BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 5 2010, pages 661–667
doi:10.1093/bioinformatics/btq002

Data and text mining

Advance Access publication January 6, 2010

Disambiguating the species of biomedical named entities using natural language parsers

Xinglong Wang[1,2,*], Jun'ichi Tsujii[1,2,3] and Sophia Ananiadou[1,2]

[1]National Centre for Text Mining, [2]School of Computer Science, University of Manchester, Manchester, UK and [3]Department of Computer Science, University of Tokyo, Tokyo, Japan

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Text mining technologies have been shown to reduce the laborious work involved in organizing the vast amount of information hidden in the literature. One challenge in text mining is linking ambiguous word forms to unambiguous biological concepts. This article reports on a comprehensive study on resolving the ambiguity in mentions of biomedical named entities with respect to model organisms and presents an array of approaches, with focus on methods utilizing natural language parsers.

Results: We build a corpus for organism disambiguation where every occurrence of protein/gene entity is manually tagged with a species ID, and evaluate a number of methods on it. Promising results are obtained by training a machine learning model on syntactic parse trees, which is then used to decide whether an entity belongs to the model organism denoted by a neighbouring species-indicating word (e.g. *yeast*). The parser-based approaches are also compared with a supervised classification method and results indicate that the former are a more favorable choice when domain portability is of concern. The best overall performance is obtained by combining the strengths of syntactic features and supervised classification.

Availability: The corpus and demo are available at http://www.nactem.ac.uk/deca_details/start.cgi, and the software is freely available as U-Compare components (Kano *et al*., 2009): NaCTeM Species Word Detector and NaCTeM Species Disambiguator. U-Compare is available at http://-compare.org/

Contact: xinglong.wang@manchester.ac.uk

Received on May 6, 2009; revised on December 29, 2009; accepted on December 30, 2009

1 INTRODUCTION

1.1 Overview

The objective of text mining is to automatically extract information from unstructured text and store the information in a form that can be easily accessible by users (Ananiadou *et al*., 2007; Hunter and Cohen, 2006). Storing information in the form of words can cause ambiguity, because a string of words often refers to different meanings in different context. Therefore, a more sensible way, as adopted by many biomedical databases and ontologies, is to organize information by *concept*, where a concept has unambiguous meaning and can be associated with a unique identifier. To make text mining useful for the community of biomedical sciences, one crucial step is to link the *hidden* and *ambiguous* mentions of named entities in text to unique concepts in knowledge bases.

This article presents our study on tackling one source of ambiguity in entity mentions: model organisms. Model organisms are species studied to understand particular biological phenomena. Biological experiments are often conducted on one species, with the expectation that the discoveries will provide insight into the workings of others, including humans, which are more difficult to study directly. From viruses, prokaryotes, to plants and animals, there are dozens of organisms commonly used in biological studies, such as *Escherichia coli*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens* and hundreds more are frequently mentioned in biological research papers. Given an article, it is often essential for readers to understand what organisms the biomedical entities (e.g. proteins) belong to, and on what organisms the experiments are carried out.

1.2 Background and motivation

In biomedical articles, entities of different species are commonly referred to using the same name, causing difficulty for software applications that link an entity to a specific species. For example, without context, '*tumor protein p53*' may associate to over 100 proteins across 23 species.[1] One way to find the species information is to look for MeSH headings, which are a set of keywords attached to a published article. However, not all articles have MeSH headings, and for the ones that have, many do not contain species keywords. Also, MeSH headings cover only the main species reported in the paper, and do not provide information on other species mentioned, whereas for many text mining applications, knowing the species for *every* entity mention is necessary. For example, to identify the proteins (i.e. the underlined terms) in the following sentence, knowing the 'focus' species of the article is not sufficient, as they belong to three different organisms: *human*, *mouse* and *rat*.[2]

> The amounts of *human* and *mouse* CD200R-CD4d3+4 and *r*CD4d3+4 protein on the microarray spots were similar ...

The importance of distinguishing model organisms has been recognized by the community of biomedical text mining. Chen *et al.* (2005) collected gene names from various source databases and calculated intra- and inter-species ambiguities. Overall, only 25 (0.02%) official symbols were ambiguous within the organisms. However, when official symbols from all 21 organisms were

*To whom correspondence should be addressed.

[1]The search was performed over the RefSeq database on July 1, 2009 and the number of species was manually counted.

[2]Prefix 'r' in 'rCD4d3+4' indicates that it is a *rat* protein.

combined, the ambiguity increased substantially to 21279 (14.2%) symbols. Hakenberg *et al.* (2008) and our previous work (Wang and Matthews, 2008) showed that species disambiguation was one of the most important steps for term normalization and identification, which concerned automatically associating mentions of biomedical entities in text to unique database identifiers (Krallinger *et al*., 2008). Also, the task of extracting protein–protein interaction (PPI) in the recent BioCreAtIvE Challenge II workshop (Krallinger *et al*., 2008) required protein pairs to be recognized and normalized, which inevitably involved species disambiguation. More recently, Kappeler *et al*. (2009) discussed a method that identified organism names (referred to in this article as *species words*), with an aim to detect the 'focus' species at document level. The results showed that organism detection was helpful for disambiguation, but their work did not attempt to link organisms to gene entities.

As the technology of natural language parsing advances, it has been successfully adopted for several information extraction tasks, such as automatically finding PPIs in text. The idea is that syntactic structures linking interacting biological entities may have common characteristics that can be exploited by similarity measures or machine learning algorithms. For example, Erkan *et al.* (2007) used the shortest path between two genes according to edit distance in a dependency tree to define a kernel function for extracting gene interactions. Miwa *et al.* (2008) comparably evaluated a number of kernels for incorporating syntactic features, including the bag-of-word kernel, the subset tree kernel (Moschitti, 2006) and the graph kernel (Airola *et al*., 2008), and concluded that combining all kernels achieved better results than using any individual one. Miyao *et al.* (2009) used syntactic paths as one of the features to train a support vector machines (SVMs) model for PPIs and also discussed how different parsers and output representations affected the performance. Targeting the task of disambiguating model organisms at entity level, this article exploits parsing technology and proposes a novel approach that employs syntactic features and transforms a multi-way supervised classification task to a less complex binary relation classification one.

1.3 Task specification

The task concerned in this article is as follows: given a text, in which mentions of biomedical named entities are annotated, we assign a *species tag* to every entity mention. The types of entities studied in this work are genes and gene products (e.g. proteins), and species tags are identifiers from the NCBI Taxonomy (taxon) of model organisms (http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy). Taxon IDs are widely used in major protein and gene databases (e.g. RefSeq, UniProt, GenBank, etc.) and have become the 'canonical' forms to denoting organisms. On the other hand, the technique presented in this article is general: any gazetteer of model organisms can replace the NCBI Taxonomy in the framework. This article focuses on species disambiguation and makes the assumption that the named entities are already recognized. In practice, an automated named entity recognizer [e.g. ABNER (Settles, 2005)] should be used before applying the systems.

2 METHODS

2.1 Species word detection

An informative indicator for species are words that denote names of model organisms in the surrounding context of an entity. For example, *p53* should be tagged as a *mouse* protein, if it appears in the phrase '*mouse* p53'. Another clue is the presence of the species-indicating prefixes in gene and protein names. For instance, prefix '*h*' in entity '*h*Sos-1' suggests that it is a *human* protein. Throughout this article, we refer to such indicative words (e.g. *mouse*, *hSos-1*) as 'species words'. Note that a species 'word' may contain multiple tokens, such as *E.coli*.

We devised a program (Wang and Grover, 2008) to detect such species words: it marks up a word in a document as a species word if it matches an entry in a list of names of organisms. Each entry in the list contains a species word and its corresponding taxon ID, and the list is merged from two dictionaries: the NCBI Taxonomy and the UniProt controlled vocabulary of species (http://www.expasy.ch/cgi-bin/speclist). The NCBI portion is a flattened NCBI Taxonomy (i.e. without hierarchy) including only the identifiers of *genus* and *species* ranks. In total, the merged list contains 356 387 unique species words and 272 991 unique species IDs. The ambiguity in species words is low: 3.86% of species words map to multiple IDs, and on average each word maps to 1.043 IDs. Therefore, we use a simple dictionary look-up method for species word detection.[3] In addition, entity names with prefixes '*h*', '*r*', '*m*', '*d*' and '*y*' are also marked as 'species words'. In biomedical publications, however, a variety of terms, such as names of diseases (e.g. 'breast cancer') and cell lines, can imply organisms. Indeed, one future research direction is on automatic recognition of species-indicating terms.

2.2 Heuristic baselines

One simple approach to assigning a species tag to an entity is by looking for the species words in its context. More specifically, we assign species IDs using *one of* the following rules, each of which is then used as a baseline system:

(1) *previous species word*: if the word preceding an entity is a species word, assign the species ID indicated by that word to the entity.

(2) *species word in the same sentence*: if a species word and an entity appear in the same sentence, assign its species ID to the entity. When more than one species word co-occurs in the sentence, priority is given to the species word at the entity's left with the smallest distance. If all species words occur to the right of the entity, take the nearest one.

(3) *majority vote*: assign the most frequently occurring species ID in the document to all entity mentions.

It is expected that the first rule would produce good precision. However, it can only disambiguate the fraction of entities that happen to have a species word to their *immediate* left. The second rule relaxes the first by allowing an entity to take the species indicated by its nearest species word in the same sentence, which should increase recall but decrease precision. Statistics from our dataset (Section 3.1) show that only 8.22% entities can potentially be resolved by rule 1 and 36.04% by rule 2, while the coverage of *majority vote* is 86.41%.

2.3 Supervised classification baseline

The problem can also be approached as a classification task. Given an entity mention and its surrounding context, a machine learning model classifies the entity into one of the classes, where each class corresponds to a species ID. The model can be trained on a corpus, in which each occurrence of named entities is tagged with a species ID by domain experts. Many machine learning algorithms would fit in this classification framework and we apply a maximum entropy model (http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html). Features used include contextual words, neighbouring species IDs, morphological features of named entities (e.g. prefixes), and all the

[3]When a word maps to multiple IDs, we assign to it the *species* instead of *genus* ID, and between multiple *species* IDs, we choose the most frequent one, as estimated from the BioCreAtIvE II IPS corpus (Section 3.1).

Table 1. Parsers and their input and output format

Parser	Input	Output
C&C (Clark and Curran, 2007)	POS-tagged	GR
ENJU (Miyao and Tsujii, 2008)	POS-tagged	PAS
ENJU-Genia (Hara *et al*., 2007)	POS-tagged	PAS
Minipar (Lin, 1998)	Sentence-detected	Minipar
Stanford (Klein and Manning, 2003)	POS-tagged	SD
Stanford-Genia	POS-tagged	SD

Fig. 1. Predicate-argument structure.

```
(ENJU(noun_arg1(SPECIESWORD orthologue))
(prep_arg12(of orthologue))
(prep_arg12(of ENTITY)))
```

Fig. 2. A syntactic feature obtained from the ENJU parser.

species IDs occurring in the document. Function words and words that consist of only digits and punctuation are filtered out. See Wang and Matthews (2008) for more details on this approach.

This method suffers from a problem that is common for supervised machine learning techniques: a learned model tends to bias towards the dataset that it is trained on (Japkowicz, 2000). In the context of our task, the model would work well on disambiguating the organisms having abundant training data, whereas creating sufficient amounts of training instances for the vast number of organisms would be infeasible. Section 3.2 provides more discussion on this matter.

2.4 Disambiguating species using parsers

We extend the rule-based system described in Section 2.2 by utilizing the paths between words in a syntactic parse tree, and assume that if a path exists between a species word and a named entity, then the entity has the species indicated by the species word. We empirically evaluate a number of parsers by measuring their performance on this task. This task-oriented evaluation approach was also taken by Miyao *et al.* (2009) on the task of extracting PPIs. The parsers used are summarized in Table 1, where ENJU-Genia and Stanford-Genia were trained on the GENIA corpus (Tateisi *et al*., 2005), a treebank of biomedical text.

In more detail, we first select the sentences in which an entity mention and a species word co-occur, and then parse the sentences. If a syntactic path exists between an entity and a species word, the entity is assumed to be of the species indicated by the species word. In cases where there is more than one path between an entity and a species word, the shortest path is chosen.

There are several practical issues to consider when using parsers for this task. First, the text needs to be linguistically preprocessed, which includes sentence boundary detection, tokenization and part-of-speech (POS) tagging. Some parsers supply preprocessing programs, but to ensure a fair parser comparison, we use the same tools (Alex *et al*., 2008) whenever possible. The middle column in Table 1 shows how the input text is linguistically preprocessed with respect to parsers. A POS-tagged text implies that it is also sentence boundary detected and tokenized, and a tokenized text implies that it is sentence detected. All parsers take POS-tagged text as input except for Minipar, which takes only sentences.

Second, the output representations of the parsers are different and we prefer a format that depicts relations between words instead of syntactic constituents. In total, four representations are used: grammatical relation (GR; Briscoe *et al*., 2006), Stanford typed dependency (SD; de Marneffe *et al*., 2006), Minipar's own representation (Lin, 1998) and ENJU's predicate-argument structure (PAS), where a dependency triple (i.e. GR, SD and Minipar) consists of head, dependent and relation, and a PAS triple contains predicate, argument and relation. The right-most column in Table 1 lists the output representation of each parser, and Figure 1 shows a sentence parsed by ENJU in PAS representation.

Third, we store parse trees as graphs and augment nodes on the graphs with biomedical annotation, such as whether a node is part of a species word or entity. This process is non-trivial for Minipar output, because Minipar uses its own tokenizer, which breaks a sentence into tokens differently. For example, protein 'kinesin-14' is treated as one token by our tokenizer, but is split as 'kinesin', '-' and '14' by Minipar. To alleviate this problem, we code rules by hand to make Minipar's tokenization more consistent to ours.

When nodes in a parse tree are annotated, the disambiguation task becomes finding the shortest path between the nodes of entities and the nodes of species words. When an entity or a species word consists of a group of nodes (i.e. tokens), we identify the syntactic head of the entity, and the path connecting to the head node is regarded as the path to the group.

2.5 Classifying relations of entities and species words

A syntactic link between an entity and a species word does not guarantee that the entity has the species indicated by the species word. For example, for the sentence shown in Figure 1, the method presented in Section 2.4 would assign both proteins 'Kip3' and 'Klp67A' the species of *Drosophila*. However, only 'Klp67A' is a *Drosophila* protein. Therefore, we define a species-entity relation as a pair $r = \langle e, s \rangle$, where e is an entity mention and s is a species word, and r is a positive relation if e is of the species indicated by s, and a negative relation otherwise. With manually curated examples, a relation classification model can be trained to rule out negative relations. More specifically, from the sentences in the training dataset that at least one entity and one species word co-occur, we extract pairs of entity and species word and create a set of relations. Then each relation is assigned with a binary label: a relation is positive if the species ID inferred from the species word matches the gold standard species annotation, and is negative otherwise. For example, for the sentence in Figure 1, relation ⟨Kip3, TaxonID: 7215⟩ is a negative instance and the pair ⟨Klp67A, TaxonID: 7215⟩ is a positive one, where TaxonID: 7215 is the species ID for *Drosophila*. From our dataset (Section 3.1), 2154 relations are extracted, of which 74.05% are positive.

For each relation, two types of features are extracted. The first are bag-of-word features, i.e. the words before, between and after the pair of entities, where the words are lemmatized, and the second are syntactic features obtained from parse analysis. Following the PPI extraction method proposed in Sætre *et al*. (2007), we apply a SVM model. For bag-of-word features, a linear kernel is used, and for syntactic paths, a subset tree kernel (Moschitti, 2006) is adopted, for which a path is represented in a flat tree format. The syntactic features used in the final systems (i.e. RELATION and HYBRID in Table 3) are predicate-argument paths obtained from ENJU-Genia.[4] Figure 2 shows a flat tree feature for the negative instance ⟨Kip3, TaxonID: 7215⟩ from Figure 1. Note that all species words (e.g. *Drosophila*) are normalized to 'SPECIESWORD', and entities (e.g. Kip3) to 'ENTITY', which not only reduces the noise in the features, but also makes the model more species generic. In other words, the relation classification model should work on any species including the ones that do not appear in the training portion of the dataset.

[4]We conducted classification experiments using only bag-of-word features, and using bag-of-word features in conjunction with syntactic features from each parser shown in Table 1. The combination of bag-of-word and ENJU-Genia PAS features yielded the best accuracy, and hence was used.

To identify the species of an entity in unseen text, we first parse the sentence and extract pairs of species words and entities, along with the bag-of-word and syntactic features. The trained model is then applied to classify the species–entity relations. The entity mention in a positive relation is tagged with the ID indicated by the species word, while the mentions in negative relations are left untagged. This way, the relation classification approach transforms a complex multi-classification task into a binary classification one. In addition, it can achieve better domain adaptability, because the relation classification model learns the relations between entities and species words, irrespective of their names.

2.6 Spreading strategies

Except for the *majority vote* rule, the approaches described in Sections 2.2, 2.4 and 2.5 are expected to yield low recall, because the rule- and parser-based systems can only detect intra-sentential relations, and hence are only applied to the entities having at least one species word appearing in the same sentence. To improve recall, we 'spread' the species from the disambiguated mentions to their 'relatives', where an entity mention $\bar{e}$ is defined as another mention e's relative under either of the following conditions: (i) if $\bar{e}$ has the same surface form with e; or, (ii) if $\bar{e}$ is an abbreviation or an antecedent of e, where abbreviation/antecedent pairs are detected using the algorithm described in Schwartz and Hearst (2003). Given the set of disambiguated mentions, we then 'spread' their species IDs to their relatives in the same document. After this process, the mentions that do not have any disambiguated relatives would still be missed by the system. In such cases, we use the species determined by the rule of *majority vote* (Section 2.2). We also create a 'hybrid' system (i.e. HYBRID) by applying both the supervised classification and the relation classification models, and take the answer given by the latter when the two systems disagree. To achieve higher precision, the relation classification model in HYBRID does *not* use 'spreading' or 'majority vote' rules.

3 RESULTS

3.1 Data annotation

Among publicly available resources, the corpora provided in the BioCreAtIvE I and II normalization tasks (Hirschman *et al*., 2005; Morgan and Hirschman, 2007) are probably the closest to what we need, in that each abstract is assumed to be species specific. The corpus for BioCreAtIvE I Task 1B (BC1) consists of three subsets, respectively, covering *fly*, *mouse* and *yeast*, while that for BioCreAtIvE II gene normalization (BC2) task covers only *human*. By merging the four datasets, one can create a corpus consisting of the above four organisms. However, there are two reasons that prevent us from performing species disambiguation experiments on the merged dataset as it is. First, entity mentions in text are not manually annotated, and therefore we cannot carry out *entity-level* disambiguation. Secondly, all entities in an abstract are assumed to belong to a specific organism, and this simplifying assumption cannot serve the purpose of this work, which is to show that individual entities may belong to organisms other than the 'focus' species of the document.

We addressed the above problems by manual annotation. As shown in Table 2, in total 730 abstracts were selected from the BC1 and BC2 datasets and merged into one corpus, where genes and gene products were automatically annotated using case-insensitive longest match against the species-specific vocabulary supplied with the respective source dataset. For each gene mention, domain experts were asked to choose one from a list of frequent taxon IDs.

Table 2. Data sources

Main Organism	Source	Abstracts
fly	BC1 Devtest	108
mouse	BC1 Devtest	250
yeast	BC1 Devtest	110
human	BC2 Test	262

The frequent taxon IDs were estimated from the training corpus for the BioCreAtIvE II Protein Interaction Pairs task (IPS), where each article is associated with pairs of UniProt IDs, from which taxon IDs can be easily derived. The IPS training corpus contains 628 full texts, with 6378 UniProt IDs belonging to 62 different species. The diversity of organisms in this corpus highlights the fact that a primary consideration when developing a species disambiguation system should be its ability to disambiguate a wide range of species with minimal additional manual effort. The organisms were then ranked by frequency and the top 10 were selected (as shown in Table 5) for annotators to choose from. The majority of the organisms covered are animals, with only a couple of bacteria and plants. Given the size of the IPS corpus, we believe this frequency list is representative. Meanwhile, we acknowledge that biologists' favorite models may vary greatly. For example, scientists studying *plants* may be more interested in documents on, e.g. *Arabidopsis thaliana*, than those on *human*. During the annotation process, the domain experts can also choose 'Other', when none of the 10 most frequent species apply, or 'not an entity', when he/she believes the automatically recognized entity is a false positive.

As the dictionary-based named entity tagging was unlikely to obtain perfect recall, and the annotators were only allowed to correct false positives but not false negatives, the resulting corpus was expected to miss some gene names.[5] The time saved on annotating gene names, however, was invested in creating more mappings between species IDs and gene names.

We appointed three PhD level biologists to perform annotation, and on average an annotator spent 4 min on each abstract. To avoid being misled, during annotation, they were not aware of the source of the file (i.e. *fly*, *human*, *mouse* or *yeast*), but were allowed to seek help from search engines such as Google and PubMed. To see human experts' performance on this task, 10% of the abstracts were doubly annotated by different annotators. By randomly taking one set of annotation as gold standard, and the other as system output, we calculated the inter-annotator agreement with an F_1 score at 93.58%, indicating that human annotators have high agreement when assigning species to biomedical entities.

In summary, 6402 genes and gene products are automatically identified using the dictionary-based named entity recognizer, where 86 out of 730 abstracts do not appear to contain any entity and are hence removed from the dataset. Also, 2.80% entities are false positives as judged by the annotators (i.e. 'not an entity'). The rest 6223 genes are manually assigned with either a taxon ID or an

[5]We did not use the gold standard text excerpts of genes because the BC1 annotation guidelines state that 'Genes are required to come from the appropriate organism for the specific database' (Colosimo *et al*., 2005), indicating that the curators were asked to annotate only the genes belonging to the organism in question.

'Other' tag, with *human* (9606) being the most frequent at 50.30%. Table 5 shows the species distribution of this dataset.

3.2 Evaluation results

Evaluation was carried out with 5-fold cross-validation, and the systems were compared using averaged precision, recall and F_1 over each species. Micro- and macro-averages of the scores were obtained, where micro-average is the mean of the summation of contingency metrics for all model organisms, so that scores of the more frequent species influence the mean more than those of less frequent ones, and macro-average is the mean over all labels, thus attributing equal weights to each species, and measuring a system's adaptability across different organisms. Table 3 shows the evaluation results. The parser-based (e.g. C&C), relation classification (i.e. RELATION) and the hybrid (i.e. HYBRID) methods are compared with the rule-based (e.g. RULE-MAJORITY) and supervised classification (i.e. ML) baselines. Note that the parser-based systems and relation classification used the spreading strategies as described in Section 2.6. We performed statistical significance tests using randomization (Noreen, 1989) on a number of pairs of methods, and Table 4 shows the results. '+', '−' and 'N' symbols indicate that the method in the corresponding row is significantly better than, worse than or not different ($P < 0.05$) from the method in the column. The six metrics compared are micro-precision, micro-recall, micro-F_1, macro-precision, macro-recall and macro-F_1. For example, the top-left cell in Table 4 shows that using the parser ENJU-Genia significantly improved macro-precision and macro-F_1 over RULE-SP, but decreased the micro-scores and did not make a difference in macro-recall.

Table 3. Averaged 5-fold cross-validation evaluation results

	micro-avg.	macro-avg.
RULE-MAJORITY	72.20 / 62.39 / 66.94	27.77 / 46.67 / 29.32
RULE-SP	74.09 / 64.03 / 68.69	29.77 / 53.81 / 32.20
RULE-SPSENT	72.94 / 63.03 / 67.63	30.22 / 54.76 / 32.93
C&C	73.82 / 63.79 / 68.44	30.51 / 53.59 / 33.43
ENJU	72.98 / 63.06 / 67.66	31.35 / 55.00 / 34.61
ENJU-Genia	73.00 / 63.08 / 67.68	30.11 / 53.42 / 32.97
Minipar	73.02 / 63.10 / 67.69	30.19 / 53.56 / 33.10
Stanford	73.67 / 63.66 / 68.30	31.17 / 56.35 / 34.35
Stanford-Genia	73.48 / 63.50 / 68.13	30.61 / 55.61 / 33.78
ML	82.69 / 82.69 / 82.69	27.01 / 27.84 / 27.37
RELATION	75.24 / 63.99 / 69.16	31.97 / 55.61 / 34.80
HYBRID	83.80 / 83.80 / 83.80	57.56 / 49.72 / 49.90

Precision/recall/F_1-score, in %.

The rule-based systems set high baselines. In terms of micro-averaged scores, the performance of the parser-based approaches were slightly worse than RULE-SP and comparable with RULE-SPSENT. However, they excelled the rule-based ones as measured by macro-averages. Among the parsers tested, the levels of micro-averaged scores vary slightly, with C&C (Clark and Curran, 2007) in lead. RELATION achieved better micro- and macro-averages as compared with the parser- and rule-based systems, thanks to its relation classification model, which alleviated the problems caused by the oversimplified assumption made by the parser-based approaches: an entity belongs to the species denoted by its closest species word on a syntactic path.

ML outperformed the rule- and parser-based approaches in terms of micro-averaged precision. However, the parser-based, relation classification and hybrid approaches have a clear advantage over ML on macro-averages, indicating their capability in tackling a wider range of organisms. Figure 3 shows the performance of ML, RELATION and HYBRID on individual organisms. The labels on the x-axis denote organisms, ordered by frequency, with smaller numbers indicating more frequent ones. Table 5 lists details of their performance on the most frequent 10 organisms. These statistics reveal that ML can only disambiguate five species that have relatively large amount of training instances, and fails completely on others. This is because the model used by ML was trained on a dataset in which occurrences of some species [e.g. *B.taurus (9913)*] are very sparse. In other words, this is a multi-classification task on heterogeneous and imbalanced datasets, a challenge for a supervised classification model to learn to discriminate enough between classes.

On the other hand, RELATION achieved comparable performance on the frequent organisms, and also worked relatively well on rare ones, displaying its good adaptability across domains. Overall, HYBRID obtained the highest points in nearly every scoring

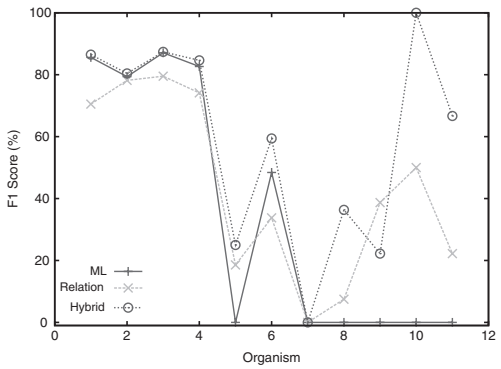

Fig. 3. Performance of ML, RELATION, HYBRID over individual organisms.

Table 4. Results of statistical significance tests between pairs of methods

	ENJU-Genia	C&C	ML	RELATION	HYBRID
RULE-SP	+ / + / + / − / N / −	+ / + / + / − / N / −	− / − / − / + / + / +	− / − / − / − / − / −	− / − / − / − /− / −
ENJU-Genia		− / N / − / N / N / −	− / − /− / + / + / +	− /− / − / − / − / −	− / − / − / − / − /−
C&C			− / − /− / + / + / +	− / − / − / − / − / −	− / − / − / − / − / −
ML				+ / + / + / − / − / −	− / − / − / − / − / −
RELATION					− / − / − / − / + / −

Table 5. The percentage of the species and the micro-averaged F_1 scores (%) of ML, RELATION and HYBRID with respect to each species

Species Name (TaxonID)	Pct (%)	ML	RELATION	HYBRID
Homo sapiens (9606)	50.30	85.60	70.51	86.48
Mus musculus (10090)	26.70	79.38	78.17	80.41
Drosophila melanogaster (7227)	10.01	87.07	79.53	87.37
Saccharomyces cerevisiae (4932)	7.79	82.66	74.13	84.64
Other	1.01	0.00	18.56	25.00
Rattus norvegicus (10116)	0.78	48.42	33.77	59.41
Escherichia coli K-12 (83333)	0.28	0.00	0.00	0.00
Xenopus tropicalis (8364)	0.12	0.00	7.50	36.36
Caenorhabditis elegans (6239)	0.11	0.00	38.71	22.22
Bos taurus (9913)	0.04	0.00	50.00	100.00
Arabidopsis thaliana (3702)	0.03	0.00	22.22	66.67

category, indicating the success in applying relation classification in conjunction with ML.[6] HYBRID integrated the two methods in a crude way, leaving ample room for exploring better combination approaches in the future.

4 DISCUSSION

There are several causes to the large variance in the performance of RELATION and HYBRID on individual species (as shown in Fig. 3). In the case of RELATION, the spreading rule, which relies on the 'one sense per discourse' heuristic, can be overly aggressive and propagate a wrong decision across the whole document. As for HYBRID, since the relation classification model did not use the spreading rule (described in Section 2.6), it would only attempt to correct the entities that have co-occurring species words in the same sentence, affecting 36.04% entities, and therefore only marginally improved the performance over ML. For both systems, errors made by the species word detection program (described in Section 2.1) may also result in false positives and false negatives. For example, the program cannot detect any species word for *E.coli K-12 (83333)* and therefore none of the systems successfully disambiguated this model organism.

The current systems do not tackle coordination, which often infers that more than one species ID applies to a gene/protein mention (e.g. '*mouse* and *human* SPO11'). Feedback from the annotators suggests that this problem is particularly common for the *mammalian* organisms, such as *human*, *mouse* and *rat*. It is the future work to extend the framework and allow assignment of multiple species IDs, possibly also taking into account the hierarchy of model organisms.

5 CONCLUSIONS

This article reports on our experiments and discoveries on the task of disambiguating the model organisms of gene and gene products. We addressed disambiguating the species of entities, instead of that of documents, and a number of approaches were implemented and compared. For evaluation, we developed a gold standard corpus, consisting of 644 MEDLINE abstracts, in which each occurrence of a gene name is manually tagged with an NCBI taxonomy ID, indicating its model organism. As measured by micro-averaged F_1 score, the systems solely relying on syntactic parse analysis did not outperform a baseline system that determines an entity mention's species by looking for its nearest species word in the same sentence. Nevertheless, the parser-based systems achieved higher macro-averaged scores.

A supervised multi-classification approach was also tested, and yielded the second best micro-averaged performance. However, it can only disambiguate the species that have abundant training instances, resulting in a low macro-averaged score of 27.37%. To fix this problem, we proposed a binary relation classification model. Trained on word and syntactic features, the model can filter out erroneous species–entity relations and achieve significantly better micro-averages than the rule- and parser-based systems, and better macro-averages than the supervised classification approach.

Relying on informative keywords in the context, the proposed approaches potentially can detect *any* species. Developing a relation classification model also requires training data. However, a generic binary model can be trained and applied to new domains without the need for extra manually annotated data. This is evidently advantageous over the multi-classification method (i.e. ML), which requires fresh annotated datasets to adapt to new domains. The best overall performance was obtained by combining the strengths of a syntactic parser (i.e. ENJU-Genia), a relation classification model, and a supervised classification model.

[6]Combining ML and RULE-SP yielded a micro-F_1 score of 81.05%, which was significantly lower than the HYBRID system presented here.

ACKNOWLEDGEMENTS

The authors would like to thank the biologists who annotated the species corpus, and Yoshinobu Kano for his help in making the software available in U-Compare.

Funding: Pfizer Ltd.; Joint Information Systems Committee (to UK National Centre for Text Mining).

Conflict of Interest: none declared.

REFERENCES

Airola,A. *et al*. (2008) A graph kernel for protein-protein interaction extraction. In *Proceedings of BioNLP*, Columbus, Ohio.

Alex,B. *et al*. (2008) Assisted curation: does text mining really help? *Pac. Symp. Biocompu.*, **13**, 556–567.

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

Briscoe,E. *et al*. (2006) The second release of the RASP system. In *Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics, Interactive Presentation Sessions*, Sydney, pp. 77–80.

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

Clark,S. and Curran,J.R. (2007) Wide-coverage efficient statistical parsing with CCG and log-linear models. *Comput. Linguist.*, **33**, 493–552.

Colosimo,M. *et al*. (2005) Data preparation and interannotator agreement: BioCreAtIvE task 1B. *BMC Bioinformatics*, **6** (Suppl. 1), S11.

de Marneffe,M.-C. *et al*. (2006) Generating typed dependency parses from phrase structure. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy.

Erkan,G. *et al*. (2007) Semi-supervised classification for extracting protein interaction sentences using dependency parsing. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, Czech Republic, pp. 228–237.

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

Hara,T. *et al*. (2007) Evaluating impact of re-training a lexical disambiguation model on domain adaptation of an HPSG parser. In *Proceedings of the 10th International Conference on Parsing Technology*, Prague, Czech Republic, pp. 11–22.

Hirschman,L. *et al*. (2005) Overview of BioCreAtIvE task 1B: normalised gene lists. *BMC Bioinformatics*, **6** (Suppl. 1), S11.

Hunter,L. and Cohen, K.B. (2006) Biomedical language processing: what's beyond PubMed. *Mol. Cell*, **21**, 589–594.

Japkowicz,N. (2000) Learning from imbalanced data sets: a comparison of various strategies. In *Proceedings of the AAAI Workshop on Learning from Imbalanced Data Sets*, Austin, Texas.

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

Kappeler,F. *et al*. (2009) TX task: automatic detection of focus organisms in biomedical publications. In *Proceedings of BioNLP*, Boulder, Colorado.

Klein,D. and Manning,C.D. (2003) Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, pp. 423–430.

Krallinger,M. *et al*. (2008) Evaluation of text-mining systems for biology: overview of the Second BioCreAtIvE community challenge. *Genome Biol.*, **9** (Suppl. 2), S1.

Lin,D. (1998) Dependency-based evaluation of Minipar. In *Proceedings of Workshop on the Evaluation of Parsing Systems*, Granada, Spain.

Miwa,M. *et al*. (2008) Combining multiple layers of syntactic information for protein-protein interaction extraction. In *Proceedings of the 3rd International Symposium on Semantic Mining in Biomedicine*, Turku, Finland, pp. 101–108.

Miyao,Y. and Tsujii,J. (2008) Feature forest models for probabilistic HPSG parsing. *Comput. Linguist.*, **34**, 35–80.

Miyao,Y. *et al*. (2009) Evaluating contributions of natural language parsers to protein-protein interaction extraction. *Bioinformatics*, **25**, 394–400.

Morgan,A.A. and Hirschman,L. (2007) Overview of BioCreAtIvE II gene normalisation. In *Proceedings of the BioCreAtIvE II Workshop*, Madrid, Spain.

Moschitti,A. (2006) Making tree kernels practical for natural language learning. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, pp. 113–120.

Noreen,E.W. (1989) *Computer Intensive Methods for Testing Hypothesis - An Introduction*. Wiley-Interscience, New York.

Sætre,R. *et al*. (2007) Syntactic features for protein-protein interaction extraction. In *Proceedings of the 2nd International Symposium on Languages in Biology and Medicine*, Singapore.

Schwartz,A.S. and Hearst,M.A. (2003) A simple algorithm for identifying abbreviation definitions in biomedical text. *Pac. Symp. Biocompu.*, **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.

Tateisi,Y. *et al*. (2005) Syntax annotation for the GENIA corpus. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, Jeju Island, Korea, pp. 220–225.

Wang,X. and Grover,C. (2008) Learning the species of biomedical named entities from annotated corpora. In *Proceedings of the International Conference on Language Resources and Evaluation*, Marrakech, Morocco.

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