Automated Georeferencing of Antarctic Species

Jamie Scott
Massey University, Auckland, New Zealand

Kristin Stock
Massey Geoinformatics Collaboratory, Massey University, Auckland, New Zealand

Fraser Morgan
Manaaki Whenua Landcare Research, Auckland, New Zealand

Brandon Whitehead
Manaaki Whenua Landcare Research, Auckland, New Zealand

David Medyckyj-Scott
Manaaki Whenua Landcare Research, Auckland, New Zealand

Abstract

Many text documents in the biological domain contain references to the toponym of specific phenomena (e.g. species sightings) in natural language form "In <LOCATION> Garwood Valley summer activity was 0.2% for <SPECIES> Umbilicaria aprina and 1.7% for <SPECIES> Caloplaca sp. ..."

While methods have been developed to extract place names from documents, and attention has been given to the interpretation of spatial prepositions, the ability to connect toponym mentions in text with the phenomena to which they refer (in this case species) has been given limited attention, but would be of considerable benefit for the task of mapping specific phenomena mentioned in text documents.

As part of work to create a pipeline to automate georeferencing of species within legacy documents, this paper proposes a method to: (1) recognise species and toponyms within text and (2) match each species mention to the relevant toponym mention. Our methods find significant promise in a bespoke rules- and dictionary-based approach to recognise species within text (F1 scores up to 0.87 including partial matches) but less success, as yet, recognising toponyms using multiple gazetteers combined with an off the shelf natural language processing tool (F1 up to 0.62).

Most importantly, we offer a contribution to the relatively nascent area of matching toponym references to the object they locate (in our case species), including cases in which the toponym and species are in different sentences. We use tree-based models to achieve precision as high as 0.88 or an F1 score up to 0.68 depending on the downsampling rate. Initial results out perform previous research on detecting entity relationships that may cross sentence boundaries within biomedical text, and differ from previous work in specifically addressing species mapping.

2012 ACM Subject Classification Computing methodologies → Information extraction; Computing methodologies → Classification and regression trees; Applied computing → Life and medical sciences

Keywords and phrases Named Entity Recognition (NER), Taxonomic Name Extraction, Relation Extraction, Georeferencing

Digital Object Identifier 10.4230/LIPIcs.GIScience.2021.II.13

Funding *Jamie Scott*: Research for this project was funded by Manaaki Whenua Landcare Research.

1 Introduction

A significant amount of biodiversity knowledge is locked up in textual descriptions within documents, often in free text form, without coordinates with which to georeference. Many research papers in the biological domain refer to specific species and their location, and the development of methods to extract species-toponym pairs can enable biodiversity mapping, with a range of related societal and economic benefits.

However, textual relations can be complex with one-to-many relationships between entities and relationships occurring sometimes across sentence boundaries, and even some distance from each other. To illustrate, Example 1 is a snippet from a journal article [29] containing four relationships between four different species mentions (one just as a broader genus term and the other two in abbreviated G. species form) to the same toponym one, two or three sentences away.

▶ **Example 1.** "There are also examples of growth forms that make use of the increasing precipitation at <LOCATION> Livingston Island. Most obvious is the very extensive growth of the <SPECIES> *Usnea* lichens whose fruticose form allows them to benefit from both rain, snow and fog (Rundel 1978). Within the mosses, <SPECIES> *A. gainii* and <SPECIES> *H. crispulum* are able to achieve high activities by growing as clumps on the rock surface and storing water from precipitation, and <SPECIES> *A. gainii* is much warmer than other species because of its darker colour."

A significant body of work has addressed the task of extracting toponym references from documents using named-entity recognition [2, 22], and then disambiguating the mentioned place names to identify the coordinates of the named place, requiring resolution of duplicate place names, place name abbreviations or complex place name sequences [5, 21]. Similarly, efforts have been made to develop automated methods to identify biological species in text [3, 17]. However, little attention has been given to the task of matching specific species mentions to their toponym references, and thus georeferencing (identifying the coordinate location) for specific species mentioned in text documents. Without this step, while we can map toponyms mentioned in a document (as for example in [25] and [1]), we cannot map specific items mentioned in text. In the case of biological text documents, this means that we cannot map species distribution or conduct spatial analysis using data "locked away" in text documents.

The task of georeferencing specific phenomena in text has been addressed in some other domains, including disaster management for event georeferencing [13], and significant attention has been given to extracting spatial relations and the reference and located objects to which they refer [18]. However, the former relies on multiple mentions of the same event, using clustering to converge on its probable toponym, and the latter relies on spatial prepositions to connect located and reference objects. A range of work in the relation extraction field of NLP is also relevant, but until recently the focus has been on relationships that occur within the same sentence [7, 33]. Recent studies that also look for relations that may cross sentence boundaries [28, 8] do not include toponym or species and may have limited transferability, or identify relations that express location in a non-geographic context [19].

There remains a gap in solving the difficult problem of associating species mentions with locations for georeferencing. We address this gap by proposing a method that has three steps:– a) extracting species mentions in text; b) extracting and disambiguating toponym mentions in text and c) predicting which, if any, pairs of those species and toponyms represent an actual <species> 'present in' or 'found at' <location> relationship. To perform the latter step, we use a machine learning classifier, in which we classify all species-toponym pairs in the document to identify those that are correct matches. We test and evaluate a range of features for our classifier on a corpus of seven documents (44037 tokens) from the Antarctic domain, including a mixture of journal papers, theses and supplementary material, written by authors from around the world.

Preliminary results using tree-based classifiers achieved precision of up to 0.88. This compares favourably to studies within the biomededical field where precision scores have varied from 0.39 to 0.65 [27, 8, 32, 19]. Furthermore, our bespoke rules- and dictionary-based approach for extraction of species achieved an F1 score of 0.87 including partial matches.

The contributions of our work are two-fold. Firstly, we apply machine learning to solve the difficult and unaddressed problem of associating species mentions with locations for georeferencing. Relation extraction in natural language is regarded as a hard problem generally, and little progress has been made on solving it for the fundamental GIScience task of georeferencing the detailed content of documents, with [1, p.1] pointing out that "Research on this problem is still in its infancy". Generic methods for relation extraction have not been applied to the particular problem of matching species and locations, and cannot handle text where the subject and object of the relation are widely separated, and that involve complex co-references. Furthermore, text that describes location differs from generic language in a number of ways, making the challenge more difficult. Specifically:

1. The ultimate goal of this work is to georeference species mentions, and thus location mentions identified by NER must be tied to specific locations using gazetteers. However, many locations identified by NER are not found in gazetteers, and thus the texts contain many items that look like location references but are actually 'red herrings' (not georeferencable), imposing additional demands on standard relation classifiers.
2. The role of toponyms as universal labels for locations presents different language uses, including common but unusual abbreviations (e.g. MDV for McMurdo Dry Valleys), as well as more standard forms (e.g. Mt for Mount).
3. The role of toponyms as a reference frame results in common repetition of place names (or even only once at the beginning of the paper), interspersed with other place names, and the need to disentangle the correct place name references for multiple mentions of different species is especially challenging.

Our second contribution is the method of application of machine-learning to the problem. Our contribution is in the binary classification model using pairwise matches (e.g. in contrast to dependency driven approaches), and the selection of features.

This paper is structured as follows: Section 2 reviews relevant literature, Section 3 outlines our methodology, Section 4 presents results, which are then discussed in Section 5, before conclusions are summarised in section 6.

2 Related Work

2.1 Recognising Species and Toponyms as Named Entities

The task of detecting location mentions in text is addressed by standard named entity recognition (NER) methods of popular modern natural language processing (NLP) Python packages such as Natural Language Toolkit (NLTK)[1] and spaCy[2]. Recent evaluation of six NER tools on Twitter content indicates precision in the low 90s, with F1 values in the 70s (as recall figures are typically not as high as precision) [16], with slightly lower values in a review on text documents [10]. Following NER, place names must also be resolved to identify the coordinates of the places to which they refer (known as toponym resolution or disambiguation), a task that is challenging due to duplicate place names, unusual abbreviations and spelling variations. Methods for resolving toponyms have included selecting the place with the highest population on the basis that it is more likely to be the one referred to due to size; associated place names mentioned in surrounding text; feature types and language models [4, 15].

[1] http://www.nltk.org
[2] https://spacy.io

In contrast to location, the task of *species* detection is not specifically addressed in standard NER tools. The linguistic structure of taxonomic names does however, allow for the development of automated methods to find these names within natural language text [17]. Linnaean rules for taxonomic names of organisms dictate genus names precede species (and any sub species) name with the former capitalised and all names in either Latin or Greek [17, 23]. Where abbreviated in text, the species name is preceded by the capitalised first letter of the genus and then a period.

Others in the biodiversity domain have tackled the species detection challenge with a mixture of approaches based on matching terms in a dictionary of known entities, [9, 20, 24], isolating taxonomic-looking strings not otherwise in a English lexicon [17, 24], or machine learning approaches like Naïve Bayes [3] or ensemble classifiers that include neural networks [24].

All those methods mentioned above focus on finding scientific names apart from [9] which also finds some common names with their method.

Rule and pattern-based methods can recognise new scientific names in text [17] but can generate false positives for non-scientific text that appears to be in a Linnaean format [24]. Words found in both scientific names and vernacular text cause problems for dictionary and machine learning approaches alike, while machine learning is sensitive to text encoding algorithms [24].

Some dictionary approaches [20] can find new name combinations from existing terms but all are limited by how comprehensive and up-to-date the dictionaries are at time of use, something especially noteworthy in the biodiversity field where thousands of species are discovered or reclassified each year [3, 17, 24].

Recall and precision figures higher than 0.8 and up to 0.97 have been shown to be possible when testing these methods on biodiversity datasets with Quaesitor[24], TaxonFinder[20], and NetiNeti[3] generally performing best in tests [24] conducted.

Despite the collective and respective informativeness of their approaches, none of the options reviewed were suitable for our pipeline for one or more reasons, including the following:

- They were not available as an easy-to-access Python package.
- They returned a single instance of a species mention instead of every mention along with position in the document.
- They separately annotated species and genus for combined terms.
- While available via an online interface or API, they were not appropriate to process entire documents of text, let alone a large collection, due to limits on the amount of text that could be parsed in each use (and while references to DOIs would remove the need to upload entire documents in some cases, a focus of our wider project is processing large batches of legacy documents, some of which do not have DOIs).

2.2 Georeferencing Phenomena in Text

The problem of georeferencing of text has been addressed in other domains. Methods for georeferencing documents by identifying their geographic location have been reviewed by [25], but these methods identify a toponym for the entire document, rather than toponyms for specific items mentioned in the document text. Methods have been developed for extracting place mentions in a document addressing a particular topic, assuming that they describe locations relating to the topic. For example, [1] identifies locations of orchards and cancer cases by extracting place names from specific portions of documents (e.g. the methods section), [6] similarly map locations connected to specific historical events across a document,

and [2] apply a similar approach for biological specimens. However, while this approach identifies locations of phenomena, these works do not distinguish between different types of phenomena mentioned in the text, as is our goal.

In the disaster management field, methods to georeference specific events have involved the collection of multiple documents (e.g. news reports) describing a single disaster event, and have clustered locations mentioned in connection with the event to georeference it [13]. However, this approach relies on multiple documents addressing a specific event, rather than individual mentions of a phenomena in a single document, as we address here.

Another significant body of work has focused on extracting place names and more complex place references of biological specimen collections, but this work addresses place references within databases, in which the species that is being located is already known and stored in another database attribute [11, 12, 14], unlike our challenge in which the link between species and location must be identified.

2.3 Identifying Relationships

In the Natural Language Processing (NLP) literature, the task of relation extraction is aimed at detecting semantic dependencies between items mentioned in text [26]. A number of relation extraction methods have been developed and increasingly methods have been designed to look for relationships that may extend beyond the sentence boundary [8, 28, 31, 32, 19], since as much as 28-30 per cent of relations in certain corpora are inter-sentential [30, 32], and our corpora contains many examples of complex inter-sentence relations between species and locations, such as that shown in Example 1.

While these methods have been applied before on genetics pathways text using graph Long Short Term Memory (LSTM) [27] and on biochemical text using Bi-affine Relation Attention Networks [32], they have had low success rates, or performed erratically depending on type of entity pair [31]. Other work has addressed only a limited the range of text, concentrating on specific biomedical entity types [27, 28, 32] or using a limited number of adjacent sentences [28]. Recent promising work from [8] looks at a large and varied dataset but one that does not identify species among its named entities. The most comparable work looks at linking bacterial species to habitats (e.g. locations within the human body) within biomedical text [19]. Their method has some commonalities with ours in that they consider syntactic features like the presence of verbs and prepositions, but it does not focus on specific toponyms that can be georeferenced, which is the ultimate goal of this study. Furthermore, while they show significant improvement over baselines, the success metrics achieved are relatively low, demonstrating the challenging nature of the problem.

3 Method

3.1 Consolidated and Compartmentalised Approaches

The ultimate aim of this project is to lay some of the groundwork for a complete and automated pipeline to extract and process text from legacy documents, identify mentions of species and toponyms within those documents and distinguish which pairs of species-toponym describe an actual geospatial relationship in which a species was found at a particular location. Figure 1 shows the consolidated end-to-end process, but in this paper, we focus on entity recognition and relationship modelling.

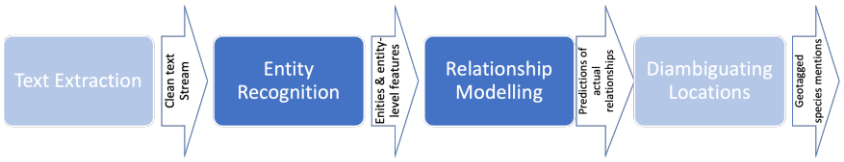


Figure 1 Consolidated pipeline process.

3.2 Entity Recognition

Whilst different natural language processing packages were considered, SpaCy was chosen over Python's Natural Language Toolkit (NLTK) for this work, due to its adaptability, especially the option to train new classes for entity recognition via transfer learning which may be useful in future development of this pipeline. To ensure the entity recognition processes were working on clean streams of text, manually pre-cleaned versions of text were copied from PDF formats avoiding images, tables and reference lists to leave only body text, titles and section headings.

3.2.1 Species Extraction

Our corpus was entirely within the Antarctic geographic area, and we employed Manaaki Whenua's[3] list of Antarctic species to assist with species extraction. The list contains over 2100 known Antarctic animals and organisms with full taxonomic names (including kingdom, phylum, class, order, family, genus and species components). For this initial study, we restricted our focus to scientific rather than common names. This restriction, and particularly the Linnean structure of those taxonomic names led us to adopt a rules-based approach to extraction of species mentions.

Working on the basis that a name would appear in text in the form of *Genus species*, *G. species* or possibly just *Genus*, the full list of taxonomic names was reduced and separated into two Python lists of unique genus and species terms of 744 and 1489 terms respectively which were then used as separate look-up lists to match tokens, bigrams or trigrams to full or abbreviated forms of species names in the list, even when the n-gram is in a combination of genus and species terms not previously found together in the original list. The algorithm also identifies mentions of genus only as well as instances where a genus is followed by a term not originally found in the species list but is found elsewhere in the document in the abbreviated G. species form.

3.2.2 Toponym Recognition and Disambiguation

In contrast to species, as location is a standard entity in the NER tools of NLP packages including SpaCy's, our approach for these entities built upon these available tools.

Using SpaCy's largest English model we applied the NER tool to tokenised documents to identify potential toponyms. These were then checked in a range of gazetteers, with particular emphasis on New Zealand and Antarctic gazetteers (due to the area of interest origin of research) to filter out misidentified toponyms and to pave the way for a genuine toponym resolution process in the future which would link confirmed toponyms to coordinates and thus enable species to be georeferenced.

[3] `https://www.landcareresearch.co.nz`

The gazetteers used were the New Zealand Gazetteer maintained by Land Information New Zealand[4], the Scientific Committee on Antarctic Research (SCAR) gazetteer[5] and the GeoNames gazetteer[6]. These contained approximately 45,000, 26,000 and 1,500,000 unique place names respectively.

A subset of the New Zealand Gazetteer relating only to toponyms in the New Zealand administered area of Antarctica (approx 5100 toponyms) was created as was a New Zealand-specific subset of the SCAR gazetteer (approx 3,500 toponyms). Similarly, subsets of the GeoNames collection relating to Antarctica (approx 18,000) or New Zealand (approx 45,000) respectively were extracted to create a total of seven gazetteers.

Toponym candidates beginning with "The" had this article removed before searching and toponyms beginning with "Mt.' or "Mt' were standardised as "Mount" to aid with matching. Some one-word place names in the broader GeoNames gazetteer that are also common English words or (e.g. Inner, Upper, Fig) or problematic for this process as a single word (South, North, Mount) were filtered from that gazetteer.

Each toponym was checked against each of the seven gazetteers for exact matches taking special note of toponyms specifically matched in an Antarctic and/or New Zealand gazetteer. All exact matches are put forward to the next stage. Additionally, toponyms not found in either Antarctica or New Zealand gazetteers (including some that were exact matches but only in the vast GeoNames gazetteer, e.g. "Portugal") are checked for close or partial matches in the more focused New Zealand and Antarctic gazetteers. This is to ensure misspellings, OCR errors, mis-tokenisation or natural variations do not lead to an Antarctic and New Zealand toponym being a) missed, or b) incorrectly linked to a different part of the world (e.g. "Victoria Land" v "Victoria") as this makes the future step of disambiguation more difficult. We excluded the non-filtered GeoNames list (1.5 million place names) for close matches to focus on matching names to the geographic areas of interest (namely Antarctica and New Zealand) and limit false positives generated during the close and partial matching process.

The close-matching process uses Python's native difflib module[7] to find the best matching entry in each of the six smaller gazetteers if the best match scores over an arbitrarily-set threshold 0.9 using difflib's get_close_matches function.

The partial-matching process looks for the biggest sub string within a potential toponym that can be exactly matched in a gazetteer. For example, the five-token candidate "Ross Sea Region of Antarctica" will be broken into two four-token strings ("Ross Sea Region of" and "Sea Region of Antarctica"), three three-token strings (e.g. "Ross Sea Region" and two others) before the algorithm matches "Ross Sea" in a gazetteer once the candidate is broken into two-token sub-strings.

Toponyms confirmed through an exact, close or partial match would then be passed through to the relationship-prediction stage of the pipeline.

3.3 Modelling

3.3.1 Conceptualisation of the classification task

The main contribution of the paper is in the method for linking specific species and toponym mentions to each other, given that documents can contain many of each, and related species and toponyms may be spread some distance from each other in the document (see Example 1).

[4] `https://gazetteer.linz.govt.nz`
[5] `https://data.aad.gov.au/aadc/gaz/scar/`
[6] www.geonames.org
[7] `https://www.landcareresearch.co.nz`

Furthermore, while every relationship is binary – linking exactly one species instance to one toponym instance in the text – any species or toponyms instance could belong to many relationships. In Example 2: the species instance Umbilicaria aprina Nyl. is in two relationships. One with the toponym instance Dry Valleys and another with Botany Bay.

▶ **Example 2.** Annual activity (% of total time) ranged from 0.2% (Umbilicaria aprina Nyl.) in the Dry Valleys (Raggio et al. 2016) through 4.6% for the same species at Botany Bay.

We thus formulate the problem on the basis that each toponym mention in the document is potentially related to every species mention in the same document and vice versa. For example, a document with 100 species mentions and 100 toponym mentions has 10,000 possible actual relationships. We implement this by creating a matrix of all possible species-toponym pairs, and the task is then to identify the actual relationships among the sea of candidates, for which we use a binary classifier.

3.3.2 Feature engineering

With the goal of predicting which relationships are genuine among the large number of possibilities, we engineered a range of features (see Table 1) that may indicate whether a species-toponym pair match (i.e. toponym x describes the toponym of species y). Some of these were entity-centric (relating to either the toponym or species) while others helped describe the connection or distance between entities in a potential relationship.

Table 1 Engineered Features.

Feature	Level	Type	Notes
Dependency_Steps	Pairwise	Integer	Length of shortest dependency path
inAbstract300_Toponym	Entity	Boolean	Explanation Below
inAbstract300_Species	Entity	Boolean	Explanation Below
inAbstract500_Toponym	Entity	Boolean	Explanation Below
inAbstract500_Species	Entity	Boolean	Explanation Below
max_TFISF_Toponym	Entity	Float	Explanation Below
max_TFISF_Species	Entity	Float	Explanation Below
Num_Nouns_Between	Pairwise	Integer	Count of nouns between entities
Num_Preps_Between	Pairwise	Integer	Count of prepositions between entities
Num_Tokens_Between	Pairwise	Integer	Count of tokens between entities
Num_Verbs_Between	Pairwise	Integer	Count of verbs between entities
Num_Words_Between	Pairwise	Integer	Count of words between entities
Preposition_Between	Pairwise	Boolean	True if preposition between entities
Same_Sentence	Pairwise	Boolean	True if entities in same sentence
Sent_Start_Toponym	Entity	Boolean	True if entity begins a sentence
Sent_Start_Species	Entity	Boolean	True if entity begins a sentence

In Abstract

Four features measured whether or not that entity, regardless of that specific instance's position in the document, was also mentioned in the abstract of a document as defined by being in either the first 300 or 500 tokens after the first mention of the word "Abstract" in a document. This is an attempt to capture some of the entity's document-level characteristics with the a priori assumption that if an entity, be it a species or toponym, is mentioned in

the abstract it is a key theme in the document and takes on increased likelihood of being in an actual relationship if mentioned subsequently. This is opposed to for example, that entity mention simply acting as a comparison to a similar species or place for reference, or representing a toponym where downstream aspects of the research, e.g. processing of samples, occurred.

Term Frequency – Inverse Sentence Frequency (TFISF)

The common measure of term frequency-inverse document frequency (TFIDF) which reflects how important a term is to a document by calculating a term's relative frequency in its own document multiplied by the logged inverse of the proportion of documents it appears in within a corpus. Instead, for our calculations, term frequency – inverse sentence frequency (TFISF) is intended to reflect how important a particular entity is to the sentence containing it.

$$tf(t,s) = f(_{t,s}) \tag{1}$$

$$isf(t,d) = log(\frac{N}{s\varepsilon d : t\varepsilon s}) \tag{2}$$

$$tfisf(t,s,d) = tf(t,d) * isf(t,d) \tag{3}$$

As shown in the equations 1 through 3, term frequency (tf) for a given term (t) is a raw count of how many times that term appears in its sentence (s) (often 1) . Inverse sentence frequency (isf) is the log of the result from dividing the number of sentences in a document (N) by the number of sentences in the document in which the term appears. The two are multiplied together to get TFISF. If an entity contains more than one word, TSISF is calculated for each word in the entity and the highest result selected.

3.3.3 Modelling Approaches

Given the modelling task is a binary classification problem with a known ground truth, a range of supervised machine learning classifiers were applied including logistic regression, AdaBoost, a neural network and three tree-based models – random forest, light gradient boosting machine and extra randomised trees. All were imported as python packages from Scikit-Learn or, in the case of the neural network, Keras.

As this was an exploratory assessment of each model's suitability for further testing and since the size of the training data is too small, at least in terms of members of the target class (i.e., actual relationships) to allow for the creation of an adequately-sized validation set in addition to the training and test splits, no parameters were fined-tuned for any models due to the risk of over-fitting to the training data. Only default parameters were used and a broad-brush summary of some of the key parameters follows.

AdaBoost used 50 estimators and a learning rate of 1, while Light GBM's learning rate was set at 0.1. Light GBM, random forest and extra randomised trees all used 100 estimators with no max depth. Gini was used as the criterion for splitting in both random forest and extra randomised trees while light GBM boosting type was set to a traditional gradient boosting tree.

Logistic regression's penalty function was set to L2 and the solver used was limited memory BFGS while the neural network, which could not be run entirely on default settings was constructed from input and one hidden layers of 64 nodes each, both with rectified linear units (ReLU) as the activation functions and an output layer using sigmoidal activation function. The loss function was binary cross entropy.

Some features were transformed for normality for use in logistic regression and the neural network but only untransformed features were used for the tree-based or boosting models as they are not affected by monotonic transformations to data.

4 Results

4.1 Data Overview

The data consisted of two tranches of documents with one used to develop the entity recognition algorithms and an annotated second set used to act as the ground truth for measuring the performance of the entity recognition and relationship prediction stages.

The first set contained nine PDF documents ranging from PhD theses to academic journal articles and supplementary material with some documents electronically borne and others scanned from printed versions. As these were not annotated, they remained unsuitable for testing and model building.

The second tranche of seven documents, all electronically borne, was annotated by one of the co-authors, a domain expert, using Tagtog, an online text annotation tool. Firstly, any combination of genus and species (or a genus and species term in isolation) were tagged as species and all toponyms were tagged as such. Next, any relationships between one species and one toponym were tagged if the text indicated the species was present or found in that toponym even if that relationship crossed the sentence boundary. Each tagged relationship was binary, linking exactly one species instance to one toponym instance in the text, but any species or toponyms instance could belong to multiple one-to-one relationships.

Table 2 Training Data Overview.

Doc.	Sentences	Tokens	Species Mentions	Location Mentions	Potential Pairs	True Pairs	True Pair %
1	214	5120	12	96	1152	11	0.955
2	252	7192	13	87	1131	12	1.061
3	335	8779	55	138	7590	48	0.632
4	194	5524	92	35	3220	11	0.342
5	82	2387	3	64	192	4	2.083
6	143	4212	8	31	248	3	1.210
7	323	10823	83	95	7885	40	0.507
Total	1543	44037	266	546	21418	129	0.602

Within this collection (as seen in Table 2), the tokenised documents ranged in length from 2300 tokens to over 10,000 tokens and from just 82 sentences to 335 sentences. Overall, the training data contained 44,037 tokens and 21,418 possible relationships between the 266 species mentions and 546 toponym mentions. This is a smaller than ideal set but indicates potential in the presented methods applied to texts describing object locations, with multiple mentions and relations that may be greatly separated.

4.2 Performance of Entity Recognition

Tables 3 and 4 show the performance of the approaches to recognising species and toponyms within the documents when matched to annotated entities, using five-fold cross-validation. We measured precision, recall and F1 score on two levels. Firstly, if there is an exact match between the entity as it was annotated and as it was extracted by the entity recognition

process (see Table 3); and secondly including partial matches where an extracted or annotated entity is contained entirely within the other (see Table 4). A simple, and easily correctable, example of a partial match for species is the algorithm failing to match the *spp.* token of *Diplosphaera spp.* because SpaCy incorrectly parses the period as a separate token but the process nonetheless recognises and extracts the genus on this example and this may still hold some value for the end user.

Table 3 Species and Toponym Recognition – Exact Matches.

Entity	Precision	Recall	F1
Species	0.9279	0.7256	0.8144
Toponym	0.9760	0.4469	0.6131

Table 4 Species and Toponym Recognition – Exact & Partial Matches.

Entity	Adjusted Precision	Adjusted Recall	Adjusted F1
Species	0.9904	0.7744	0.8692
Toponym	0.9840	0.4505	0.6180

The rules- and dictionary-based approach to species recognition correctly identifies 72.6 per cent of species exactly as they are tagged by the annotator and this recall figure rises to 77.4 per cent when instances of partial matches are included. Precision is higher at 0.9279 (0.9904 including partial matches) with an overall F1 score for the species recognition process of 0.8144 (0.8692).

The corresponding results for the toponym recognition process using SpaCy's inbuilt NER tool and a multi-level gazetteer matching process are generally lower than that for species with the exception of precision for exact matches which is higher at 0.9760. This reflects a high degree of confidence that a toponym that passes through the gazetteer-matching process as either an exact, close or partial match in a gazetteer will also have been tagged as a toponym by the annotator. This figure rises to 0.9840 when partial matches with what has been annotated are included.

However, recall for toponym is 0.4469 (0.4505 including partial matches) and for some documents, this is as low as 0.2286 for exact matches. Some ultra-specific toponyms e.g. subglacial caves like *Harry's Dream* and *22 Blue* are not recognised in gazetteers and abbreviations like *MDVs* for all subsequent mentions of *McMurdo Dry Valleys* in one particular document are filtered out by the gazetteer-matching process as false negatives.

4.3 Performance of Species-Toponym Matching Model

We performed classification with three tree-based classifiers, after tests with neural networks, logistic regression and AdaBoost proved less successful.

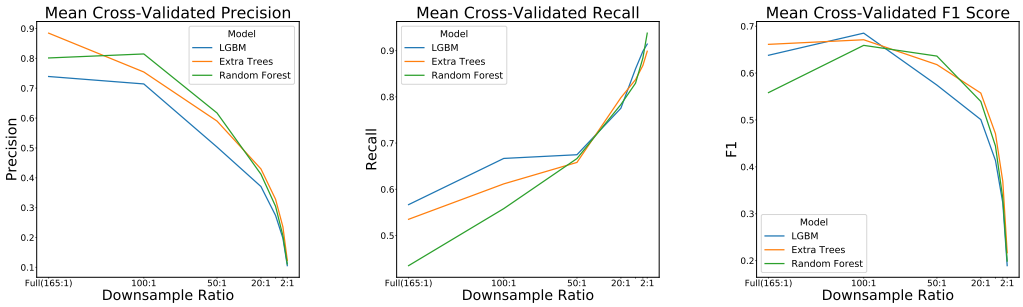
Because of the large ratio of false relationships to actual relationships, we attempt to mitigate the impact by down-sampling the data at different rates to reduce the imbalance in the data. The original ratio of 160 false relationships for every one actual relationship was reduced to 100:1, 50:1, 10:1, 5:1 and 2:1 in different iterations of testing to gauge if this approach could help a machine learning tool make better predictions.

Table 5 shows the average performances of the model and down-sample combinations with three classifiers with five-fold cross validation. Light GBM and extra randomised trees combined with a slight down-sample of 100 to 1 achieved the highest F1 scores, while the extra trees model also did comparatively well on the 'full' training sample within each cross-validation split.

Table 5 Five-fold Mean Cross Validation Scores.

Model	Down-sample Ratio	Precision	Recall	F1
Extra Trees	None (full sample)	**0.884561**	0.535077	0.661317
Extra Trees	100:1	0.754360	0.611692	0.671119
Extra Trees	50:1	0.590000	0.658154	0.618070
Extra Trees	10:1	0.327795	0.837231	0.470736
LGBM	None (full sample)	0.739273	0.566769	0.638022
LGBM	100:1	0.714369	0.666769	**0.685243**
LGBM	50:1	0.502949	0.674769	0.574422
LGBM	10:1	0.273593	**0.860615**	0.413559
Random Forest	None (full sample)	0.801411	0.434769	0.558376
Random Forest	100:1	0.814839	0.558154	0.659163
Random Forest	50:1	0.617086	0.666462	0.636299
Random Forest	10:1	0.304225	0.829538	0.444405

The trends in the effect of down-sampling can further be seen in the graphs in Figure 2 which respectively plot precision, recall and F1 across models and down sampling rate. The more aggressively the majority 'no relationship' class was down-sampled to match the minority 'actual relationship' class the higher the rate of recall as the models got better at finding all actual relationships. However, this occurred at the expense of precision and generated more false positives. The F1 score tended to increase for a minor down-sampling effort of 100:1 for all three models and fall away after that.



(a) Precision. **(b)** Recall. **(c)** F1 Score.

Figure 2 Mean Cross-Validated Metrics by Downsample Rate and Model.

4.4 Feature Importance

For an indication of feature importance in the three types of tree-based models, we extracted the feature importance rankings for the respective models when used on full samples. Importance is determined by the the number of splits using each feature as a percentage of total splits in the the respective collections of decision trees that comprised each model. Table 6 shows some commonalities among which features were used most by the various models.

The variables created to capture whether or not a species was also mentioned in the abstract (at either 300 or 500 tokens from the first mention of 'Abstract' in the document) and whether or not a preposition is between the two entities provided little to no value in the

models, but the TSISF variables for both species and toponym were in the top seven features for all three models and were the most important for the LGBM model, while TSISF for toponym was also the most important for the extra randomised trees model. All of the five *number of <type> between* distance variables rounded out the top seven in each model, and the *number of dependency steps* also appeared in the top nine in each of the models.

Table 6 Most Important Features by Model Type.

	Extra Random Trees		Light GBM		Random Forest	
Feature	**Rank**	**Imp.**	**Rank**	**Imp.**	**Rank**	**Imp.**
max_TFISF_Location	1	0.1461	1	0.2687	5	0.1222
Num_Words_Between	2	0.1317	7	0.0553	1	0.1700
Num_Tokens_Between	3	0.1316	4	0.1203	2	0.1530
Num_Preps_Between	4	0.1217	6	0.0787	4	0.1295
Num_Nouns_Between	5	0.1105	5	0.1027	3	0.1401
Num_Verbs_Between	6	0.1095	3	0.1477	6	0.0901
max_TFISF_Species	7	0.0657	2	0.1670	7	0.0721
Dependency_Steps	8	0.0589	9	0.0113	9	0.0266
Sent_Start_Location	9	0.0491	8	0.0133	8	0.05055

5 Discussion

5.1 Entity Recognition

The precision for both species and toponym recognition is high (>0.9 for both) and minor adjustments to the species method utilising existing SpaCy functionality such as out-of-vocab tags could help it learn to recognise species and genus names that aren't listed in a given dictionary. This would lift the species recall figure which sits at 0.77 including partial matches. Training a NLP engine to identify species entities is another option but not tested so far with this work due to a lack of training data.

With a recall of just 0.45, larger adjustments are required to the toponym recognition process. The task of toponym extraction in these types of documents is challenging because of issues such as "second mentions" of place names (e.g. "McMurdo Dry Valleys"), including pronouns ("it"), shortened forms ("the Valleys") and acronyms ("MDVs"), and methods to address these through coreference resolution would improve results. Loosening the gazetteer filtering process to allow more toponym candidates to pass through would improve recall, but at the cost of precision.

5.2 Relationship Extraction

The performance of the relationship modelling process with limited fine-tuning of tree-based models is promising. The highest precision for predicting actual relationship was 0.88 (for the Extra Trees model with no down-sampling), and the highest recall was 0.86 (for the LGBM model with 10:1 down-sampling). The highest F1 score of 0.68 was achieved for LGBM with 100:1 down-sampling (see Table 5). Further work on larger annotated data sets would allow for validation sets to be created in addition to train/test splits and facilitate tuning of models.

Ours is an approach that has not been applied to species-toponym relationships before and it shows promising results when compared with other work addressing related, but different problems in the biomedical fields, which use graph LTSMs [27], Bi-affine Relation

Attention Networks [32], and transformer-type networks [8]. Furthermore, out work is not limited to finding cross-sentence relations only in adjacent or near-adjacent sentences like [28]. The process has highlighted however, the potential of exploring coreference resolution as recent studies [8, 28] have done.

While some of the engineered features, namely the two TSISF measures showed promise in the modelling process, others seemingly offered little utility. The process of engineering and exploring new features (e.g. word vectors, sentence polarity etc) that help map the relationship of species and toponym should be explored further and may yield improved results as part of future work.

6 Conclusion

In this paper, we have described a method to extract mentions of species and place names from text documents, and then to determine which place names describe the toponyms of which species.

We have demonstrated a rules- and dictionary-based approach for the extraction of species in the Antarctic context, and applied existing place name extraction methods, with a set of gazetteers to identify toponyms. Our main contribution is the development of a method that uses tree-based classifiers to match toponyms and species mentions in order to identify the toponym of specific species, for georeferencing and mapping, with a precision of 0.88 (highest F1 of 0.68). In future work, we plan to include additional features in the model, employ a larger corpus for training and tuning; and improve efficiency through filtering some species-toponym combinations (those that are unlikely) before applying the classifier.

This research contributes to the goal of georeferencing text mentions of specific species on two specific fronts. Firstly, little to no research is available on automatically extracting species toponyms from text documents, and the work described in this paper is among the first to provide a method for extracting specific mention toponyms in the biological domain. Secondly, cross-domain methods for georeferencing mentions of different kinds of phenomena (and being able to identify which kinds of phenomena are where) in a document have been limited thus far. While this is early work, the method shows promise, particularly for dealing with relations that may cross sentence boundaries and contain other kinds of complexities such as abbreviations and a many to many relationship between species and toponyms.

References

1 Elise Acheson and Ross S. Purves. Extracting and modeling geographic information from scientific articles. *PLOS ONE*, 16(1):e0244918, January 2021. doi:10.1371/journal.pone.0244918.

2 Moises Acuna-Chaves and José Araya. Extraction of geographic entities from biological textual sources. In *2017 XLIII Latin American Computer Conference (CLEI)*, pages 1–8, 2017. doi:10.1109/CLEI.2017.8226422.

3 Lakshmi Manohar Akella, Catherine N. Norton, and Holly Miller. NetiNeti: discovery of scientific names from text using machine learning methods. *BMC Bioinformatics*, 13(1):211, 2012. doi:10.1186/1471-2105-13-211.

4 Ana Bárbara Cardoso, Bruno Martins, and Jacinto Estima. Using recurrent neural networks for toponym resolution in text. In *EPIA Conference on Artificial Intelligence*, pages 769–780. Springer, 2019.

5 Arthur D Chapman and John R Wieczorek. *Georeferencing Best Practices*. GBIF Secretariat, Copenhagen, 2020. doi:10.15468/doc-gg7h-s853.

6 Rachel Chasin, Daryl Woodward, Jeremy Witmer, and Jugal Kalita. Extracting and displaying temporal and geospatial entities from articles on historical events. *The Computer Journal*, 57(3):403–426, 2014.

7 Hong-Woo Chun, Yoshimasa Tsuruoka, Jin-Dong Kim, Rie Shiba, Naoki Nagata, Teruyoshi Hishiki, and Jun'Ichi Tsujii. Extraction of Gene-Disease Relations from Medline using Domain Dictionaries and Machine Learning. In *Biocomputing 2006*, pages 4–15, Maui, Hawaii, December 2005. World Scientific. doi:10.1142/9789812701626_0002.

8 Markus Eberts and Adrian Ulges. An End-to-end Model for Entity-level Relation Extraction using Multi-instance Learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3650–3660, 2021. URL: https://www.aclweb.org/anthology/2021.eacl-main.319.

9 Martin Gerner, Goran Nenadic, and Casey M. Bergman. LINNAEUS: A species name identification system for biomedical literature. *BMC Bioinformatics*, 11(1):85, 2010. doi:10.1186/1471-2105-11-85.

10 Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. What's missing in geographical parsing? *Language Resources and Evaluation*, 52(2):603–623, 2018.

11 Qinghua Guo, Yu Liu, and John Wieczorek. Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach. *International Journal of Geographical Information Science*, 22(10):1067–1090, 2008.

12 Robert P Guralnick, John Wieczorek, Reed Beaman, Robert J Hijmans, the BioGeomancer Working Group, et al. Biogeomancer: automated georeferencing to map the world's biodiversity data. *PLoS Biol*, 4(11):e381, 2006.

13 Felix Hamborg, Corinna Breitinger, and Bela Gipp. Giveme5W1H: A Universal System for Extracting Main Events from News Articles. *arXiv:1909.02766 [cs]*, September 2019. URL: http://arxiv.org/abs/1909.02766.

14 Andrew W Hill, Robert Guralnick, Paul Flemons, Reed Beaman, John Wieczorek, Ajay Ranipeta, Vishwas Chavan, and David Remsen. Location, location, location: utilizing pipelines and services to more effectively georeference the world's biodiversity data. *BMC bioinformatics*, 10(14):1–9, 2009.

15 Yiting Ju, Benjamin Adams, Krzysztof Janowicz, Yingjie Hu, Bo Yan, and Grant McKenzie. Things and strings: improving place name disambiguation from short texts by combining entity co-occurrence with topic modeling. In *European Knowledge Acquisition Workshop*, pages 353–367. Springer, 2016.

16 Morteza Karimzadeh, Scott Pezanowski, Alan M. MacEachren, and Jan O. Wallgrün. GeoTxt: A scalable geoparsing system for unstructured text geolocation. *Transactions in GIS*, 23(1):118–136, 2019. doi:10.1111/tgis.12510.

17 Drew Koning, Indra Neil Sarkar, and Thomas Moritz. TaxonGrab: Extracting Taxonomic Names From Text. *Biodiversity Informatics*, 2, 2005. doi:10.17161/bi.v2i0.17.

18 Parisa Kordjamshidi, Martijn Otterlo, and Marie-Francine Moens. Spatial role labeling: Towards extraction of spatial relations from natural language. *TSLP*, 8:4, December 2011. doi:10.1145/2050104.2050105.

19 Parisa Kordjamshidi, Dan Roth, and Marie-Francine Moens. Structured learning for spatial information extraction from biomedical text: bacteria biotopes. *BMC bioinformatics*, 16:129, April 2015. doi:10.1186/s12859-015-0542-z.

20 Patrick R. Leary, David P. Remsen, Catherine N. Norton, David J. Patterson, and Indra Neil Sarkar. uBioRSS: Tracking taxonomic literature using RSS. *Bioinformatics*, 23(11):1434–1436, 2007. doi:10.1093/bioinformatics/btm109.

21 Jochen L. Leidner. Toponym resolution in text: annotation, evaluation and applications of spatial grounding. *ACM SIGIR Forum*, 41(2):124–126, December 2007. doi:10.1145/1328964.1328989.

22 Jochen L. Leidner and Michael D. Lieberman. Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special*, 3(2):5–11, 2011. doi:10.1145/2047296.2047298.

23 Carl Linnaeus. *Species Plantarum*. Laurentius Salvius, Stockholm, Sweden, 1753.

24 Damon P. Little. Recognition of Latin scientific names using artificial neural networks. *Applications in Plant Sciences*, 8(7):e11378, 2020. doi:10.1002/aps3.11378.

25 Fernando Melo and Bruno Martins. Automated Geocoding of Textual Documents: A Survey of Current Approaches. *Transactions in GIS*, 21(1):3–38, 2017. doi:10.1111/tgis.12212.

26 Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. Named entity recognition and relation extraction: State-of-the-art. *ACM Computing Surveys (CSUR)*, 54(1):1–39, 2021.

27 Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. Cross-Sentence N-ary Relation Extraction with Graph LSTMs. *Transactions of the Association for Computational Linguistics*, 5:101–115, 2017. doi:10.1162/tacl_a_00049.

28 Chris Quirk and Hoifung Poon. Distant Supervision for Relation Extraction beyond the Sentence Boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1171–1182, Valencia, Spain, April 2017. Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/E17-1110.

29 Burkhard Schroeter, T. G. Allan Green, Ana Pintado, Roman Türk, and Leopoldo G. Sancho. Summer activity patterns for mosses and lichens in Maritime Antarctica. *Antarctic Science*, 29(6):517–530, December 2017. doi:10.1017/S095410201700027X.

30 Kumutha Swampillai and Mark Stevenson. Inter-sentential Relations in Information Extraction Corpora. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

31 Kumutha Swampillai and Mark Stevenson. Extracting relations within and across sentences. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 25–32, 2011.

32 Patrick Verga, Emma Strubell, and Andrew McCallum. Simultaneously Self-Attending to All Mentions for Full-Abstract Biological Relation Extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 872–884, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi:10.18653/v1/N18-1080.

33 Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. Classifying relations via long short term memory networks along shortest dependency paths. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1785–1794, 2015.