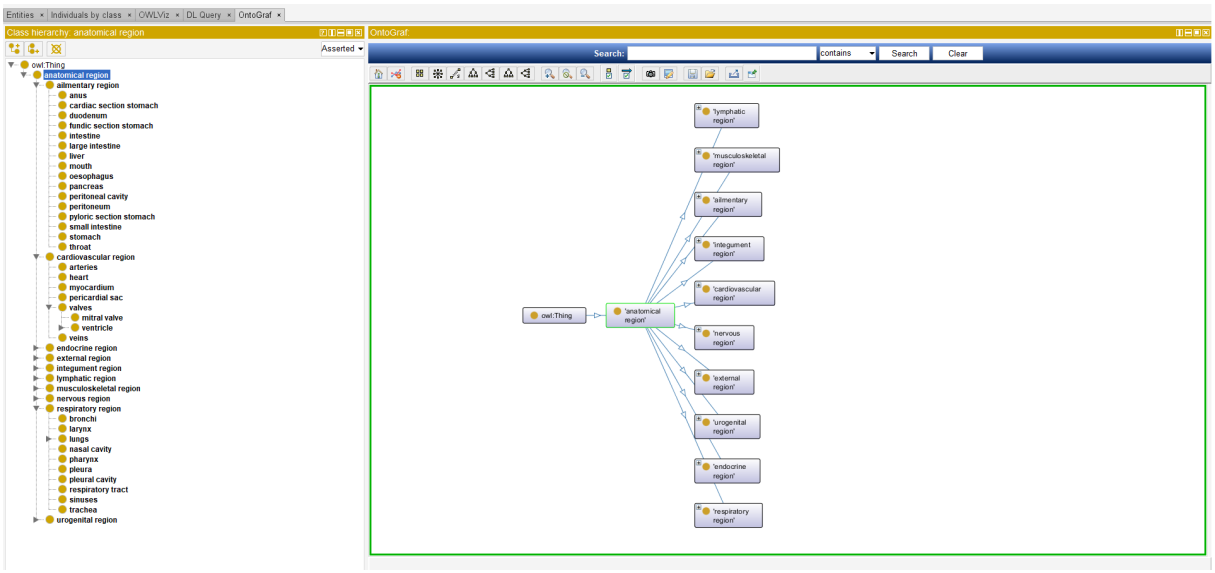

Figure 3: Screenshot of Protégé IDE for ontology development. The OntoGraf plugin [5] was used for ontology visualisation.

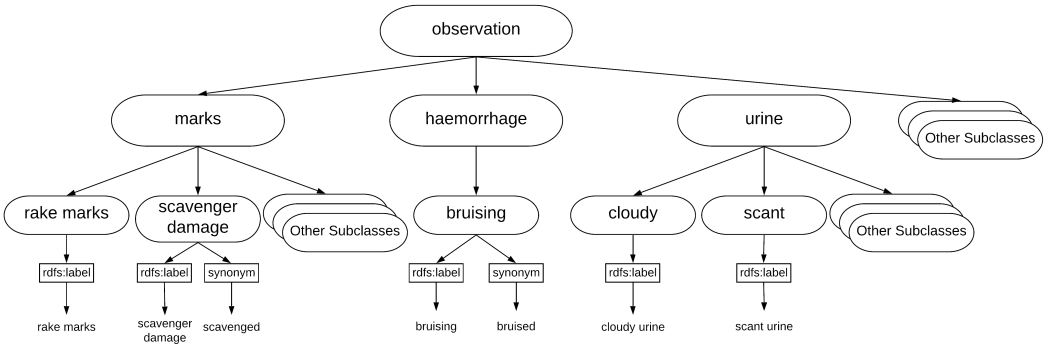

Figure 4: Structure of the observation ontology demonstrated using the marks, haemorrhage and urine class examples.

that precede causes of death. They were then categorised based on semantic similarity. For example, the "physical trauma" class represent cases where there is evidence of blunt force or penetration to the skin which appear to have been detrimental to the animal's health. This category captures conditions such as boat strikes, bottlenose dolphin attack trauma and entanglement (where rope, line or netting has wrapped around the animal).

3.3.3 Anatomy Ontology. Finally, the anatomy ontology was created based on the anatomical terms which were used to convey observations within the reports. The highest level of the "anatomical region" tree contains classes which relate to different organ systems within the body, or anatomical regions which are semantically linked. The latter situation applies, for example, to the "integument region" (relating to the skin) and the "external region" which mostly refers to external observations out with the main scope of those captured in the integument class.

The next level of subclasses generally represents different types of these regions. The decision was taken to make a distinction between having a parent-child relationship and an "isPartOf" attribute: it is not accurate, for example, to represent regions such as "the left valve of the heart" as a subclass of "heart", but it is still desirable to capture the relationship between these two regions. The "isPartOf" object property is transitive and asymmetric. This allows for instances such as the duodenum to be more accurately represented. The duodenum "isPartOf" the small intestine, and the small intestine "isPartOf" the intestines. Therefore, we can deduce that the duodenum "isPartOf" the intestines also. The structure of this ontology is shown in Figure 6, which shows some example anatomies in the ailmentary system. The anatomy ontology was manually populated and structured based on the headings used for sections of gross pathology reports (shown in Figure 2).

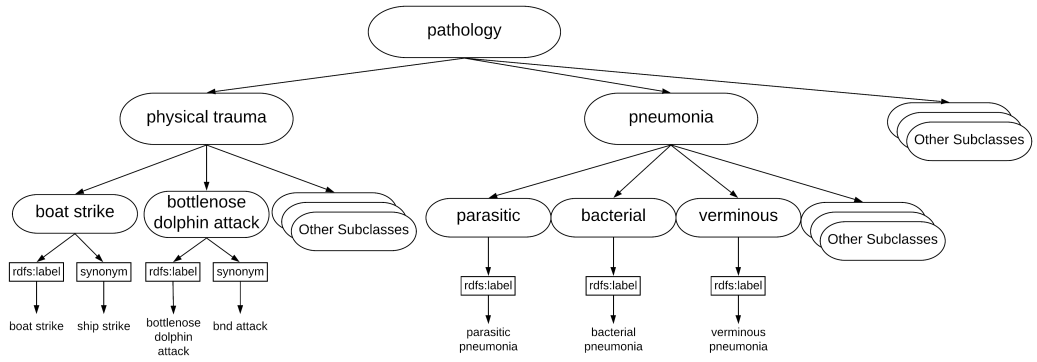

Figure 5: Structure of the pathology ontology demonstrated using the physical trauma and pneumonia class examples.

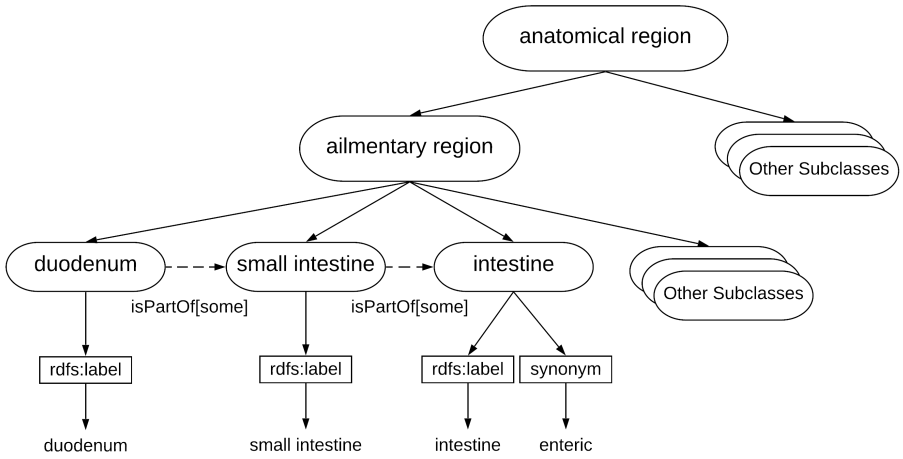

Figure 6: Hierarchical structure pf anatomy ontology demonstrated using example ailmentary region subclasses.

3.4 Information Retrieval

Information retrieval consists of several individual components within the framework including, *(a)* lexicon, *(b)* entity-relationship extraction engine, *(c)* anatomy, observation and presence recognition and *(d)* formatting to extract anatomical features, observations and pathologies. This pipeline is shown in the flowchart in Figure 1 as well as Algorithm 1.

3.4.1 Lexicons. Our anatomy, pathology and observation ontologies (refer Section 3.3) are used to identify entities. Two lexicons of key terms are generated by parsing the three ontology xml files and extracting their "rdfs:label", and "synonym" attributes. The observation lexicon was created using the observation and pathology ontologies, while the anatomy lexicon was created using the anatomy ontology. Pathological terms are incorporated in the observation lexicon because they can used to both represent the pathology of the specimen as a whole, and the condition of anatomical region. An example would be a case of physical trauma cause by entanglement which is a subclass of physical trauma within the pathology ontology. The inclusion of the term "entangled" could be treated as both an observation and pathology based on the representation within our ontology.

3.4.2 Entity-Relationship Extraction Engine. Reports are first segmented at the sentence level. As shown in Figure 2, sections are not always delimited by a full stop as one would expect. As such, we assume that sentences could also be delimited by new line characters ('\n'). Word level tokenisation is also performed. Words and punctuation are tagged using NLTK's POS Tagger library [14]. Words are then grouped together through "Noun Phrase Chunking" (NP-chunking). Empirically, we developed a simple feature-based grammar of tag patterns which represent entities and entity-relationships. Our grammar is passed into NLTK's *Regexparser* library to create chunks of entities and entity-relationships. The regular expression based grammar we defined for this task can be seen below.

```
NP: {<DT>?<JJ>*<VB.*>*<JJ>*<NN.*>+}
NP: {<NP><CC><NP>}
NP: {<VBD|VBN>}
NP: {<CD><RB>}
```

```
NP: {<NP><NP>}
IN: {<IN>}
REL:{<NP><IN><NP>}
REL:{<NP><TO><NP>}
REL:{<NP><:><NP>}
```

Entity chunks are grouped together as noun phrases (NP). The first rule captures any case where there is at least one noun preceded by any adjectives (JJ) or verbs (VB) and may include a determiner ("the", "a" etc.) denoted by 'DT'. If any past tense or past participle verbs are used separately, they is also chunked as a noun phrase to account for cases such as "right eye: scavenged". Lastly, NPs can be linked into a single NP where they are separated by coordinating conjunction terms such as 'and'.

NPs are then linked together into relationship (REL) chunks based on several conditions. Simple adjacency of two NPs is the first relationship as proximal entities are likely to relate. Prepositions (*e.g.*, 'in') were also of particular interest as they represent a relationship between that which precedes and follows them. The word 'to' is another good link between NPs given that phrases such as "damage to left flank" are very common. Lastly, we use the colon to capture cases where the anatomical entity is stated, then observations follow. An example of this is shown in Figure 2: "Blubber: NAD". This grammar is designed to capture relatively simple expressions, but can be expanded to incorporate more complex entity-relationships in future.

3.4.3 Anatomy, Observation and Presence Recognition. For each sentence in a report, each REL chunk is parsed and compared to the anatomical and observational lexicons. Where a NP chunk contains a sub-string that occurs in either lexicon, it is identified as anatomy or observation accordingly.

The implementation deliberately only incorporates NP - NP relationships (as defined in our grammar) as when only the relation subsection of the sentence is used for the marking of negated terms, one reduces the number of falsely negated terms. This means a relatively simple process for identifying negated words can be used, to a high degree of accuracy.

We use the NLTK mark_negated package for this purpose. The package adds '_NEG' as a suffix to any word between a negation and certain punctuation marks. For each REL chunk with an identified observation, a negated version of the statement is generated. The NP chunk containing the free text representation of the observation is compared to the same chunk after negated terms are marked. In the event an observational term is negated, it is considered to be an absent case. An example would be "no obvious rake marks on flank", where rake marks would be identified as an observational entity. When compared to the negation marked version of the text ("there are no obvious_NEG rake_NEG marks_NEG on_NEG flank_NEG") the negation of the observation would become apparent. In this event "rake marks" would be identified as "absent". The benefit of this approach is that one only marks negated terms at the relationship level. If one were to mark at the sentence level, unrelated negated terms would incorrectly cause for a classification of absence rather than presence.

When a recognised observational term is not attributed to an anatomical entity, it is still recorded as either present or absent and is not attributed to an anatomical region. There are a number of reasons why an anatomical entity might not be identified: the term used is not represented within the anatomical ontology; the observation is not used in relation to an anatomical entity; or the grammar used for chunking fails to capture relevant NP chunks within a relationship.

3.4.4 Formatting Findings. The information extracted is summarised in a dictionary implemented in Python, which can then be used for analysis or classification systems. The anatomy and observation terms are represented as strings, while the presence or absence of an observation is stored as a Boolean value. Some examples are shown below:

```
{
    'anatomy': 'right pectoral fin',
    'observation': 'scavenger damage',
    'presence': True
}
{
    'anatomy': 'epidermis',
    'observation': 'rake marks',
    'presence': False
}
{
    'anatomy': 'skull',
    'observation': 'nad',
    'presence': True
}
```

4 USE CASE, RESULTS AND ANALYSIS

To analyse the effectiveness of our approach, we identified observations which could negate or confirm a chosen pathological finding, and used the presence or absence of these observations as a means of classification. Bottlenose Dolphin attacks (BDA) on harbour porpoises are a very common cause of death within the dataset, with 50 of the 193 cases listing BDA as the key finding in the SMASS database.

A deterministic classifier was created using extracted empirical observations. The classes chosen were either explicit mentions of BDA, or strong indicators such as "rake marks". Where BDA is mentioned, it's absence or presence is sufficient to classify the case as "Non-BDA" or "BDA". The observation "rake marks", however, can also be used to describe some grey seal attacks (GSA). As such, we then filter all observations relevant to seal attacks and claw marks (an indicator of a GSA). We make the assumption that if there is evidence of both a GSA and a BDA, an explicit mention of BDA should be found. If the document has not been classified using these rules, presence of "rake marks" results in "BDA" classification. The deterministic classifier's sequence of decisions is shown in Algorithm 2. The cause of death stored in the SMASS database was used as ground truth for classifier evaluation.

The results for the Bottlenose Dolphin attacks use case are shown in Table 1. The approach achieved an overall accuracy of 83.4% and an F1-score of 0.83. BDA classification achieved a precision of 0.70, a recall of 0.64, and F1-score of 0.67. Non-BDA classification achieved a precision of 0.88, a recall of 0.90, and F1-score of 0.89. Of the 193 reports used, 50 were cases of BDA and 143 were Non-BDA cases based on the cause of death stored in the SMASS database.

Algorithm 2: Deterministic BDA classification process based on presence or absence of observations.

```
Result: prediction
if Any observation is a BDA term then
    if observation present then
        prediction <- "BDA";
    else
        prediction <- "Non-BDA";
    end
    return prediction;
else
    if Any present observation is a GSA or claw mark term
     then
        prediction <- "Non-BDA";
        return prediction;
    else
        if Any present observation is a rake mark then
            prediction <- "BDA";
            return prediction;
        end
        prediction <- "Non-BDA";
        return prediction;
    end
end
```

Metrics	Cumulative	BDA	Non-BDA
Accuracy	0.83	-	-
ROC-AUC	0.77	-	-
Recall	-	0.64	0.90
Precision	-	0.70	0.88
F1-score	0.83	0.67	0.89
Support	193	50	143

Table 1: BDA classifier performance evaluation metric scores.

A Receiver Operating Characteristic (ROC) curve (Figure 7) was generated using the BDA precision and recall values listed above which achieved the Area Under Curve (AUC) score of 0.77. The disadvantage of labelling based on cause of death is that there are many instances where a BDA has occurred but a separate finding has been identified as the cause of death. This leads to some false positive (FP) classifications as BDA terms and indicators are still described. This can be seen in the confusion matrix in Figure 8. Given the deterministic nature of our classifier, there are three possible causes of incorrect classifications. The first is the presence of a separate more significant finding which caused death, even though a BDA occurred; the second is the use of an significant term outwith the scope it was intended; and the third is that a significant finding is not successfully identified by the entity-relationship engine.

When analysing cases of FPs, some statements such as "rake marks, assumed bird" lead to an incorrect detection. This reflects

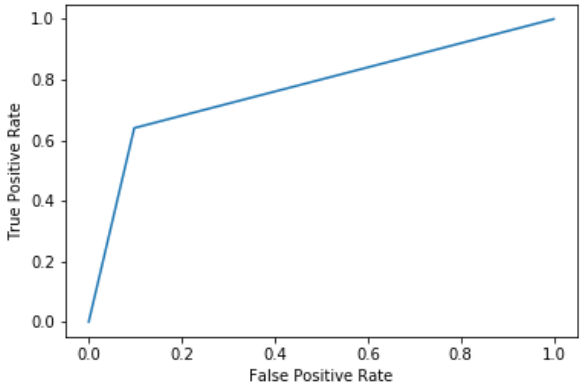

Figure 7: ROC-AUC curve of BDA classifier predictions. AUC = 0.771

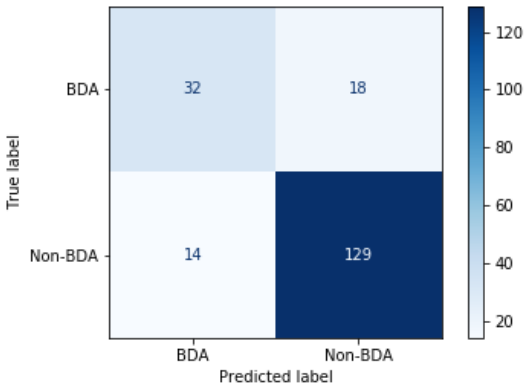

Figure 8: Confusion matrix of BDA classifier predictions.

the difficulty of varying terminology usage between pathologists. Several FPs included instances where BDA rake marks were "healed" or "healing". Some false negatives (FNs) were caused by the lack of explicit mentions of BDAs and there being "no obvious rake marks". This suggests that there were other indicators of BDA despite the absence of rake marks.

The recall (0.64) and precision (0.70) scores of the BDA classifications are relatively low due to the simplicity of the feature-based grammar, and the total number of true positives (TPs) being underestimated. This being said, analysis of FPs and FNs also showed some cases where significant phrases were not captured by the feature-based grammar in the entity-relationship engine.

The precision and recall scores associated with Non-BDAs (0.88 and 0.90 respectively) are considerably higher. When analysed, it was found that several cases were correctly identified as Non-BDA due to the exclusion of GSAs in the deterministic classifier. This shows that the creation of an inclusion/exclusion based determiner can be used to increase trust in positive classifications, meaning any insight obtained is more robust to scrutiny. While the classifier had

minor shortcomings due to the grammar used, the results are very promising for future work. Using cause of death as a label leads to lower performance metrics than anticipated. The use of a manually labelled dataset would naturally produce more realistic results, but in its absence, we can still get a good understanding of the classifier characteristics.

Another thought to consider is that one necropsy report contains many other fields pertaining to morphology, confidence in diagnosis and other free text sections such as the histopathology report and conclusion sections. By incorporating relevant fields and applying a similar information retrieval process to other free text sections, more accurate, complex, and inclusive determiners would be defined, meaning a higher confidence in positive or negative results.

5 CONCLUSIONS

We proposed Ir-Man, an information retrieval framework for marine animal necropsy analysis. The framework applied and adapted information retrieval techniques to reports in a previously unexplored domain. Necropsy reports of stranded marine mammals provide a unique insight into marine ecology; the ability to access and aggregate this information will allow for more useful epidemiological analysis. Despite the challenges associated with mining semi-structured gross pathology reports, our information retrieval framework achieved a baseline accuracy of 83.4% when classifying BDAs on harbour porpoises. Future work will include the incorporation of more complex feature-based grammar representations which would identify structure within text more effectively; expanding the ontologies to incorporate other cetaceans; and defining and detecting further nuance between different observation classes. Most importantly, the framework will be used to further pathological and epidemiological understanding within the marine mammal domain.

ACKNOWLEDGEMENT

We acknowledge the support of a joint PhD studentship by University of Stirling and Scotland's Rural College (SRUC).

REFERENCES

[1] Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research* 32 Database issue (2004), D267–70.

[2] Nathan Bollig, Lorelei Clarke, Elizabeth Elsmo, and Mark Craven. 2020. Machine learning for syndromic surveillance using veterinary necropsy reports. *PloS one* 15, 2 (2020), e0228105.

[3] 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 (2001), 301–310.

[4] Albert Comelli, Luca Agnello, and Salvatore Vitabile. 2015. An ontology-based retrieval system for mammographic reports. In *IEEE Symposium on Computers and Communication (ISCC)*. IEEE, 1001–1006.

[5] Sean Falconer. 2010. Ontograf protege plugin. *Place: Available at: http://protegewiki. stanford. edu/wiki/OntoGraf [Accessed: 21/03/2014]* (2010).

[6] Jeffrey Friedlin, Malika Mahoui, Josette Jones, and Patrick Jamieson. 2011. Knowledge discovery and data mining of free text radiology reports. In *IEEE First International Conference on Healthcare Informatics, Imaging and Systems Biology*. IEEE, 89–96.

[7] Lenz Furrer, Susanne Küker, John Berezowski, Horst Posthaus, Flavie Vial, Fabio Rinaldi, Thierry Poibeau, and Pamela Faber. 2015. Constructing a syndromic terminology resource for veterinary text mining. (2015).

[8] Hongyuan Gao, Erin J Aiello Bowles, David Carrell, and Diana SM Buist. 2015. Using natural language processing to extract mammographic findings. *Journal of biomedical informatics* 54 (2015), 77–84.

[9] Zelalem Gero and Joyce C Ho. 2019. NamedKeys: Unsupervised Keyphrase Extraction for Biomedical Documents. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. 328–337.

[10] Lejun Gong, Jie Yan, Jiacheng Feng, and Ronggen Yang. 2015. A dictionary-based approach to identify biomedical concepts. In *12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*. IEEE, 1091–1095.

[11] Lejun Gong, Ronggen Yan, Quan Liu, Haoyu Yang, Gene Yang, and Kaiyu Jiang. 2016. Extraction of biomedical informtion related to breast cancer using text mining. In *12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*. IEEE, 801–805.

[12] Donghyeon Kim, Jinhyuk Lee, Chan Ho So, Hwisang Jeon, Minbyul Jeong, Yonghwa Choi, Wonjin Yoon, Mujeen Sung, and Jaewoo Kang. 2019. A neural named entity recognition and multi-type normalization tool for biomedical text mining. *IEEE Access* 7 (2019), 73729–73740.

[13] Susanne Küker, Celine Faverjon, Lenz Furrer, John Berezowski, Horst Posthaus, Fabio Rinaldi, and Flavie Vial. 2018. The value of necropsy reports for animal health surveillance. *BMC veterinary research* 14, 1 (2018), 191.

[14] Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. *arXiv preprint cs/0205028* (2002).

[15] Vajenti Mala and DK Lobiyal. 2015. Concepts extraction for medical documents using ontology. In *International Conference on Advances in Computer Engineering and Applications*. IEEE, 773–777.

[16] Sebastian Mate, Felix Köpcke, Dennis Toddenroth, Marcus Martin, Hans-Ulrich Prokosch, Thomas Bürkle, and Thomas Ganslandt. 2015. Ontology-based data integration between clinical and research systems. *PloS one* 10, 1 (2015).

[17] Christopher J Mungall, Carlo Torniai, Georgios V Gkoutos, Suzanna E Lewis, and Melissa A Haendel. 2012. Uberon, an integrative multi-species anatomy ontology. *Genome biology* 13, 1 (2012), R5.

[18] Shamkant B Navathe. 2007. Text Mining and Ontology Applications in Bioinformatics and GIS. In *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*. IEEE, xviii–xix.

[19] Sarah E Nelms, James Barnett, Andrew Brownlow, NJ Davison, Rob Deaville, Tamara S Galloway, Penelope K Lindeque, D Santillo, and Brendan J Godley. 2019. Microplastics in marine mammals stranded around the British coast: ubiquitous but transitory? *Scientific Reports* 9, 1 (2019), 1–8.

[20] Natalya Fridman Noy, Monica Crubézy, Ray W Fergerson, Holger Knublauch, Samson W Tu, Jennifer Vendetti, and Mark A Musen. 2003. Protégé-2000: an open-source ontology-development and knowledge-acquisition environment.. In *AMIA... Annual Symposium proceedings. AMIA Symposium*, Vol. 2003. American Medical Informatics Association, 953–953.

[21] P Sudeshna, S Bhanumathi, and MR Anish Hamlin. 2017. Identifying symptoms and treatment for heart disease from biomedical literature using text data mining. In *International Conference on Computation of Power, Energy Information and Commuincation (ICCPEIC)*. IEEE, 170–174.

[22] Hao Wei, Mingyuan Gao, Ai Zhou, Fei Chen, Wen Qu, Chunli Wang, and Mingyu Lu. 2019. Named entity recognition from biomedical texts using a fusion attention-based BiLSTM-CRF. *IEEE Access* 7 (2019), 73627–73636.

[23] Rosie Williams, Mariel ten Doeschate, David J Curnick, Andrew Brownlow, Jonathan L Barber, Nicholas J Davison, Robert Deaville, Matthew Perkins, Paul D Jepson, and Susan Jobling. 2020. Levels of Polychlorinated Biphenyls Are Still Associated with Toxic Effects in Harbor Porpoises (Phocoena phocoena) Despite Having Fallen below Proposed Toxicity Thresholds. *Environmental Science & Technology* 54, 4 (2020), 2277–2286.

[24] Rosie S Williams, David J Curnick, Jonathan L Barber, Andrew Brownlow, Nicholas J Davison, Rob Deaville, Matthew Perkins, Susan Jobling, and Paul D Jepson. 2020. Juvenile harbor porpoises in the UK are exposed to a more neurotoxic mixture of polychlorinated biphenyls than adults. *Science of the Total Environment* 708 (2020), 134835.

[25] Yujie Yang, Yunpeng Cai, Wenshu Luo, Zhifeng Li, Zhenghui Ma, Xiaolu Yu, and Haibo Yu. 2013. An ontology-based approach for text mining of stroke electronic medical records. In *IEEE International Conference on Bioinformatics and Biomedicine*. IEEE, 288–291.

[26] Sendong Zhao, Meng Jiang, Ming Liu, Bing Qin, and Ting Liu. 2018. CausalTriad: toward pseudo causal relation discovery and hypotheses generation from medical text data. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. 184–193.