BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 2 2012, pages 254–260
doi:10.1093/bioinformatics/btr640

Data and text mining

Advance Access publication November 30, 2011

Which species is it? Species-driven gene name disambiguation using random walks over a mixture of adjacency matrices

Nathan Harmston[1], Wendy Filsell[2] and Michael P. H. Stumpf[1,*]

[1]Centre for Bioinformatics, Division of Molecular Biosciences, Imperial College London, London SW7 2AZ and [2]Unilever R&D, Colworth Science Park, Sharnbrook, Bedfordshire MK44 1LQ, UK

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: The scientific literature contains a wealth of information about biological systems. Manual curation lacks the scalability to extract this information due to the ever-increasing numbers of papers being published. The development and application of text mining technologies has been proposed as a way of dealing with this problem. However, the inter-species ambiguity of the genomic nomenclature makes mapping of gene mentions identified in text to their corresponding Entrez gene identifiers an extremely difficult task. We propose a novel method, which transforms a MEDLINE record into a mixture of adjacency matrices; by performing a random walkover the resulting graph, we can perform multi-class supervised classification allowing the assignment of taxonomy identifiers to individual gene mentions. The ability to achieve good performance at this task has a direct impact on the performance of normalizing gene mentions to Entrez gene identifiers. Such graph mixtures add flexibility and allow us to generate probabilistic classification schemes that naturally reflect the uncertainties inherent, even in literature-derived data.

Results: Our method performs well in terms of both micro- and macro-averaged performance, achieving micro-F_1 of 0.76 and macro-F_1 of 0.36 on the publicly available DECA corpus. Re-curation of the DECA corpus was performed, with our method achieving 0.88 micro-F_1 and 0.51 macro-F_1. Our method improves over standard classification techniques [such as support vector machines (SVMs)] in a number of ways: flexibility, interpretability and its resistance to the effects of class bias in the training data. Good performance is achieved without the need for computationally expensive parse tree generation or 'bag of words classification'.

Contact: m.stumpf@imperial.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 11, 2011; revised on November 12, 2011; accepted on November 15, 2011

1 INTRODUCTION

The scientific literature contains a wealth of information about biological organisms that is relevant for biomedical scientists: from descriptions of protein–protein interactions (PPIs) (Hoffmann *et al.*, 2005) to parameters for models of biological systems (Hakenberg *et al.*, 2004). The ever-increasing publication rate now means that it has become impossible for manual curators to extract this information comprehensively. This situation has led to a sustained interest in the development and application of text mining (TM) methods to the biological and biomedical domain (Ananiadou *et al.*, 2006; Harmston *et al.*, 2010).

Many organisms are used to investigate biological processes and phenomena (Fields and Johnston, 2005) with the expectation that any understanding obtained can be applied to other organisms (e.g. the zebrafish immune system is used as a model of the human immune system). However, in some cases this mapping of experimental results is not justified due to inter-species variation (i.e. a protein may be subject to neo/subfunctionalization), which could lead to erroneous inferences being drawn. Therefore, it is important to identify the species to which any genes (and other entities of interest) mentioned in a paper belong, and in which species the experiments were performed in order to reduce the chances of this kind of mistake occurring. Here, we provide a new method that is able to achieve this automatically embedded in a TM framework.

A typical TM pipeline works by first recognizing or tagging named entities (NEs) of interest (e.g. gene mentions or species mentions) in a block of text and then normalizing or mapping these mentions to canonical unique identifiers (e.g. Entrez gene identifiers or NCBI Taxonomy identifiers). This normalization process facilitates summarizing and integration of results obtained using TM with other types of biological knowledge. However, the ambiguity and variability of the genomic nomenclature makes this task extremely challenging. A single gene mention can potentially refer to many genes within a species or across species.

The inter-species ambiguity of the genomic nomenclature (Chen *et al.*, 2005; Mons, 2005) means that identifying the correct species of a gene mention is an important subtask of gene normalization (Hakenberg *et al.*, 2008; Wang and Matthews, 2008), but despite of this there has been only limited work in this area. Kappeler *et al.* (2009) investigated methods to identify the focal organisms of an abstract using a custom species tagger and various heuristics. This work showed that identifying the species mentioned in an abstract was beneficial both for gene normalization and PPI extraction. Genes from several species can be mentioned in a single article (e.g. an article on evolutionary analysis of a gene may refer to its homologues in multiple species) so it is important to perform species disambiguation at the level of an individual gene mention. Not performing species-driven gene name disambiguation led to poor performance in the BioCreative (BC)

*To whom correspondence should be addressed.

II PPI extraction task (Krallinger *et al.*, 2008) as gene mentions participating in interactions were required to be normalized to identifiers from various species. The simplest method to perform species disambiguation is to assign taxonomy identifiers based on a majority vote, where the most frequent taxonomy identifier in a document is assigned to all genes in that document. Another simple way to perform species disambiguation at the document level is to use the MeSH terms attached to a MEDLINE record. MeSH is a controlled vocabulary where terms are assigned to new documents by a team of annotators. However, not all records have MeSH identifiers and there is an implicit assumption of both the completeness of MeSH vocabulary (with respect to species names) and the reliability of the manual annotations.

Both the BC II.5 (Leitner *et al.*, 2010) and BC III gene normalization task were concerned with inter-species gene normalization at the document level. Inter-species gene name disambiguation is a difficult task, as is evident from the large difference in performance between the best results reported in the BC II evaluation (which only looked at identifiers from human) and BC II.5. Verspoor *et al.* (2010) investigated different strategies for assigning species corresponding to genes mentions, e.g. assigning gene mentions to a species based on their proximity to species words, or by considering which species was mentioned first in the document. The publicly accessible GNAT (Hakenberg *et al.*, 2011) and GeneTUKit (Huang *et al.*, 2011) systems both perform species-driven gene name disambiguation to achieve high levels of performance in gene name normalization tasks. Methods which only look a global features of a document lead to poor performance at the level of individual mentions. In order to improve performance, the syntactical/contextual relationship (i.e. co-occurrence in a sentence or noun phrase) of words suggestive of a certain species with an individual gene mention needs to be used.

Recently, Wang *et al.* (2010) produced the freely available DECA corpus (available for download from http://www.nactem.ac.uk/deca/) allowing the development and evaluation of methods for species disambiguation at the mention level. The DECA corpus is an annotated corpus of 636 abstracts from the BC I and BC II gene normalization tasks, consisting of pairs of gene mentions and NCBI taxonomy identifiers. Gene mentions were identified using dictionary-based named entity recognition, and taxonomy identifiers were manually assigned to the identified mentions by PhD level biologists. The annotators were only allowed to choose from 10 potential taxonomy identifiers (those identified as the most frequent based on the BC II Protein Interaction Pairs task (Krallinger *et al.*, 2008), if the gene mention was found to belong to another species then it was annotated as 'Other'. Several methods were proposed and evaluated (RELATION, ML and HYBRID methods) on this corpus. A hybrid method combining a syntactic parser with both a relation and supervised classification model was found to have the best overall performance.

We propose a novel method, which transforms a MEDLINE record into a mixture of adjacency matrices and by using a random walkover this mixture model performs multi-class supervised classification (where the potential classes are defined by the contents of a record). Each class corresponds to a single NCBI Taxonomy identifier. This method does not require training data for all potential classes in order to achieve high performance and does not only perform classification but also provides a probability, which serves to quantify the certainty attached to a classification.

Table 1. Corpora statistics: all instances of *ncbitaxon:-1* were excluded from counts of documents, classes and annotations

	Corpus		
	Original	A	B
Documents	636	634	629
Classes	11	11	33
Annotations	6227	6218	5958
Re-annotations	–	9	646

Re-annotations is the number of annotations that have changed compared to the original DECA annotations.

2 METHODS

2.1 Preprocessing of the DECA corpus

Two of the abstracts in the original DECA corpus were found to be duplicates and were removed (corpus A). Error analysis of the results produced by our method led us to identify several annotation mistakes in the corpus. This led to extensive manual re-curation of the corpus. Where possible if the species was proposed in the original corpus as 0 (other species), it was re-annotated with its correct taxonomy identifier (corpus B). A representative selection of these re-annotations were checked by another annotator and found to be correct. This led to a large increase in the number of potential classes to distinguish between. Statistics regarding the resulting corpora are provided in Table 1. We evaluated our method on both the original corpus (with duplicate documents removed) and on the re-annotated version (Supplementary Material I). Where available, MeSH terms for all abstracts were obtained from the MEDLINE database. Titles were also obtained from MEDLINE when they were absent from the original BC corpora.

2.2 Processing pipeline

The resulting text was then split into sentences using the Julie Sentence Boundary detector (Hahn *et al.*, 2008). After this, shallow parsing using the GENIA tagger (Tsuruoka *et al.*, 2005) was then performed on these sentences providing information on the boundaries of noun and verb phrases. We use the gene mentions provided in the DECA corpus.

As in previous work, we collectively refer to words suggestive of the species of a gene entity (such as species mentions, cell line mentions, MeSH terms and prefixes) as *species words*. We evaluate our method using two different methods for identifying these *species words*. In order to provide as fair a comparison as possible to the results in the original DECA paper, we used the U-Compare (Kano *et al.*, 2009) NaCTeM Species Word Detector in order to identify species mentions. This component both identifies species mentions and normalizes them to their corresponding NCBI Taxonomy identifier. We obtained the mapping used by Wang *et al.* (2010) in order to map non-species identifiers to species identifiers (i.e. allowing the mapping of *mammalian* to 9606 and *Drosophila* to 7227). We denote the use of this set of *species words* on corpus A as A_1. We also evaluate how the performance depends on the set of *species words* used. A custom *species word* detection and normalization component was used, which identifies a variety of *species words* (species mentions, cell line mentions and MeSH terms) and normalizes them to NCBI taxonomy identifiers. We denote the use of this set of *species words* on corpus A as A_2.

Species mentions were identified using the LINNAEUS species tagger (Gerner *et al.*, 2010) combined with an implementation of the Schwartz and Hearst algorithm (Schwartz and Hearst, 2003) to identify and disambiguate abbreviations. Species mentions which were part of experimental methods mentions were then filtered out (i.e. yeast from yeast two-hybrid method). Experimental methods were identified using exact text matching against a dictionary of experimental methods extracted from the PSI:MI ontology (Supplementary Material II). Species mentions were then normalized to

NCBI Taxonomy identifiers using dictionary lookup and a set of heuristics to help handle mentions that mapped to multiple identifiers.

One type of species mention that can lead to this problem is the use of a shortened form of the Linnaean binomial species name (*D. miranda*) without the explicit declaration of the long form (*Drosophila melanogaster*). Authors may do this when talking about a number of species from the same genus, the first species mention been written with its full Linnaean binomial name, followed by mentions with an abbreviated genus component of the name. This observation allows the use of a heuristic where an ambiguous mention is normalized by using the taxonomic tree. Given an ambiguous mention, we identify the potential genera it could be associated with and look in the abstract for any unambiguous mentions with these potential genera. If only one exists, we propose this to be the genus of the mention and use this to normalize the species mention to a unique taxonomy identifier e.g. if an abstract contains mentions of *D.melanogaster* and *D.miranda* then *D.miranda* would be mapped as *Drosophila miranda* rather than *Diaspis miranda* as *Drosophila* is the genus of both *Drosophila melanogaster* and *Drosophila miranda*. In the cases where this heuristic could not be applied and the ambiguous abbreviation potentially referred to a major model organism (e.g. *Caenorhabditis elegans* maps to 41 different organisms) and a long form is not present, then the mention is mapped to the most probable model organism [*C.elegans* (6239)]. This module was evaluated on the Linnaeus-100 corpus achieving 0.9683 precision and 0.9606 recall. When no further rules could be applied to ambiguous mentions, a uniform probability was assigned to all remaining potential taxonomy identifiers. While this does not decrease the number of ambiguous species mentions found in the DECA corpus, it does have a large impact on reducing the number of ambiguous mentions found when normalizing species mentions tagged in the entire MEDLINE database.

Cell line mentions were tagged using both the ABNER (Settles, 2005) and the GENIA tagger. These mentions were then mapped to taxonomy identifiers using a manually curated dictionary, which was created by integrating CLKB (Sarntivijai *et al.*, 2008), ATCC, HyperCLDB (Romano *et al.*, 2009), MeSH ontology and the DMSZ catalogues (Supplementary Material III). If a cell line mention could be normalized to multiple taxonomy identifiers, then the same probability was assigned to all the potential identifiers that the mention could refer to. MeSH terms from the B subtree of the MeSH ontology were mapped to taxonomy identifiers using exact dictionary matching against the NCBI taxonomy database (Supplementary Material IV).

Sometimes prefixes (e.g. *At*, *Zm*, *Os*, *h*, *r*, *m*, *d*, *y*) are attached to gene mentions to indicate the species that the gene symbol is referring to. Typically, this is only allowed by a limited number of journals, and only to help avoid confusion during cross-species comparisons. Prefixes can also be used to indicate whether the gene mention refers to a mutant version (e.g. **m**Sos-1) or that it is a ribosomal gene (e.g. **r**Chromatin); these are, of course, ambiguous with the prefix for mouse (m) and rat (r), respectively. Where possible we distinguish the correct sense of the prefix by using the long form (if present) identified by the implementation of the Schwartz and Hearst algorithm.

2.3 Random walks over a mixture of adjacency matrices

We propose a novel method to solve the species-driven gene name disambiguation problem by using a random walkover an additive mixture model of adjacency matrices (Fig. 1). For a specific document, i, we denote the set of gene mentions for that document as M_i, the set of *species words* associated with that document as N_i and the set of taxonomy identifiers to which the *species words* in that document can be normalized to as Oi. We assume that different authors state the species of a gene mention over all documents in a similar way irrespective of the actual species being proposed.

We represent the relationships between these *species words* and gene mentions as a directed bipartite graph, $G^{(i,j)}$, with edges, $e \in E^{(i,j)}$, from nodes in M_i to nodes in N_i.

$$G^{(i,j)} = (M_i, N_i, E^{(i,j)}). \quad (1)$$

Ly9 is a mouse cell membrane antigen found on all lymphocytes and coded for by a gene that maps to chromosome 1. We previously described the isolation and characterization of a full-length cDNA clone for mouse Ly9. Using cross-species hybridization we isolated cDNA clones encoding the human homologue Humly9. PMID:8537117

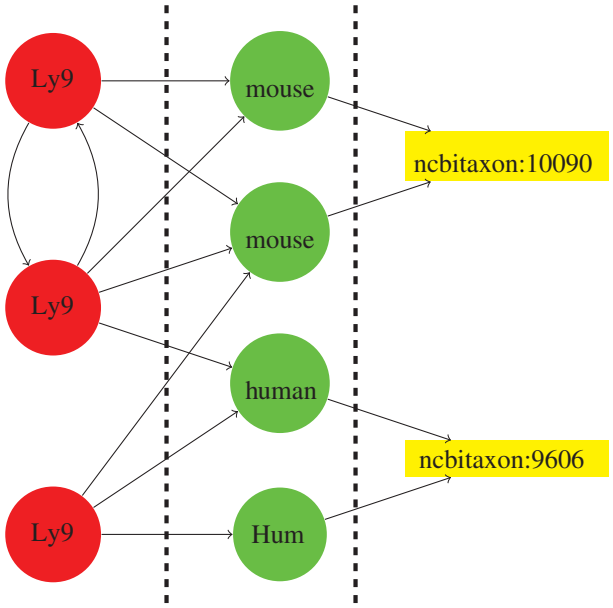

Fig. 1. Example representation of the method. A block of text is processed into a tripartite graph shown above. Gene mentions are shown as red circles. Species mentions as green circles and taxonomy identifiers as yellow rectangles. 10 090 refers to the taxonomy identifier for *Mus musculus* and 9606 for *Homo sapiens*. Not all edges shown.

The adjacency matrix $A^{(i,j)}$ is the corresponding adjacency matrix for graph $G^{(i,j)}$ with entries defined by,

$$A^{(i,j)}_{mn} = \begin{cases} 1, & \text{if } J(m,n) \text{ is true and } m \in M_i \text{ and } n \in N_i \text{ and } mn \in E^{(i,j)} \\ 0, & \text{else.} \end{cases} \quad (2)$$

Each matrix, $A^{(i,j)}$, represents the adjacency matrix for a specific relationship type j in document i (Table 2) and is generated by combining the entities and information generated by the processing pipeline. An edge exists between a gene mention (m), species word (n) pair for a specific feature (j) if that feature holds for that pair [i.e. $J(m,n)$ evaluates to true]. This is equivalent to its corresponding entry in $A^{(i,j)}_{mn}$ being equal to 1. If m and n are in the same sentence, then the corresponding entry for adjacency matrix 1 is 1. These are then combined together using an additive mixture model. The normalization of species mentions to taxonomy identifiers is represented by a different directed bipartite graph, $\tilde{G}^i$,

$$\tilde{G}^{(i)} = (N_i, O_i, \tilde{E}^{(i)}), \quad (3)$$

where there are edges, $e \in \tilde{E}^{(i)}$, from nodes in N_i to nodes in O_i. The adjacency matrix $B^{(i)}$ is the adjacency matrix for graph $\tilde{G}^{(i)}$ with entries defined using Equation (5). The edge weights are the probabilities assigned by the species normalizer that a species mention normalizes to a specific taxonomy identifier,

$$\text{Neigh}(n) = \{o \mid o \in O_i \text{ and } n \in N_i \text{ and } (n,o) \in \tilde{E}^i\}, \quad (4)$$

and

$$B^{(i)}_{no} = \begin{cases} \frac{1}{|\text{Neigh}(n)|} & \text{if } |\text{Neigh}(n)| > 0 \\ 0, & \text{else.} \end{cases} \quad (5)$$

Table 2. Features used in the mixture model

	No.	Feature	CV_{A_2}	CV_B
J	1	Occurs in same sentence	1.2475	0.9615
	2	Occurs in same noun phrase	1.0104	1.2249
	3	Document has MeSH term	1.3507	0.7491
	4	Occurs in the same sentence to the left	0.6068	0.7395
	5	Occurs in the same sentence to the right	1.3007	0.8818
	6	Occurs to the left in the abstract	0.6424	0.0068
	7	Occurs to the right in the abstract	0.6496	0.7109
	8	Has prefix	0.0000	0.0000
	9	Occurs in document	1.0112	0.6858
K	1	Same entity word left	0.4916	0.4784
	2	Same entity word right	0.7236	0.2265

Features are divided into two sets J and K, for use in the two-step and three-step random walk, respectively. Also shown is the coefficient of variation (CV) for the estimated parameters on the both versions of the corpus using our custom *species word* detection and normalization component (A_2 and B).

The union of these two bipartite graphs is then taken, resulting in a tripartite graph. We define the union operator on adjacency matrices to give the same result as a graph join (Harary, 1994) on the graphs they encode,

$$\begin{aligned} A^{(i,j)} \cup B^{(i)} &\equiv G^{(i,j)} \cup \tilde{G}^{(i)} \\ &\equiv (M_i, N_i, O_i, E^{(j)}, \tilde{E}^{(i)}). \end{aligned} \tag{6}$$

The corresponding adjacency matrix is then row-normalized to form a transition matrix, and random walks of either length 2 or 3 depending on the adjacency matrix are performed.

For example, an author may refer to *mouse Ly9* and then later just *Ly9* while still referring to the mouse version of Ly9. The second term in Equation (7) allows the transfer of information between gene mentions with the same surface form. Here the second Ly9 (Fig. 1) can reach identifier 10 090 through a three-step random walk by first walking to a previous mention of Ly9 and then on from there. This allows use of the contextual information of the previous Ly9 in making a decision about the second mention. We allow spreading of information in two distinct directions (from earlier/later mentions to later/earlier mentions).

The two tripartite graphs are then combined with the value at position *mo* stating the probability of ending at taxonomy identifier *o* starting from gene mention *m*. The probability that a specific gene mention *m* is associated with taxonomy identifier *o* is given by,

$$P(O=o \mid M=m, D_i, \boldsymbol{\theta}) \propto \left(\left(C_1 \sum_{j \in J} \left(\theta_j A^{(i,j)} \right) \cup B^{(i)} \right)^2 + \sum_{\substack{M_i \equiv M_j \\ i \neq j}} \sum_{k \in K} \left(C_k \left(\left(\theta_k A^{(i,k)} + \sum_{j \in J} \left(\theta_j A^{(i,j)} \right) \right) \cup B^{(i)} \right)^3 \right) \right)_{mo}. \tag{7}$$

The first term in the likelihood equation corresponds to the two-step random walks through the network and the second term to three-step random walks. A gene mention is assigned the taxonomy identifier with the highest probability in its row in the final tripartite adjacency matrix. If the maximum value is not unique (i.e. there is a tie), no classification is made. We only allow edges between gene mentions where they have the same surface form (e.g. $M_i \equiv M_j$). Random walks longer than length 3 are not possible in this network topology. C_1 is the normalization vector for the transition matrix for the two-step random walk random and C_k is the normalization vector for the transition matrices for three-step random walk (where each matrix represents a different direction that the walk can proceed in). Each entry in the normalization vector is calculated as the reciprocal of the equivalent row sum for each row in A^*. We then can the calculate the likelihood using Equation (7).

The parameters (θ_*) of the individual components of the mixture model $A^{(i,*)}$ are estimated by minimizing the negative log-likelihood (Lk) of the model [Equation (8)]. T_i is the set of annotations for document i in the portion of the corpus that is used as a training set T,

$$Lk = -\sum_i \sum_{mo \in T_i} \log P(O=o \mid M=m, D_i, \boldsymbol{\theta}) \tag{8}$$

3 RESULTS

Our method was evaluated using 5-fold cross-validation over the DECA corpus. We assume that gene mentions are provided and the aim is to assign a single taxonomy identifier to individual gene mentions. We calculated the micro-precision, micro-recall, micro-F1, macro-precision, macro-recall and macro-F1 over the different versions of the DECA corpus. Precision is calculated as the total number of correctly assigned gene mention–taxonomy identifier pairs divided by the total number of assigned pairs. Recall is calculated as the total number of correctly assigned 'gene mention-taxonomy identifier' pairs divided by the total number of pairs annotated in the corpus. F_1 is the unweighted harmonic mean of precision and recall. Micro-averaged metrics are calculated globally over all potential classes. These metrics provide a measure of how well our method is doing over all the instances in the corpus. However, training data for multi-classification tasks can exhibit a class bias, where certain classes are over-represented. This can lead to high micro-averaged performance as the method may work well on the classes with a large number of instances, but the method may show poor performance on rarer classes with few or no instances in the training set. Macro-averaged metrics provide measures where this class bias is taken into consideration by explicitly assuming that the class sizes are equal. This is achieved by taking the mean of the desired performance metric over the potential classes. Methods that report high micro-average and low macro-average performance suggest that the method is susceptible to local over-fitting.

Our method performs well compared to the current state-of-the-art methods reported previously by Wang *et al.* (2010), in terms of both macro- and micro-averaged metrics (Table 3). Randomization testing was performed and the method was found by significantly outperforming the previously reported ML method in terms of macro-averaged performance and the RELATION method in terms of both macro- and micro-averaged performance (Table 4).

In comparison to the method in Wang *et al.* (2010), which only used a species tagger and prefix resolution system, using a different species tagger and additionally using cell lines and MeSH terms as extra sources of evidence was found to improve overall performance. A number of abstracts contain no species mentions and the only clue with regard to the species is the assigned MeSH terms. If MeSH terms were not used in the model, the micro-averaged F_1 would decrease to 0.7344 and 0.7572, for corpus A_2 and B, respectively [we note that our method resolves many more species than the method of Wang *et al.* (2010) and hence a decrease in micro-averages is expected]. Not including cell line mentions in the evaluation resulted in a decrease to 0.8459 and 0.8840 for corpus A_2 and B in terms on F_1. This shows that in addition to identifying species names, using other sources of evidence improves the overall performance. Our method enables the integration of these additional sources of evidence in a coherent and easily understandable manner.

Table 3. Micro and macro-averaged 5-fold cross-validation performance metrics (precision, recall, F_1) over versions A and B of the DECA corpus

	Corpus	Precision	Recall	F_1
Micro-average	A_1	0.8243 (0.6768–0.9406)	0.6998 (0.4967–0.6998)	0.7561 (0.5729–0.8773)
	A_2	0.8693 (0.8215–0.9108)	0.8261 (0.7495–0.9085)	0.8466 (0.7937–0.9097)
	B	0.9112 (0.8535–0.9571)	0.8596 (0.7571–0.9309)	0.8842 (0.8125–0.9438)
Macro-average	A_1	0.3412 (0.2746–0.4457)	0.5205 (0.4303–0.6027)	0.3621 (0.2939–0.4852)
	A_2	0.3252 (0.2200–0.4034)	0.5072 (0.4254–0.5615)	0.3689 (0.2791–0.4383)
	B	0.4748 (0.3467–0.5881)	0.6260 (0.5012–0.7527)	0.5148 (0.4007–0.6264)

Values inside brackets show the range of the performance measure over the 5-folds. A_1 shows the performance of the method using the *species words* identified by the NaCTeM Species Word Detector. A_2 using our custom *species word* detection and normalization component.

Table 4. Significance test results between the various performance metrics (micro and macro average **Pr**ecision, **Re**call and **F1**) on A_1 using our method and the various methods detailed in Wang *et al.* (2010)

	Micro-average			Macro-average		
	Pr	Re	F1	Pr	Re	F1
ML	-	-	-	+	+	+
RELATION	+	+	+	+	+	+
HYBRID	-	-	-	-	+	N

- denotes that our method performs significantly worse than a previously proposed method, + denotes significantly better than and N denotes no significant difference was observed ($P < 0.05$). Details of the mean and variance for each performance metric are provided in Supplementary Material V.

Table 5. Percentage frequency (%), precision, recall and F_1 for individual classes on corpus A averaged over 5-folds using the *species words* obtained using the NaCTeM species word detector (A_1)

Name (Taxonomy ID)	%	Precision	Recall	F_1
Homo sapiens (9606)	51.70	0.6158	0.5360	0.5304
Mus musculus (10090)	27.42	0.7432	0.7884	0.7387
Drosophila melanogaster (7227)	10.31	0.3965	0.6459	0.3978
Saccharomyces cerevisiae (4932)	8.03	0.2181	0.4338	0.2231
Other (0)	1.12	0.1251	0.4607	0.1899
Rattus norvegicus (10116)	0.80	0.4355	0.8098	0.5251
Escherichia coli K-12 (83333)	0.29	0.0000	0.0000	0.0000
Xenopus tropicalis (8364)	0.13	0.2778	0.5333	0.3651
Caenorhabditis elegans (6239)	0.11	0.1364	0.5000	0.2143
Bos taurus (9913)	0.05	0.0000	0.0000	0.0000
Arabidopsis thaliana (3702)	0.03	0.0000	0.0000	0.0000

We show the performance measures on the individual classes for A_1, A_2, B in Tables 5, 6 and 7, respectively. It should be noted that the method can disambiguate species even when the species has only a small number of instances in the corpus. High performance is achieved even when members of a class are absent from the training data for that fold (as certain classes are only present in one document in the corpus).

Examination of the parameter estimates over the different folds shows that they are relatively stable, with most having a small coefficient of variation (CV) (Table 2). The rank of the parameters

Table 6. Percentage frequency (%), precision, recall and F_1 for individual classes on corpus A averaged over 5-folds using *species words* obtained using our custom *species word* detection and normalization component (A_2)

Name (Taxonomy ID)	%	Precision	Recall	F_1
Homo sapiens (9606)	51.70	0.5778	0.5350	0.5445
Mus musculus (10090)	27.42	0.7064	0.8445	0.7340
Drosophila melanogaster (7227)	10.31	0.5590	0.5593	0.5576
Saccharomyces cerevisiae (4932)	8.03	0.3497	0.5811	0.4011
Other (0)	1.12	0.2010	0.6922	0.3039
Rattus norvegicus (10116)	0.80	0.3463	0.8813	0.4563
Escherichia coli K-12 (83333)	0.29	0.0000	0.0000	0.0000
Xenopus tropicalis (8364)	0.13	0.0000	0.0000	0.0000
Caenorhabditis elegans (6239)	0.11	0.0000	0.0000	0.0000
Bos taurus (9913)	0.05	0.0000	0.0000	0.0000
Arabidopsis thaliana (3702)	0.03	0.0000	0.0000	0.0000

is notably similar over all folds (Spearmans $\rho = 0.9621$ and 0.9718 for corpus A_2 and B, respectively). This shows that the relative importance of the features is highly consistent over all folds, supporting the assumption that authors state the species of a gene in a consistent manner irrespective of the actual species.

4 DISCUSSION

Unlike other classification (e.g. SVMs) approaches, our method does not provide a hard classification ('yes'/'no') and instead provides a probability associated with each classification it makes. This probability reflects both the uncertainty in assigning a *species word* to a gene mention, and in the normalization of a *species word* to an NCBI taxonomy identifier, allowing uncertainty to be propagated. Thus, the reliability of the proposed species identifier can be assessed directly, and we do not artificially inflate confidence in a given classification. Where there is insufficient evidence in the abstract to distinguish between the species of an entity, a Maximum Entropy (MaxEnt) approach is followed where all potential taxonomy identifiers are assigned a uniform probability (no classification is made). Like the relation method proposed by Wang *et al.* (2010), our method does not require separate training data for each class/species and instead assumes that authors propose the species of an entity in a uniform manner over all species. The method is conceptually similar to construction-integration model proposed by Kintsch (1988).

Table 7. Percentage frequency (%), precision, recall and F_1 for individual classes on corpus B averaged over 5-folds

Name (Taxonomy ID)	%	Precision	Recall	F_1
Homo sapiens (9606)	52.64	0.7214	0.6652	0.6739
Mus musculus (10090)	25.04	0.7414	0.8694	0.7822
Drosophila melanogaster (7227)	9.84	0.4601	0.5816	0.4985
Saccharomyces cerevisiae (4932)	8.81	0.5413	0.8382	0.6279
Rattus norvegicus (10116)	1.58	0.5572	0.9529	0.6700
Xenopus laevis (8355)	0.34	0.0833	0.2500	0.1250
Gallus gallus (9031)	0.23	0.3750	0.5000	0.4167
Bos taurus (9913)	0.17	0.3571	0.5000	0.4167
Drosophila subobscura (7241)	0.15	1.0000	1.0000	1.0000
Other (0)	0.13	0.0648	0.5000	0.1148
Caenorhabditis elegans (6239)	0.12	0.0000	0.0000	0.0000
Drosophila virilis (7244)	0.10	0.6667	1.0000	0.8000
Ovis aries (9940)	0.08	0.8333	1.0000	0.9091
Escherichia coli (562)	0.07	0.8000	1.0000	0.8889
Brassica napus (3708)	0.07	0.6667	1.0000	0.8000
Canis lupus (9615)	0.07	1.0000	1.0000	1.0000
Zea mays (4577)	0.05	0.2727	1.0000	0.4286
Schizosaccharomyces pombe (4896)	0.05	1.0000	1.0000	1.0000
Candida albicans (5476)	0.05	0.5000	0.3333	0.4000
Danio rerio (7955)	0.05	1.0000	1.0000	1.0000
Pseudomonas aeruginosa (287)	0.03	1.0000	1.0000	1.0000
Arabidopsis thaliana (3702)	0.03	0.0000	0.0000	0.0000
Ricinus communis (3988)	0.03	0.0000	0.0000	0.0000
Scaptodrosophila lebanonensis (7225)	0.03	1.0000	1.0000	1.0000
Drosophila mauritiana (7226)	0.03	0.6667	1.0000	0.8000
Drosophila simulans (7240)	0.03	0.0000	0.0000	0.0000
Oryctolagus cuniculus (9986)	0.03	0.5000	0.5000	0.5000
AMV (11866)	0.03	0.3333	1.0000	0.5000
Mortierella alpina (64518)	0.03	0.0000	0.0000	0.0000
Oryza sativa (4530)	0.02	1.0000	1.0000	1.0000
Drosophila sechellia (7238)	0.02	0.5000	1.0000	0.6667
Teleost fish (70862)	0.02	0.0000	0.0000	0.0000
Tetraodon nigroviridis (99883)	0.02	0.0000	0.0000	0.0000

In the absence of any *species words*, it is not possible to propose a taxonomy identifier for any gene mentions within that abstract. In future, it may be possible to use an imputation step where we could probabilistically assign an identifier based on the frequency of species in PubMed, or by looking at its candidates in BioThesaurus (Liu *et al.*, 2006) and assigning a uniform probability over all potential taxonomy identifiers.

The performance of any species disambiguation method is highly dependent on the performance of the components responsible for tagging and normalizing *species words*. Here, we used both the NaCTeM Species Word Detector and the LINNAEUS tagger for species named entity recognition (NER), and any publicly available/custom-built species tagger could be used [e.g. TaxonGrab (Koning *et al.*, 2005) or the OrganismTagger (Naderi *et al.*, 2011)]. One of the issues with species name identification is that authors can use the genus name as shorthand for the species name. Performance of the method would improve if genus names were identified and normalized to the major model organism in that genus [e.g. identifying and normalising *Arabidopsis* to *Arabidopsis thaliana* (3702)]. However, this would require a species tagger capable of distinguishing whether the genus mention is referring to an individual species or the genus as a whole in order to not adversely impact precision.

In cases involving co-ordination (e.g. the human and mouse KLHLI genes (PMID:10888605), both elements of the coordination are assigned to the named entity with similar (but not necessarily equal) probabilities of being the correct one. In some cases, this leads to false positives. Unlike the previous work by Wang *et al.* (2010), we have not investigated the use of the computationally expensive syntactic parse trees; however, these could possibly be used as an additional feature in our model and enable us to handle coordinated entities. We find, however, that good performance can be achieved without incurring the heavy computational burden of generating parse trees of sentences in the MEDLINE abstracts. We believe that in future species disambiguation should be considered as a multi-label classification problem, where a gene mention can have many taxonomy identifiers assigned to it, although it would require another complete re-curation (and extension) of the DECA corpus to make this possible. It is not possible to include 'bag of words' style features in the proposed graph mixture model and so we have not investigated their use here.

The results show that our method attains perfect precision and recall for some classes and performs poorly for others. Some of the low incidence species in Table 7 occur in only 1-fold (sometimes in only one document). Even though the training information for these specific classes is non-existent, our method achieves remarkable performance on some of these classes. The main reason for poor performance on these low incidence classes is that the evidence for other classes in an abstract overwhelms the evidence for that class (i.e. the correct class has a single *species word* while the incorrect class has many *species words* close to the gene mention). Examination of misclassifications revealed two situations where our method can perform poorly: in cases where the abstract describes the result of a transfection experiment where a gene from one species is transfected into another species; and when an abstract is talking about the interactions between viral and host genes, the host species is implicitly mentioned in the name of the virus and is not identified by the species NER component, leading to possible misclassification of host genes as viral genes. Beyond this no systematic sources of bias were detectable.

In principle, we could also use higher order information in this task. For example, it has previously been noted that other *species words* such as the names of diseases (e.g. Creutzfeldt--Jakob disease and Li–Fraumeni syndrome) are indicative of human; however, we leave the identification, mapping and evaluation of such clues for future work. Previous work in gene normalization has found that examining co-authorship networks (Farkas, 2008) can improve performance in gene normalization tasks. As researchers would tend to work on a limited number of organisms during their career (e.g. 'researcher *A* only works on species '*X*'), it seems that this information could also be exploited in distinguishing the species of named entities, but this requires us to identify authors reliably (which would be simplified immeasurably by researcher IDs).

5 CONCLUSION

Our method performs well in terms of both macro- and micro-averaged performance of previous methods and shows significant improvements over the RELATION method proposed by Wang *et al.* (2010). Most importantly, our approach, once the reliable

corpora are in place, can be applied in an automatic fashion without any further user intervention, which will greatly aid its employment in the context of novel organisms or contexts which consider broad phylogenetic panels. We are currently integrating this approach with our TM pipeline (`Serendipity`) in order to generate species specific PPI networks for a variety of organisms (*Saccharomyces cerevisiae*, *Escherichia coli* and for various host–pathogen interaction/infection networks). The method is quite generic and could potentially be used for a host of relation classification tasks, such as assigning genes to PATO identifiers using phenotype clues identified in an abstract potentially allowing the automatic association of genes to phenotypic information. This is already a taxing problem for human annotators, but computational approaches such as proposed here are the only really viable option in order to cope with the wealth of information being generated in genomics and systems biology.

ACKNOWLEDGEMENTS

We thank Xinglong Wang (National Centre for Text Mining, UK) for his help regarding questions about the DECA corpus and assistance in its recuration, and to Paul Kirk and William Kelly for comments and discussion.

Funding: BBSRC/Unilever CASE studentship (to N.H.); Royal Society Research Merit Award holder (to M.P.H.S.).

Conflict of Interest: none declared.

REFERENCES

Ananiadou,S. *et al.* (2006) Text mining and its potential applications in systems biology. *Trends Biotechnol.*, **24**, 571–579.

Chen,L. *et al.* (2005) Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*, **21**, 248–256.

Farkas,R. (2008) The strength of co-authorship in gene name disambiguation. *BMC Bioinformatics*, **24**, i126–i132.

Fields,S. and Johnston,M. (2005) Cell biology. Whither model organism research? *Science*, **307**, 1885–1886.

Gerner,M. *et al.* (2010) LINNAEUS: a species name identification system for biomedical literature. *BMC Bioinformatics*, **11**, 85.

Hahn,U. *et al.* (2008) An overview of JCoRe, the JULIE lab UIMA component repository. In *Proceedings of the LREC'08 Workshop Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP*, European Language Resources Association, pp. 1–7.

Hakenberg,J. *et al.* (2004) Finding kinetic parameters using text mining. *Omics J. Integr. Biol.*, **8**, 131–152.

Hakenberg,J. *et al.* (2008) Inter-species normalization of gene mentions with GNAT. *Bioinformatics*, **24**, i126–i132.

Hakenberg,J. *et al.* (2011) The GNAT library for local and remote gene mention normalization. *Bioinformatics*, **27**, 2769–2771.

Harary,F. (1994) *Graph Theory*. Addison-Wesley, Reading, MA.

Harmston,N. *et al.* (2010) What the papers say: Text mining for genomics and systems biology. *Hum Genomics*, **5**, 17–29.

Hoffmann,R. *et al.* (2005) Text mining for metabolic pathways, signaling cascades, and protein networks. *Sci. STKE*, **2005**, pe21.

Huang,M. *et al.* (2011) GeneTUKit: a software for document-level gene normalization. *Bioinformatics*, **27**, 1032–1033.

Kano,Y. *et al.* (2009) U-Compare: share and compare text mining tools with uima. *Bioinformatics* , **25**, 1997–1998.

Kappeler,T. *et al.* (2009) TX task: automatic detection of focus organisms in biomedical publications. In *BioNLP '09: Proceedings of the Workshop on BioNLP*, Association for Computational Linguistics.

Kintsch,W. (1988) The role of knowledge in discourse comprehension: a construction-integration model. *Psychol. Rev.*, **95**, 163–182.

Koning,D. *et al.* (2005) TaxonGrab: extracting taxonomic names from text. *Biodivers. Informat.*, **2**, 79–82.

Krallinger,M. *et al.* (2008) Overview of the protein-protein interaction annotation extraction task of Biocreative II. *Genome Biol.*, **9** (Suppl. 2), S4.

Leitner,F. *et al.* (2010) An overview of Biocreative II.5. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **7**, 385–399.

Liu,H. *et al.* (2006) Biothesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics*, **22**, 103–105.

Mons,B. (2005) Which gene did you mean? *BMC Bioinformatics*, **6**, 142–145.

Naderi,N. *et al.* (2011) OrganismTagger: detection, normalization, and grounding of organism entities in biomedical documents. *Bioinformatics* , **27**, 2721–2729.

Romano,P. *et al.* (2009) Cell line data base: structure and recent improvements towards molecular authentication of human cell lines. *Nucleic Acids Res.*, **37**, D925–D932.

Sarntivijai,S. *et al.* (2008) A bioinformatics analysis of the cell line nomenclature. *Bioinformatics*, **24**, 2760–2766.

Schwartz,A.S. and Hearst,M.A. (2003) A simple algorithm for identifying abbreviation definitions in biomedical text. *Pac. Symp. Biocomput.*, **8**, 451–462.

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

Tsuruoka,Y. *et al.* (2005) Developing a robust part-of-speech tagger for biomedical text. *Lect. Notes Comput. Sci.*, **3746**, 382–392.

Verspoor,K. *et al.* (2010) Exploring species-based strategies for gene normalization. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **7**, 462–471.

Wang,X. and Matthews,M. (2008) Distinguishing the species of biomedical named entities for term identification. *BMC Bioinformatics*, **9** (Suppl. 11), S6.

Wang,X. *et al.* (2010) Disambiguating the species of biomedical named entities using natural language parsers. *Bioinformatics*, **26**, 661–667.