Gene Name Disambiguation using Multi- Scope Species Detection

Journal:	*IEEE/ACM Transactions on Computational Biology and Bioinformatics*
Manuscript ID:	TCBBSI-2013-10-0295
Manuscript Type:	SI: GIW 2013
Keywords:	H.2.8.l Text mining < H.2.8 Database Applications < H.2 Database Management < H Information Technology and Systems, I.2.7.h Text analysis < I.2.7 Natural Language Processing < I.2 Artificial Intelligence < I Computing Methodologies, J.3.a Biology and genetics < J.3 Life and Medical Sciences < J Computer Applications

Gene Name Disambiguation using Multi-Scope Species Detection

Jui-Chen Hsiao[1†], Chih-Hsuan Wei[2†] and Hung-Yu Kao[1,2*]

Abstract—Species detection is an important topic in the text mining field. According to the importance of the research topics (e.g., species assignment to genes and document focus species detection), some studies are dedicated to an individual topic. However, no researcher to date has discussed species detection as a general problem. Therefore, we developed a multi-scope species detection model to identify the focus species for different scopes (i.e., gene mention, sentence, paragraph, and global scope of the entire article). Species assignment is one of the bottlenecks of gene name disambiguation. In our evaluation, recognizing the focus species of a gene mention in four different scopes improved the gene name disambiguation. We used the species cue words extracted from articles to estimate the relevance between an article and a species. The relevance score was calculated by our proposed Entities Frequency-Augmented Invert Species Frequency (EF-AISF) formula, which represents the importance of an entity to a species. We also defined a relation guide factor (RGF) to normalize the relevance score. Our method not only achieved better performance than previous methods but also can handle the articles that do not specifically mention a species. In the DECA corpus, we outperformed previous studies and obtained an accuracy of 88.22%.

Index Terms—Biomedical text mining, gene name disambiguation, focus species detection

1 Introduction

INFORMATION extraction from biomedical literature sources has been studied for twenty years. One important topic, which has been discussed for several years, is the “focus discussion subject detection” for a specific target, such as “detecting the focus species for articles” [1, 2], “gene function assignment for human genes” [3], or “protein-protein interaction evidence in sentences” [4].

There are two issues that previous studies have not discussed in much detail. The first is document triage for the organism group. Users working with organism groups need to separate articles into specific species categories to narrow down the articles that they need to survey [5-7]. However, NCBI taxonomy (http://www.ncbi.nlm.nih.gov/taxonomy), which includes 220,000 species as of June 10, 2013, has a high dimension structure [8] and lacks a corpus. Therefore, species detection has not been well-researched [1, 2, 9]. The other issue is species assignment, which is a critical issue of gene name normalization. Most methods cannot handle neglected species well. In addition, previous methods focused only on abstracts rather than on the full text of articles.

One of the critical issues of focus species detection is document focus species identification, which identifies the topic species of a particular article. Two previous methods of handling this challenge are dictionary-based matching with a voting strategy [9, 10] and a statistic-based method with an incremental mining strategy [1, 2]. Dictionary-based matching cannot identify the focus species if no species is mentioned. The statistic-based method depends on the training corpus. Based on the existing focus species corpora, this method can only handle four species (i.e., human, fly, yeast, and mouse).

Another important topic in species detection is species assignment for gene mentions because it is a very important step in gene normalization. The Critical Assessment of Information Extraction Systems in Biology (BioCreative), a bi-yearly competition in the field of biological text mining, includes several important biomedical text mining issues. The goal of the GN tasks in BioCreative II, II.5, and III [11-13] is to map the genes or proteins mentioned in the literature to standard database (Entrez Gene) identifiers. In BioCreative II.5 and III gene normalization (GN) tasks, many participants note that accurate species assignment is one of the critical keys in avoiding gene normalization ambiguity [11-13].

In previous studies, gene normalization studies have focused on the case where the species information is provided. Hakenberg [13, 14] developed a dictionary-based gene-name normalization system (GNAT) and obtained the best performance for the GN task in BioCreative II. GNAT is the first method to focus on cross-species normalization, and it can handle 13 different species with an F-measure of 81.4%. Wermter [15] also developed a statistical method, GENO, by applying a TF-IDF weighting scheme and then calculating semantic similarity scores to resolve ambiguous terms. Unlike GNAT, GENO only focuses on the human gene. Thus, it is developed by samples and is easy to rebuild.

Since the difficulty of species assignment of genes has

- *Jui-Chen Hsiao is with the Institute of Medical Informatics, National Cheng Kung University, Tainan, Taiwan, R.O.C. E-mail: vincent770608@gmail.com*
- *Chih-Hsuan Wei is with the Institute of Computer Science and Information, National Cheng Kung University, Tainan, Taiwan, R.O.C. E-mail: chwei@ikmlab.csie.ncku.edu.tw*
- *Hung-Yu Kao is with the Institute of Computer Science and Information and Institute of Medical Informatics, Engineering, National Cheng Kung University, Tainan, Taiwan, R.O.C. E-mail: hykao@mail.ncku.edu.tw*
- **Corresponding author*
- *†These authors contributed equally to this work.*

been addressed, the interactor normalization task (INT) of BioCreative II.5 is the first international competition that is dedicated to inter-species gene normalization without providing the species information. Because the characteristics of inter-species GN are difficult to address and the task uses the full text, the normalization results appear to be low[16]. Hakenberg et al. [17] refined their previous work and integrated GNAT and BANNER. They obtained the highest precision in BioCreative II. In the same competition, Chen et al. [18] developed a Biological Literature Miner (BioLMiner) system to handle the INT and IPT (interaction pair task) tasks. Their system is based on a support vector machine (SVM) and a conditional random field with designed informative features. Verspoor et al. [19] defined a fuzzy dictionary to detect mentions of proteins, and they also described several heuristic strategies to disambiguate species. The AkaneRE [20] is based on the U-Compare system, and it includes sentence boundary detection, tokenization, parsing, named-entity recognition, generation of potential relations, and generation of features for each relation. AkaneRE also assigns confidence scores and ranking of candidate relations, and it obtained the highest recall (68.3%) in whole participations. Dai et al. [21] defined a three-stage normalization algorithm using a ranking method to handle the task and obtained the best AUC, 0.4347.

In 2010, the gene normalization task of BioCreative III [8] focused on an issue similar to the INT task of BioCreative II.5. Kuo [22] developed two context-based dynamic strategies to select dictionary identifiers related to the specific species that appear in a paper and to generate a set of overlapping gene mention variants with nearly perfect recall. Tsai [23] developed a multi-stage gene normalization procedure and a ranking method that exploited information from different paragraphs of a paper. Huang [12] developed a document-level gene normalization software, GeneTuKit, which employs both the local context surrounding gene mentions and the global context in a machine learning classifier from an entire full-text document. Separate from GeneTuKit, Wei [11] developed an inference network method to handle the gene normalization task and obtained a 46.56% F-measure in a manually annotated corpus.

The major cause for the low performance of gene normalization in many studies is poor species assignment. Orthologous genes and/or proteins that belong to different species are identified by different NCBI Entrez Gene identifiers. Before normalizing genes to specific gene identifiers, the species to which they belong must be detected. Current research [24] has developed a multi-level approach for gene normalization. It not only identifies unique genes in textual mentions but also assigns them to families.

Because of the importance of species assignment in gene mentions, several studies have been dedicated to this topic. Wang et al. [25] proposed a hybrid method that combined a supervised classification with a relation extraction model. Their approach can identify the intra-sentential relation between species and gene mention in a sentence. However, their method does not function well if no species mention co-occurs with the gene mentions in a sentence. Similarly, their method could not handle the articles that had no species mentions (17% in the DECA corpus) and simply assigned “human” as a default. To address this problem, Harmston et al. [26] used MesH terms, which were annotated manually to obtain the additional species information for species assignment. Mu et al. [27] defined a dictionary-based prototype for calculating the matrix similarity between tokens and species and applied an imbalanced learning method to learn the evidence from referring to a specific species from the dictionary and the training corpus. SR4GN [28], an upgraded version of the species assignment module of GenNorm [11], is a hybrid of a statistical method and dictionary-based matching with a heuristic strategy. By the successful combination of a statistical method and a heuristic method, state-of-the-art results can be obtained.

The above conclusion on the importance of species detection in different cases (e.g., focus species document triage and species assignment for gene mentions) led us to define a multi-scope species detection method that can handle different scopes in literature for different applications. Our primary goal was to identify the species for each gene mention even when there is no species information in the article. We proposed a relational guide factor (RGF) to enhance the capability of the species detection method for species assignment of gene mentions. Our method resolves the mapping problem between gene and species. Unlike previous studies [10, 27, 28], our approach focuses on the full-text article (e.g., PMC articles) structure, including several paragraphs. More specifically, our defined method considers the focus species evidence for different scopes (i.e., the scope of whole paper, paragraph scope, sentence scope, and noun phrase scope.). This method can identify the most-discussed species for a target (e.g., gene, paragraph) in an article.

This study is useful in the biomedical text mining field. The method can do more than assign species for gene mentions. For different purposes, it can detect the focus species for document triage for different organism groups (e.g., TAIR, RGD, Wormbase), or it can be used to detect the animal that has been used in the experiments in vivo by mining the experiment paragraph.

2 Method

2.1 Overview of Our Method

Briefly, our method consists of three steps. The first part is the pre-processing of each gene mention. This step includes tokenization, cue word extraction, and distillation. The second step is the estimation of focus species by our defined coefficient, the entity frequency-augmented invert species frequency (EF-AISF), to calculate the relevance between cue words and species. The species with the highest correlation coefficient is chosen as the probable focus species. However, some orthologous genes are usually discussed in the same research articles, such as the human gene, which uses mice for wet experiments. Considering the co-occurrence of species pairs, we defined a relational guide factor (RGF) to normalize the species coefficient, thus enhancing the capability of species detection. The purpose of the last step is to assign appropriate species to gene mentions. We defined a multi-scope species assignment strategy to find the most suitable species for a gene mention. For each gene mention, the strategy collects the species evidence in the different scopes. Then, the species with the strongest evidence is assigned to the gene mentions.

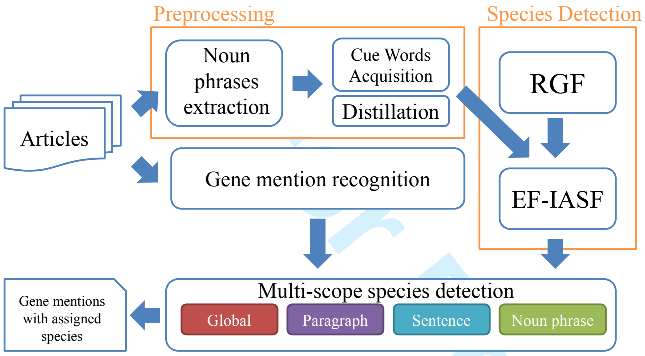

Fig. 1. Flow chart of our proposed multi-scope species detection method

2.2 Species Indication for Multiple Scopes

For the different purposes of species detection for different targets in an article, we defined four different scopes (i.e., global, paragraph, sentence, and noun phrase). Global scope includes all cue words in the article. In our method, MeSH terms are also considered as cue words in the global scope. Paragraph (e.g., abstract and figure captions) and sentence scopes are more specific with respect to the cue words in one paragraph and one sentence. The noun phrase scope is the basic scope that considers all N-grams (N=1-3) as cue words. The global, paragraph and sentence scopes use only the cue words in noun phrases.

2.3 Noun Phrases Extractions

To extract the noun phrases, we used the Perl module Lingua-EN-Tagger [29], which is a part-of-speech tagger for English natural language processing. To avoid redundancies and over-processing, all noun phrases that are substrings of other noun phrases were ignored. As an example, in PMID 10022127, the noun phrases “presents several functional differences”, “TIF1gamma presents several functional differences”, “functional differences”, “differences”, “presents”, and “several functional differences” are the substrings of the longest noun phrase: “TIF1gamma presents several functional differences”. We only retained the longest noun phrase in this step.

2.4 Cue Word Acquisition from Noun Phrases

Then, we extracted cue words from the noun phrases. There are 3 types of cue words, i.e., tokens, N-grams (N=2, 3) and noun phrases. The longest entity unit is the noun phrase, and the shortest entity unit is the token. For example, the noun phrase “several mammalian tin-ag orthologues” (belong to the human category) should be segmented to “several”, “mammalian”, “by”, “tin”, “ag”, “orthologues”, ”several mammalian”, “mammalian tin”, “tin ag”, “ag orthologues”, “several mammalian tin”, “mammalian tin ag”, ”tin ag orthologues” and “protein encoded by otof”. Those cue words are stored in the human category.

2.5 Distillation of the Cue Words

Most cue words do not provide focus species evidence and may mislead the detection. To focus on the helpful evidence related to the focus species, we defined four rules to filter out the unnecessary words. The first rule is that if a word represents three species or more, e.g., if “interact” represents three species (human, fly, and mouse), then it should be filtered out. The second rule is that if the word only appears in one article, or if it represents different kinds of species in different articles, then it is removed. For example, this rule applies when the phrase “heterotyp interact”, which means “heterotypic interaction”, represents two species (human and fly) in two different articles. The third rule is that if the word appears too many times in noun phrases that do not include gene mentions, then it should be removed. For the last rule, we defined thresholds for each species. We assigned the cue word that appears the majority of the times in only one species to be a standard of this species. For example, “human” has the greatest number (74 times) of mentions in the noun phrases that include human genes, and it appears 511 times in all noun phrases. Therefore, we set the threshold for human as 0.14 (74/511). Using the same calculation for all cue words to generate the species indication evidence, the words are filtered out if that number is under the threshold. As an example, the entity “transport” appears 7 times in the noun phrases that include human genes, but it is mentioned 65 times in all noun phrases. Because the species indication evidence of that word is not stronger than the species threshold, we assume the word “transport” is not a good cue word for the species.

2.6 Entities Frequency-Augmented Invert Species Frequency (EF-AISF)

Unlike species names, most cue words cannot indicate a specific species (e.g., “muscle” indicates all mammals). Therefore, we propose using the entities frequency-augmented invert species frequency (EF-AISF) to estimate the relevance between a cue word and a species. If an entity has a high frequency in a species and rarely appears in other species, this entity is suitable for disambiguating species. The entities frequency (EF_{ij}) is the frequency of occurrence of the entity e_i in the species s_j. The idea behind the augmented invert species frequency $(AISF_{ij})$ is the diversity of the entity e_i in species s_j. A higher AISF may indicate that the entity e_i is a significant species distinguishing entity. The formula is shown below. We define n_{ij} as the number of occurrences of an entity i in a species j, and we define $MAX(n_j)$ as the maximum number of occurrences of all entities. To normalize the distribution of the different species of n_{ij}, n_{ij} is divided by $MAX(n_j)$ as follows:

$$EF_{ij} = \frac{n_{ij}}{MAX(n_j)}$$

Then, we define $AISF_{ij}$ as below, where n_i is the sum

($\sum_j^s n_{ij}$) of the entities i in all species, S_i is the number of species that contain the entity i, and S is the number of species in the entire corpus. AISF is a measure of whether the entity e_i is common or rare across all species, as follows:

$$AISF_{ij} = \log(\frac{S}{S_i} \times \frac{n_{ij}}{n_i})$$

Our defined AISF was designed by referring to the inverse document frequency (IDF). Different from the document frequency $\frac{n_{ij}}{n_i}$ of IDF, we designed $\frac{S}{S_i} \times \frac{n_{ij}}{n_i}$ instead. For example, if two entities, x and y, appear the same number of times within a species, but x appears fewer times than y within another species, AISF can emphasize the evidence of x.

We proposed a noun phrase focus species confidence by summarizing the EF-AISF for all the entities in the noun phrase format and a sentence focus species confidence by summarizing the EF-AISF for all the entities in sentence format. The paragraph focus species confidence is the combination of all sentence focus species confidences in the paragraph. Different from the paragraph focus species confidence, we add the MeSH terms to obtain the global focus species confidence.

After measuring the EF-AISF for each species and cue words pair, we summed the EF-AISF scores of the cue words from the same scope (e.g., $EFAISF_j^S = \sum_i^M EFAISF_{ij}^S$). As shown in Fig. 2, cue words e_2 and e_3 are in sentence 1, which focuses to species s_3. Thus, the two cue words are part of the evidence for s_3 to gene mention G.

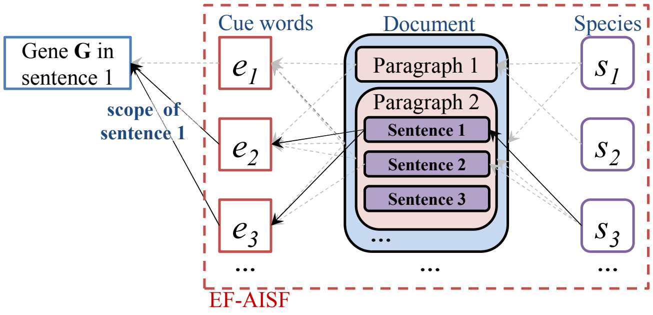

Fig. 2. An example of an EF-AISF calculation for Gene G in the sentence scope

2.7 Relational Guide Factor (RGF)

After completing the calculation, the closest species is assigned to the gene mention. As a sample solution, we assigned the species with the highest score as the scope focus species. However, the focus species in a different scope may not be the same. Therefore, we proposed a relational guide factor (RGF) that can infer the relation between any two species.

The approach we presented can guide the scope's focus species confidence by calculating the RGF of the gene mention pairs in the sentence and paragraph. If $x_{i_{(s1,s2)}}$ is the number of gene mentions of pairs associated with species S1 and S2 in sentence i, $y_{j_{(s1,s2)}}$ is the number of gene mention pairs associated with species S1 and S2 in paragraph j, M is the total number of sentences in the corpus, and N is the total number of paragraphs in the corpus, then the relational guide factor formula is as follows:

$$S\mu_{(S1,S2)} = \frac{\sum_i^M x_{i_{(s1,s2)}}}{M}, P\mu_{(S1,S2)} = \frac{\sum_j^N y_{j_{(s1,s2)}}}{N}$$

$$S\sigma_{(S1,S2)} = \sqrt{\frac{\sum_i^M x_{i_{(s1,s2)}}^2}{M} - \left(\frac{\sum_i^M x_{i_{(s1,s2)}}}{M}\right)^2}$$

$$P\sigma_{(S1,S2)} = \sqrt{\frac{\sum_j^N y_{j_{(s1,s2)}}^2}{N} - \left(\frac{\sum_j^N y_{j_{(s1,s2)}}}{N}\right)^2}$$

where $S\mu_{(S1,S2)}$ and $P\mu_{(S1,S2)}$ are the means of $x_{i_{(s1,s2)}}$ and $y_{j_{(s1,s2)}}$, and $S\sigma_{(S1,S2)}$ and $P\sigma_{(S1,S2)}$ are the standard deviations of $x_{i_{(s1,s2)}}$ and $y_{j_{(s1,s2)}}$ in the corpus. The sentence relational guide factor (SRGF) and paragraph relational guide Factor (PRGF) are as follows:

$$SRGF_{(S1,S2)} = \frac{S\sigma_{(s1,s2)}}{S\mu_{(s1,s2)}}, EFAISF_k^S = \frac{EFAISF_j^S}{SRGF_{(j,k)}}$$

$$PRGF_{(S1,S2)} = \frac{P\sigma_{(s1,s2)}}{P\mu_{(s1,s2)}}, EFAISF_k^P = \frac{EFAISF_j^P}{PRGF_{(j,k)}}$$

Therefore, when S1 and S2 are closer, the score is smaller. This shows that two more closely related species more often appear together in the same sentence or the same paragraph. An example of a co-occurrence of human gene and mouse gene mentions is "Isolation and characterization of cDNA clones for Humly9: the human homologue of mouse Ly9". This shows a human gene, "Humly9", and a mouse gene, "Ly9", occurring in the same sentence. In our observation, the co-occurrence of two human gene mentions in a sentence or paragraph is the most frequent. Following that, the co-occurrence of a human gene and a mouse gene in a sentence or paragraph are the next most frequent when the two genes are different, which occurs because many orthology genes occur both in humans and in mice. Additionally, biologists usually use mice for wet experiments to investigate the possible gene functions in the human body.

3 Results

3.1 Evaluation Dataset

In this study, we used the DECA corpus, published by Wang et al. [25]. There are 644 abstracts in the DECA corpus, collected from BioCreative I Task 1B [30] and the BioCreative II gene normalization task [31]. In total, 6406 gene mentions in the corpus were annotated using the case-insensitive longest match of the species vocabulary supplied with the respective source dataset. Each gene mention was annotated with a specific taxonomy identifier as the standard by the domain experts. Prior to our experiment, we filtered the gene mentions that were not entities (the taxonomy identifiers were assigned -1) or that were associated with "other species" (the taxonomy identifiers were assigned 0). As shown in TABLE 1, the human is the most-discussed species in the DECA corpus. The DECA resource is from the Biocreative 1B and II gene normalization task, which focused on human, mouse, fly and yeast. Therefore, over 97% of the gene mentions are associated with one of these species. After the filtering step, 5974 gene mentions in the DECA corpus remained (366 gene mentions had been tagged -1, and 66 gene mentions had been tagged 0).

TABLE 1.
Percentages of NCBI species mentions in the DECA corpus

NCBI Species	Gene mention Frequency	Rate
H. sapiens (9606)	3203	53.03%
M. musculus (10090)	1504	24.90%
D. melanogaster (7227)	636	10.53%
S. cerevisiae (4932)	508	8.41%
R. norvegicus (10116)	70	1.16%
E. coli K-12 (83333)	18	0.30%
X. tropicalis (8364)	19	0.31%
C. elegans (6239)	7	0.12%
O. cuniculus (9986)	2	0.03%
B. taurus (9913)	3	0.05%
A. thaliana (3702)	2	0.03%
Arthropoda (6656)	1	0.02%
M. zibellina (36722)	1	0.02%
Other species	66	1.09%

To highlight the difficulty of species assignment, we used GenNorm to process the DECA corpus to find the percentage of the mentions in the different scopes. We considered the assignments “First_letter” and “Previous rules” for noun phrase scope and “Front” and “Back” for sentence scope. Other mentions designated as “Major” were considered for whole article scope. As shown in TABLE 2, the percentage of gene mentions associated with the whole article scope was over 65%.

In the whole article scope, no species mention in a sentence can be used for species assignment. In addition, approximately 17% of the articles do not have any species mentions in the abstract. The only way to find the evidence for species assignment is to retrieve and analyze the entire text, which demonstrates why we defined EF-AISF to find indicators from cue words.

TABLE 2.
Percentages of species inferred from scopes in the DECA corpus

	Frequency	Percentage
Noun phrase scope	167	2.80%
Sentence scope	1885	34.35%
Whole abstract scope	3922	65.65%

3.2 Comparison of Performance on Species Assignment with Mu et.al., 2010, Wang et.al., 2009 and SR4GN

To evaluate our method, we applied two measures, micro- and macro-averages, to evaluate performance. Micro-averaging shows the sum of the accuracy for all gene mentions. This approach emphasizes the influence of the more frequent species (e.g., H. sapiens) over the less frequent ones (e.g., M. zibellina). Macro-averaging is the mean of all the species; thus, all species contribute with equal importance. According to the specific characteristics of the macro-average, the macro-average can be used to measure the adaptability of the method to different species. TABLE 3 shows the evaluation results. As shown in TABLE 3, our method performed better than previous methods [10, 27, 28]. Especially for the macro-average, our method is more robust for some neglected species. In our experiment, we performed a five-fold cross validation similar to that of two previous studies [10, 27]. We randomly separated the corpus into five folds. In each run, we calculated the EF-AISFs and RGFs using four of the folds and then tested using the other fold.

TABLE 3.
Comparison of the micro- and macro-averages in the DECA corpus

Species (taxonomy identifier)	Mu,2010	Wang,2009	SR4GN (2012)	Our method
H. sapiens (9606)	0.87	0.86	0.88	0.92
M. musculus (10090)	0.80	0.80	0.80	0.82
D. melanogaster (7227)	0.86	0.87	0.83	0.88
S. cerevisiae (4932)	0.90	0.85	0.89	0.93
R. norvegicus (10116)	0.69	0.59	0.69	0.76
E. coli K-12 (83333)	0.00	0.00	0.00	0.95
X. tropicalis (8364)	0.40	0.36	0.00	0.60
C. elegans (6239)	0.22	0.22	0.43	0.67
O. cuniculus (9986)	0.00	0.00	0.22	0.14
B. taurus (9913)	0.50	1.00	0.00	0.50
A. thaliana (3702)	0.00	0.67	0.14	0.00
Arthropoda (6656)	1.00	0.00	0.00	1.00
M. zibellina (36722)	0.50	0.00	0.00	0.00
Micro-Average	0.8513	0.838	0.8542	0.8822
Macro- Average	0.5196	0.4797	0.3734	0.5854

We also compared the relational guide factor (RGF) with two sample strategies. The first strategy consists of combining the four scope scores for each species and then assigning the species with the highest score as the focus species. The second strategy consists of ranking the scores without using RGF. The focus species in the selected article (PMID: 11086001) is human (taxonomy id: 9606). This article has 23 gene mentions, of which 21 mentions belong to the human species, but two mentions, “betaIV spectrin” and “betaIVSigma1 spectrin”, belong to another species (taxonomy id: 10116). When using the combined scores to assign the species to gene mentions, all of the gene mentions in the same article are assigned to the same species. In our experiment, using RGF achieves better results than using either of the two basic strategies.

TABLE 4.
Evaluation of three strategies for focus species detection

Species	Combining 4	Ranking 4	Relational

(taxonomy identifier)	scope scores w/o RGF	scope scores w/o RGF	guide factor (RGF)
H. sapiens (9606)	0.88	0.89	0.92
M. musculus (10090)	0.80	0.84	0.82
D. melanogaster (7227)	0.81	0.94	0.88
S. cerevisiae (4932)	0.86	0.96	0.93
R. norvegicus (10116)	0.86	0.69	0.76
E. coli K-12 (83333)	0.84	0.74	0.95
X. tropicalis (8364)	0.06	0.00	0.60
C. elegans (6239)	0.00	1.00	0.67
O. cuniculus (9986)	0.43	0.14	0.14
B. taurus (9913)	0.50	0.00	0.50
A. thaliana (3702)	0.50	0.00	0.00
Arthropoda (6656)	1.00	1.00	1.00
M. zibellina (36722)	0.00	0.00	0.00
Micro-Average	0.8508	0.8820	0.8822
Macro- Average	0.5819	0.5533	0.5854

3.3 Evaluation of Species Assignment for Different Scopes

To determine the performance of each scope, we conducted an experiment to assign a species for each scope. The result of each scope assignment is shown in TABLE 5. Our strategy only detected 818 gene mentions (683 correct and 135 incorrect), based on appearances in the noun phrase. The results show that the accuracy increases with the trend from noun phrase scope to paragraph scope. From our experiment, 94.79% of the gene mentions belong to the same species in one article. Therefore, the accuracy in the noun phrase scope is better than in the paragraph scope. In addition, the performance in the paragraph scope is better than in the global scope. According to our observations, if an article has two kinds of species gene mentions, such as in PMID: 11086001, then the MeSH index usually mentions the two species' names (e.g., human and rat). However, the paragraph scope does not provide enough information for correct species detection. We could not determine the focus species by using only MeSH terms. Nevertheless, the MeSH terms are useful when the article has rare species information. The results for each scope performance are shown in TABLE 6. In this experiment, we give every gene mention one answer in all scopes. If the gene mention has no answer in a scope, then we assign the species that is detected by a voting strategy. The results in TABLE 6 show that using our sentence scope or paragraph scope focus species to assign a gene mention is better than using a voting strategy.

TABLE 5.
Evaluation of each scope focus species detection

Scope	Correct answer	Incorrect answer	Scope size	accuracy
Noun phrase	683	135	818	83.50%
Sentence	3551	655	4206	84.43%
Paragraph	5226	736	5962	87.66%
Global	5170	804	5974	86.54%

TABLE 6.
Evaluation of each scope focus species detection using a voting strategy

Scope	Right answer	Incorrect answer	accuracy
Noun phrase	4182	1792	70.00%
Sentence	4697	1277	78.62%
Paragraph	5226	748	87.48%
Global	5170	804	86.54%

3.4 Applying EF-AISF on a support vector machine (SVM)

Machine learning is currently the most popular solution for classification. Because our measure is simple to add as a feature to machine learning methods, we conducted an additional experiment to determine how our measures would perform in a machine learning environment. For each species and gene mention pair, we used a classifier to estimate the relevance and then assigned the species with the highest relevance to the gene mention. We assumed that our developed measure would be useful as an additional feature for a machine learning method focused on species disambiguation. However, with respect to the difference between our method and general statistical methods, our method not only measures the confidence between a species and a gene mention (i.e., EF-AISF) but also considers the relation between any pair of species (i.e., RGF). Unlike other measures, RGF cannot be implemented directly in a machine learning method.

To understand the contribution of EF-AISF, we applied a support vector machine (SVM) to the process gene name disambiguation on the DECA corpus. We performed two runs to compare the SVM models with and without using EF-AISF. We used LibSVM [32], which is one of the most popular implementations for SVM. For the first run of the SVM model, we added five voting strategy features from our previous study [11], as shown in TABLE 7. For species detection, we used SR4GN [28], which defines two robust strategies for inferring genus names and species strains. The results are shown in TABLE 8. Using our proposed measure produced better results than not using it because our method does more than simply measure the confidence between gene and species by EF-AISF. We also calculated the association between species pairs. As shown in TABLE 8, the performance of our method is better than that of the unmodified SVM implementation.

TABLE 7.
Five voting strategy features in SVM

Features	Description
First_letter	The first lowercase letter of the gene name is an abbreviation of its species.
Previous	The species is assigned to a gene entity if the species entity appears before the gene entity.
Front	The species is assigned to the gene entity if the species entity is in front of the gene entity

	in the same sentence. The nearest species is used for assignment.
Back	The species is assigned to the gene entity if the species entity behind the gene entity in the same sentence. The nearest species is used for assignment.
Major	The most discussed species; the default is "human".

TABLE 8.
Comparison of SVM and our method

	Micro-Average
SVM (5 features)	65.21%
SVM + EF-AISF	87.34%
Our method (EF-AISF & RGF)	88.22%

4 Conclusions

Focus species detection is an important research topic for several biomedical text mining issues. This study proposed a robust method to analyze species in a novel way. This study presents two major contributions. The first is multi-scope focus species detection. According to the multi-scope strategy, our method can handle focus species detection for different scopes, including full text (global), paragraph, sentence and noun phrase. Detecting document focus species can help an organism group database society to perform document triage for literature curation. Additionally, detecting focus species in individual paragraphs can identify the animals used for in vivo experiments. Assigning species to gene mentions can also help with gene name disambiguation. The second contribution is the utilization of our multi-scope focus species detection to species assignment of gene mentions. To measure the relevance of a species to a gene mention, we defined a new coefficient, EF-AISF. We also considered the relevance between a species pair by defining a relational guide factor to normalize the confidence between a species and a cue word. The performance of our method was better (88.22% F-measure for the micro-average) and more robust than that of previous studies for many species (58.54% F-measure for the macro-average).

References

[1] C.-H. Wei and H.-Y. Kao, "Represented Indicator Measurement and corpus distillation on focus species detection," in *IEEE International conference on bioinformatics and biomedicine*, 2010, pp. 657-662.

[2] C.-H. Wei and H.-Y. Kao, "Unsupervised Corpus Distillation for Represented Indicator Measurement on Focus Species Detection," *International Journal of Data Mining and Bioinformatics*, vol. 8, pp. 413-426, 2013.

[3] C. Blaschke, E. A. Leon, M. Krallinger, and A. Valencia, "Evaluation of BioCreAtIvE assessment of task 2," *BMC Bioinformatics*, vol. 6, p. S16, 2005.

[4] M. Krallinger, M. Vazquez, F. Leitner, D. Salgado, A. Chatraryamontri, A. Winter, L. Perfetto, L. Briganti, L. Licata, M. Iannuccelli, L. Castagnoli, G. Cesareni, M. Tyers, G. Schneider, F. Rinaldi, R. Leaman, G. Gonzalez, S. Matos, S. Kim, W. J. Wilbur, L. Rocha, A. V. Tendulkar, S. Agarwal, F. Liu, X. Wang, R. Rak, K. Noto, C. Elkan, Z. Lu, R. I. Dogan, J.-F. Fontaine, M. A. Andrade-Navarro, and A. Valencia, "The Protein-Protein Interaction tasks of BioCreative III: classication/ranking of articles and linking bio-ontology concepts to full text.," *BMC Bioinformatics*, vol. 12, p. S3, 2011.

[5] T. C. Wiegers, A. P. Davis, and C. J. Mattingly, "Collaborative biocuration—text-mining development task for document prioritization for curation," *Database (Oxford)*, vol. 2012, p. bas037, 2012.

[6] C.-H. Wei, B. R. Harris, D. Li, T. Z. Berardini, E. Huala, H.-Y. Kao, and Z. Lu, "Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts " *Database (Oxford)*, vol. base041, 2012.

[7] Z. Lu and L. Hirschman, "Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II," *Database (Oxford)*, vol. 2012, p. bas043, 2012.

[8] Z. Lu, H.-Y. Kao, C.-H. Wei, M. Huang, J. Liu, C.-J. Kuo, C.-N. Hsu, R. T.-H. Tsai, H.-J. Dai, N. Okazaki, H.-C. Cho, M. Gerner, I. Solt, S. Agarwal, F. Liu, D. Vishnyakova, P. Ruch, M. Romacker, F. Rinaldi, S. Bhattacharya, P. Srinivasan, H. Liu, M. Torii, S. Matos, D. Campos, K. Verspoor, K. M. Livingston, and W. J. Wilbur, "The Gene Normalization Task in BioCreative III," *BMC Bioinformatics*, vol. 12, p. S9, 2011.

[9] T. Kappeler, K. Kaljurand, and F. Rinaldi, "TX Task:Automatic Detection of Focus Organisms in Biomedical Publications," in *Proceedings of the Workshop on BioNLP*, 2009, pp. 80-88.

[10] X. Wang, J. i. Tsujii, and S. Ananiadou, "Disambiguating the Species of Biomedical Named Entities using Natural Language Parsers," *Bioinformatics*, vol. 26, pp. 661-667, 2010.

[11] C.-H. Wei and H.-Y. Kao, "Cross-species gene normalization by species inference," *BMC Bioinformatics*, vol. 12, p. S6, 2011.

[12] M. Huang, J. Liu, and X. Zhu, "GeneTUKit: a software for document-level gene normalization," *Bioinformatics*, vol. 27, pp. 1032-1033, 2011.

[13] J. Hakenberg, C. Plake, R. Leaman, M. Schroeder, and G. Gonzalez, "Inter-species normalization of gene mentions with GNAT," *Bioinformatics*, vol. 24, pp. i126-i132, 2008.

[14] A. A. Morgan, Z. Lu, X. Wang, A. M. Cohen, J. Fluck, P. Ruch, A. Divoli, K. Fundel, R. Leaman, J. Hakenberg, C. Sun, H.-h. Liu, R. Torres, M. Krauthammer, W. W. Lau, H. Liu, C.-N. Hsu, M. Schuemie, K. B. Cohen, and L. Hirschman, "Overview of BioCreative II gene normalization.," *Genome Biology*, vol. 9, p. S3, 2008.

[15] J. Wermter, K. Tomanek, and U. Hahn, "High-Performance Gene Name Normalization with GENO," *Bioinformatics*, 2009.

[16] F. Leitner, S. A. Mardis, M. Krallinger, G. Cesareni, L. A. Hirschman, and A. Valencia, "An Overview of BioCreative II.5," *IEEE/ACM Transactions On Computational Biology And Bioinformatics*, vol. 7, pp. 385-399, 2010.

[17] J. Hakenberg, R. Leaman, N. H. Vo, S. Jonnalagadda, R. S. C. Miller, L. Tari, C. Baral, and G. Gonzalez, "Efficient Extraction of Protein-Protein Interactions from Full-Text Articles," *IEEE/ACM Transactions On Computational Biology And Bioinformatics*, vol. 7, pp. 481-494, 2010.

[18] Y. Chen, F. Liu, and B. Manderick, "BioLMiner System:

Interaction Normalization Task and Interaction Pair Task in the BioCreative II.5 Challenge," *IEEE/ACM Transactions On Computational Biology And Bioinformatics,* vol. 7, pp. 428-441, 2010.

[19] K. Verspoor, C. Roeder, H. L. Johnson, K. B. Cohen, W. A. B. Jr., and L. E. Hunter, "Exploring Species-Based Strategies for Gene Normalization," *IEEE/ACM Transactions On Computational Biology And Bioinformatics,* vol. 7, pp. 462-471, 2010.

[20] R. Saetre, K. Yoshida, M. Miwa, T. Matsuzaki, Y. Kano, and J. i. Tsujii, "Extracting Protein Interactions from Text with the Unified AkaneRE Event Extraction System," *IEEE/ACM Transactions On Computational Biology And Bioinformatics,* vol. 7, pp. 442-453, 2010.

[21] H.-J. Dai, P.-T. Lai, and R. T.-H. Tsai, "Multistage Gene Normalization and SVM-Based Ranking for Protein Interactor Extraction in Full-Text Articles," *IEEE/ACM Transactions On Computational Biology And Bioinformatics,* vol. 7, pp. 412-420, 2010.

[22] C.-J. Kuo, M. H. Ling, and C.-N. Hsu, "Soft tagging of overlapping high confidence gene mention variants for cross-species full-text gene normalization," *BMC Bioinformatics,* vol. 12, p. S6, 2011.

[23] R. Tsai and P.-T. Lai, "Multi-stage gene normalization for full-text articles with context-based species filtering for dynamic dictionary entry selection," *BMC Bioinformatics,* vol. 12, p. S7, 2011.

[24] S. V. Landeghem, J. Björne, C.-H. Wei, K. Hakala, S. Pyysalo, S. Ananiadou, H.-Y. Kao, Z. Lu, T. Salakoski, Y. V. d. Peer, and F. Ginter, "Large-scale event extraction from literature with multi-level gene normalization," *Plos One,* vol. 8, p. e55814, 2013.

[25] X. Wang, J. i. Tsujii, and S. Ananiadou, "Disambiguating the species of biomedical named entities using natural language parsers," *BIOINFORMATICS,* vol. 26, pp. 661-667, 2010.

[26] N. Harmston, W. Filsell, and M. Stumpf, "Which species is it? Species-driven gene name disambiguation using random walks over a mixture of adjacency matrices," *BIOINFORMATICS,* p. 7, 2011.

[27] T. Mu, X. Wang, J. i. Tsujii, and S. Ananiadou, "Imbalanced Classification Using Dictionary-based Prototypes and Hierarchical Decision Rules for Entity Sense Disambiguation," in *Coling*, 2010, pp. 851-859.

[28] C.-H. Wei, H.-Y. Kao, and Z. Lu, "SR4GN: a species recognition software tool for gene normalization," *Plos one,* vol. 7, p. e38460, 2012.

[29] A. Coburn, "Lingua-EN-Tagger [http://search.cpan.org/~acoburn/Lingua-EN-Tagger-0.19/Tagger.pm]."

[30] H. L, C. M, M. A, and Y. A, "Overview of BioCreAtIvE task 1B: normalized gene lists.," *BMC Bioinformatics,* vol. 6, p. 23, 5/24 2005.

[31] A. A. Morgan, Z. Lu, X. Wang, A. M. Cohen, J. Fluck, P. Ruch, A. Divoli, K. Fundel, R. Leaman, J. Hakenberg, C. Sun, H.-h. Liu, R. Torres, M. Krauthammer, WilliamWLau, H. Liu, C.-N. Hsu, M. Schuemie, K. B. Cohen, and L. Hirschman, "Overview of BioCreative II gene normalization," *Genome Biology,* vol. 9, 01 September 2008.

[32] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology,* vol. 2 p. 27, 2011.

Jui-Chen Hsiao received his M.S. degree in Computer Science at the Institute of Medical Informatics at the National Cheng-Kung University, Taiwan, R.O.C., in 2013.

Chih-Hsuan Wei received his M.S. degree in Computer Science and Information Education at the National Tainan University, Taiwan, R.O.C., in 2007. He received his Ph.D. degree in Computer Science and Information Engineering at the National Cheng-Kung University, Taiwan, R.O.C., in 2013. He joined the Biomedical Text Mining Group at the National Center for Biotechnology Information (NCBI) and has dedicated his work to bioconcept mention normalization and biocuration issues since February 2011. He is a member of IEEE.

Hung-Yu Kao received his B.S. and M.S. degrees in Computer Science from the National Tsing Hua University, Hsinchu, Taiwan, in 1994 and 1996, respectively. In July 2003, he received his PhD degree from the Electrical Engineering Department, National Taiwan University, Taipei, Taiwan. He was a post-doctoral fellow of the Institute of Information Science (IIS), Academia Sinica, from 2003 to 2004. Dr. Kao is currently an Associate Professor of Computer Science and Information Engineering at the National Cheng Kung University. His research interests include web information retrieval/extraction, search engine, knowledge management, data mining, social network analysis and bioinformatics. He has published more than 40 research papers in refereed international journals and conference proceedings. He is a member of IEEE and ACM.