Classifying Relations for Biomedical Named Entity Disambiguation

Xinglong Wang[†‡] **Jun'ichi Tsujii**[*†‡] **Sophia Ananiadou**[†‡]
[†]School of Computer Science, University of Manchester, UK
[‡]National Centre for Text Mining, UK
[*]Department of Computer Science, University of Tokyo, Japan
{xinglong.wang,j.tsujii,sophia.ananiadou}@manchester.ac.uk

Abstract

Named entity disambiguation concerns linking a potentially ambiguous mention of named entity in text to an unambiguous identifier in a standard database. One approach to this task is supervised classification. However, the availability of training data is often limited, and the available data sets tend to be imbalanced and, in some cases, heterogeneous. We propose a new method that distinguishes a named entity by finding the informative keywords in its surrounding context, and then trains a model to predict whether each keyword indicates the semantic class of the entity. While maintaining a comparable performance to supervised classification, this method avoids using expensive manually annotated data for each new domain, and thus achieves better portability.

1 Introduction

While technology on named entity recognition (NER) matures, many researchers in the field of information extraction (IE) gradually shifted their focus to more complex tasks such as named entity disambiguation and relation extraction. Both tasks are particularly important for biomedical text mining, which concerns automatically extracting facts from the exponentially growing biomedical literature (Hunter and Cohen, 2006). One type of facts is relations between biomedical named entities, such as disease-drug relation, gene-disease relation, protein-protein interaction (PPI), etc. To automatically extract these facts, advanced natural language processing techniques such as parsing have been adopted to analyse the syntactic and semantic structure of text. The idea is that linguistic structures between the 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 they concluded that combining all kernels achieved better results than using any individual one. Miyao et al. (2008) used syntactic paths as one of the features to train a support vector machines (SVM) model for PPIs and also discussed how different parsers and output representations affected the end results.

Another crucial IE task is named entity disambiguation, which concerns grounding mentions of named entities in text to unambiguous *concepts* as defined in some standard dictionary or database. For instance, given a search term *Python*, users may like to see the results grouped into the following categories: *a type of snake*, *a programming language*, or *a film* (Bunescu and Paşca, 2006). One approach to such lexical disambiguation tasks is supervised classification. However, such techniques suffer from the knowledge acquisition bottleneck, meaning that manually annotating training data is costly and can never satisfy the need by the machine learning algorithms. In addition, supervised techniques may not yield reliable results when the distributions of the semantic classes are different in the training and test datasets (Agirre and Martinez, 2004; Koeling et al., 2005). For example, on the task of word sense disambiguation, a model trained on a dataset where the predominant sense of the word *star* is "heavenly body", may not work well on text mainly composed of entertainment news. Such problems are also major concerns when developing a system to disambiguate biomedical named entities (e.g., protein,

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 1513–1522,
Singapore, 6-7 August 2009.

gene, and disease), for which some researchers rely on hand-crafted rules in addition to a small amount of training data (Morgan and Hirschman, 2007; Hakenberg et al., 2008).

This paper proposes a new disambiguation method that, instead of classifying each individual occurrence of an entity, it classifies pair-wise relations between the entity mention in question and the "*cue words*" in its adjacent context, where each cue word is assumed to bear a semantic class. We then select the cue word that has a positive relation with the entity, and pass its semantic tag to it. While an individual entity mention may belong to a large number of semantic classes, a relation can only take one of two values: positive or negative, hence transforming a complex multi-classification problem into a less complicated binary classification task. The remainder of the paper is organised as follows: Section 2 proposes the disambiguation method and Section 3 introduces the task of disambiguating the model organisms of biomedical named entities. Section 4 describes in detail our proposed method and also a number of baseline systems for comparison purposes. Section 5 shows the evaluation results and discusses the advantages and drawback of our system, and we finally conclude in Section 6.

2 Disambiguation as Relation Classification

The named entity disambiguation task is defined as follows: given a mention of a named entity in text, we automatically assign a semantic tag d to it, where $d \in D$, and D is a pre-compiled dictionary with $|D|$ entries. When $|D|$ is small, the problem can be approached by supervised classification. For example, to determine whether an occurrence of an entity is a protein, a gene or an RNA, Hatzivassiloglou et al. (2001) compared performance of 3 supervised classification methods and reported results near the human agreement rate. Nevertheless, when $|D|$ is large (e.g., > 100), the performance of classification may decrease, especially when the distribution of d in training dataset differs from that in the test set. In other words, when $|D|$ is large, named entity disambiguation becomes a multi-class classification task on heterogeneous and imbalanced datasets, which is challenging for a machine learning model to learn to discriminate enough between the semantic classes (Japkowicz, 2000).

We propose an alternative method for named entity disambiguation. Intuitively, in the surrounding context of an ambiguous entity, one can often find "cue words" that are informative indicators of the entity's semantic category. These cue words are provided by authors to remind readers the semantic identity of a named entity. For example, in an article about protein p53, phrase "human protein p53" may be mentioned, where both *human* and *protein* contain semantic information regarding p53: *human* indicates the model organism of p53, and *protein* suggests the type of this entity. Such cue words may occur infrequently in the training data, making it difficult for machine learning classifiers to capture.

Our method exploits this observation. Given a sentence, let E be the set of 'target' entities (e.g., *p53*) and W of the 'cue' words (e.g., *human*) that co-occur in a sentence, we define a relation as a pair $r = \langle e, w \rangle$, where $e \in E$ and $w \in W$, and r is a positive relation if e belongs to the semantic class indicated by w, and is a negative one if not. Then we can disambiguate e by accomplishing the following steps: 1) identify W and build a set of relations $R = \{\langle e, w_i \rangle | w_i \in W, i = 1, 2, .., n\}$, where n is the size of W; and 2) classify every $r \in R$ and assign the semantic tag of w_j to e such that $r_j = \langle e, w_j \rangle$ is positive. The first task can be tackled by a dictionary lookup, or by an NER system, if manually annotated data is available. The second is essentially a binary relation classification task, and in this work, we use an SVM model exploiting bag-of-word and syntactic features.

3 Species Disambiguation

We show the performance of the proposed method on a task of resolving one major source of ambiguity in protein and gene entities: 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 *E. coli*, *Drosophila*, *Homo sapiens*, and hundreds more are frequently mentioned in biological research papers. In biomedical articles, entities of different species are commonly referred to us-

ing the same name, causing great ambiguity. For example, searching a protein sequence database, RefSeq[1] with query "*tumor protein p53*" resulted in over 100 proteins, as the name is shared by many organisms.

The importance of distinguishing model organisms has been recognised 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 21 organisms were combined, the ambiguity increased substantially to $21,279$ (14.2%) symbols. Hakenberg et al. (2008) showed that species disambiguation is one of the most important steps for term normalisation and identification, which concerns automatically associating mentions of biomedical entities in text to unique database identifiers (Morgan et al., 2008). Also, the task of extracting PPIs in the recent BioCreative Challenge II workshop (Hirschman et al., 2007) requires protein pairs to be recognised and normalised, which inevitably involves species disambiguation.

More specifically, given a text, in which mentions of biomedical named entities are annotated, a species disambiguation system automatically assigns a *species identifier*, as in a standard database of model organisms, to every entity mention. The types of biomedical named entities concerned in this study are protein, gene, protein complex and mRNA/cDNA, and we used identifiers from the NCBI Taxonomy of model organisms.[2] The work focuses on species disambiguation and assumes that the entities are already identified. In practice, an automated named entity recogniser (e.g., ABNER (Settles, 2005)) should be used before applying the systems.

4 Approaches

This section describes a number of approaches to species disambiguation, highlighting the relation classification method proposed in Section 2.

4.1 Heuristics Baselines

The cue words for species are words denoting names of model organisms (e.g., *mouse* as in phrase "mouse p53"). Another clue is the presence of the species-indicating prefixes in gene and protein names. For instance, prefix 'h' in entity "hSos-1" suggests that it is a *human* protein. Throughout this paper, we refer to such cue words (e.g., *mouse*, *hSos-1*) as "species words". Note that a species "word" may contain multiple tokens (e.g., *E. Coli*).

We encoded this knowledge in a rule-based species tagging system (Wang and Grover, 2008). The system takes a 2-step approach. First, it marks up species words in the document using a species-word detection program,[3] which searches every word in a dictionary of model organisms and assigns a species ID to the word if a match is found. The dictionary was built using the NCBI taxonomy[4] and the UniProt controlled vocabulary of species,[5] and in total it contains 420,224 species words for 324,157 species IDs. When species words are identified, we disambiguate an entity mention using *one of* the following rules:

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 to 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 (see Section 5.1) show that only 5.68% entities can potentially be resolved by rule 1 and 22.16% by rule 2, while the *majority* rule can tackle every entity mention in the dataset.

[1] http://www.ncbi.nlm.nih.gov/RefSeq

[2] http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy

[3] The species word detector identifies the cue words and was used in all the systems studied in this paper. We could not properly evaluate the detector due to the lack of manually annotated data. Its performance, however, would not affect the comparative evaluation results, and improvement to species word detection should increase the performance of these disambiguation systems.

[4] ftp://ftp.ncbi.nih.gov/pub/taxonomy/

[5] http://www.expasy.ch/cgi-bin/speclist

4.2 Supervised Classification

The disambiguation problem can 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. We carried out experiments with two classification methods: multi-class classification and one-class classification, where a maximum entropy model[6] was used for the former and SVM-light[7] for the latter. In one-class classification, we trained a series of binary SVM classifiers, each constructing a separating hyperplane that maximises the margin between the instances of one specific species (i.e., the target class) and a set of randomly selected instances of other species (i.e., the outlier class). We used equal numbers of instances for both classes in training. The following types of features were used in both multi-class and one-class experiments, where the values of n were set empirically by cross-validation on the training data:

- *leftContext* The n word lemmas to the left of the entity ($n = 200$).
- *rightContext* The n word lemmas to the right of the entity ($n = 200$).
- *leftSpeciesIDs* The n species IDs to the left of the entity (with order, $n = 5$).
- *rightSpeciesIDs* The n species IDs to the right of the entity (with order, $n = 5$).
- *leftNouns* The n nouns to the left of the entity (with order, $n = 2$).
- *leftAdjs* The n adjectives to the left of the entity (with order, $n = 2$).
- *leftSpeciesWords* The n species word forms to the left of the entity ($n = 5$).
- *rightSpeciesWords* The n species word forms to the right of the entity ($n = 5$).
- *firstLetter* The first character of the entity itself (e.g., 'h' in *hP53*).
- *documentSpeciesIDs* All species IDs that occur in the document in question.
- *useStopWords* filter out function words.
- *useStopPattern* filter out words consisting only of digits and punctuation characters.

Feature selection was also carried out for the one-class classification experiments. We compared two feature selection methods that reportedly work well on the task of text classification: information gain (IG) (Yang and Pedersen, 1997) and Bi-Normal separation (BNS) (Forman, 2003). IG measures the decrease in entropy when the feature is given vs. absent, and is defined as: $IG(Y|X) = H(Y) - H(Y|X)$ where $H(Y)$ is the uncertainty about the value of Y (i.e., Y's entropy), and $H(Y|X)$ is Y's conditional entropy given X. The BNS is defined as: $|F^{-1}(x) - F^{-1}(y)|$, where F^{-1} is the standard Normal distribution's inverse cumulative probability function, namely, *z-score*; x is the ratio between the number of positive cases containing the feature in question, and the total number of positive cases; and y is the ratio between the number of negative cases containing the feature, and the total number of negative cases.

We computed a weight for each feature and then ranked the features according to their weight, with respect to each feature selection method. The top 10% features were used in training. Given a test instance, the one-class classification method first counts the species words in the document that the instance appears in, and then applies in sequence the binary models of each occurring species, starting from the most frequent one. For example, if a document contains 5 occurrences of *human* and 3 *mouse*, we first apply the *human* species model to judge whether an entity mention is of *human* species, and only if not, the *mouse* model was applied. The most-frequent species in the document was used as backup when none of the binary models gives positive answers.

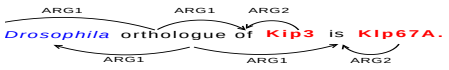

Figure 1: Predicate argument structure (PAS).

4.3 Relation Classification

4.3.1 Overview

As for the proposed relation classification method, in the training phase, we first selected the sentences in which an entity mention and a species word co-occur, and constructed pair-wise entity-species relations. We then assigned each relation a binary label: a relation is positive if the species ID inferred from the species word matches the gold-standard species annotation on the entity, and is negative otherwise. For example, for the sentence shown in Figure 1, where *Drosophila* is a species word, and Kip3 and Klp67A are proteins, relation ⟨Kip3, *Drosophila*⟩ is a negative instance and the

[6] http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

[7] http://svmlight.joachims.org/

pair ⟨Klp67A, *Drosophila*⟩ is a positive one.[8]

For each relation, a vector of features were extracted. We followed the PPI extraction method described in (Miyao et al., 2008), where two types of features were used for a SVM classifier. The first was bag-of-word features, i.e., the words before, between and after the pair of entities, where the words were lemmatised. We added an additional feature of the distance between the entity and the cue word. The other type was syntactic features obtained from parsers. For bag-of-word features, a linear kernel was used, and for syntactic ones, a subset tree kernel (Moschitti, 2006) was adopted. The syntactic features were represented in a flat tree format. Figure 2 shows such a feature for the negative instance ⟨Kip3, *Drosophila*⟩ from Figure 1. Note that all species words (e.g., *Drosophila*) were normalised to "SPECIESWORD", and entities (e.g., Kip3) to "ENTITY", which not only reduces the noise in the feature set, but also makes the model more species-generic. From the training dataset (see Section 5.1), $25,413$ relations were extracted, of which 63.3% were positive.

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

Figure 2: A syntactic feature obtained from the ENJU parser.

To identify the species of an entity in unseen text, we first parsed the sentence, and then listed all pairs of species words and entities as relations. Having extracted the bag-of-word and syntactic features from the instance, the trained model was applied to judge whether each species-entity relation was positive. The entity mention in a positive relation would be tagged with the ID indicated by the species word, while the mentions in negative relations would be left untagged. The next section describes in detail how we extracted the syntactic features from text.

4.3.2 Syntactic Features

Given a sentence, a natural language parser automatically recognises its syntactic structure and outputs a parse tree, in which nodes represent words or syntactic constituents. A path between a pair of nodes can be interpreted as a syntactic relation between sentence units, which was proved useful to infer biological relations (e.g., Airola et al., 2008; Miwa et al., 2008).

Parser	Input	Output
C&C	POS-tagged	GR
ENJU	POS-tagged	PAS
ENJU-Genia	POS-tagged	PAS
Minipar	Sentence-detected	Minipar
RASP	Tokenised	GR
Stanford	POS-tagged	SD
Stanford-Genia	POS-tagged	SD

Table 1: Parsers and their input and output format

We experimented with the following parsers (summarised in Table 1):

- **Dependency parsers** identify one word as the head of a sentence and all other words are either a dependent of that word, or else dependent on some other word that connects to the headword through a sequence of dependencies. We used Minipar (Lin, 1998) and RASP (Briscoe et al., 2006) for the experiments;
- **Constituent-structured parsers** split a sentence into syntactic constituents such as *noun phrases* or *verb phrases*. We used the Stanford parser (Klein and Manning, 2003), and also a variant of the Stanford parser (i.e., Stanford-Genia), which was trained on the GENIA treebank (Tateisi et al., 2005) for biomedical text;
- **Deep parsers** aim to compute in-depth syntactic and semantic structures based on syntactic theories such as HPSG (Pollard and Sag, 1994) and CCG (Steedman, 2000). We used the C&C parser (Clark and Curran, 2007), ENJU (Miyao and Tsujii, 2008), and a variant of ENJU (Hara et al., 2007) adapted for the biomedical domain (i.e., ENJU-Genia);

There were a number of practical issues to consider when using parsers for this task. Firstly, before parsing, the text needs to be linguistically pre-processed, and the quality of this process has a significant impact on parsers' performance. The pre-processing steps include sentence boundary detection, tokenisation and part-of-speech (POS) tagging, all of which can be tricky especially when applied to biomedical text (Grover et al., 2003). To avoid the noise that can be introduced in the pre-processing steps and to concentrate on evaluating the performance of the parsers, we used the same pre-processing tools (Alex et al., 2008a)[9] whenever possible. The middle column in Table 1 shows how the input text was linguistically pre-processed with respect to each parser. A POS-tagged text implies that it was also sentence boundary detected and tokenised Except for

[8] *Orthologues* are genes/proteins in different species but have similar sequences. In this example it implies that Klp67A is a *Drosophila* protein but Kip3 is not.

[9] These particular tools were chosen because they were adopted to pre-process the ITI-TXM dataset, which we used in our study.

RASP and Minipar, all parsers took POS-tagged text as input. RASP requires POS tags and punctuation labels that were derived from the CLAWS-7 tagset,[10] whereas our dataset uses POS labels from the Penn Treebank tagset (Marcus et al., 1994). As RASP does not recognise the Penn tagset, we used its build-in POS tagger. Minipar, on the other hand, does not support input of tokenised or POS-tagged text, and therefore took split sentences as input.

Secondly, the output representations of the parsers are different and we preferred a format that depicts relations between words instead of syntactic constituents. In total, 4 representations were 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). All the above representations define relations of words in triples, 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. Figure 1 shows a sentence parsed by ENJU in PAS representation. The right-most column in Table 1 lists the output representation of each parser. A syntactic path between an entity and a species word was represented by a sequence of triples, each following the order of head-dependent or predicate-argument. These paths were used as syntactic features for the SVM classifier.

4.4 Spreading Strategies

Except for the *majority vote* rule, the approaches described in Sections 4.1 and 4.3 were expected to yield low recall, because they can only detect intra-sentential relations, and therefore only be applied to the entities having at least one species word appearing in the same sentence.

Since our aim is to disambiguate as many entity mentions as possible, we would like to "spread" the decisions from the disambiguated mentions to their "relatives" in the same document. We define an entity mention $\bar{e}$ as another mention e's relative under either of the following conditions: a) if $\bar{e}$ has the same surface form with e; or, b) if $\bar{e}$ is an abbreviation or an antecedent of e, where abbreviation/antecedent pairs were 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 used a "default" species, as determined by the rule of *majority vote* (see Section 4.1).

5 Evaluation

5.1 Data and Ontology

The species disambiguation experiments were conducted using the ITI-TXM corpus (Alex et al., 2008b), a collection of *full-length* biomedical research articles manually annotated with linguistic and biomedical information for developing automatic information extraction systems. The corpus contains two datasets covering slightly different domains: enriched protein-protein interaction (EPPI) and tissue expression (TE). Whenever possible, protein, protein complex, gene, and mRNA/cDNA entities were tagged with NCBI Taxonomy IDs, denoting their species, and it was the species annotation that this study used.

The EPPI and TE datasets have different distributions of species. The entities in EPPI belong to 118 species with *human* being the most frequent at 51.98%. In TE, the entities are across 67 species and *mouse* is the most frequent at 44.67%.[11] The inter-annotator agreement of species annotation on EPPI and TE are 86.45% and 95.11%, respectively.

The species disambiguation systems were developed on the training portions of the EPPI and TE corpora, each containing 221 articles, and evaluated on a dataset combining the development test (DEVTEST) datasets of EPPI and TE, containing 58 and 48 articles, respectively. The combined training dataset contains $96,992$ entity mentions belonging to 138 model organisms, while the DEVTEST dataset contains $23,118$ entities of 54 species. The diversity of model organisms in this corpus highlights the fact that a primary consideration when developing a species disambiguation system is its ability to distinguish a wide range of species with minimal additional manual effort.

5.2 Results

5.2.1 Evaluation Metrics

The evaluation was carried out on the DEVTEST dataset, and the systems are compared using av-

[10] http://ucrel.lancs.ac.uk/claws7tags.html

[11] These figures were obtained from the training split of the datasets.

	micro-avg.	macro-avg.
Maxent	70.48 / 70.48 / 70.48	10.07 / 10.00 / 9.85
SVM	62.24 / 59.35 / 60.76	14.70 / 17.11 / 15.01
SVM (IG)	65.20 / 61.06 / 63.06	14.90 / 19.53 / 16.09
SVM (BNS)	43.61 / 42.63 / 43.11	11.99 / 10.05 / 9.34

Table 2: Evaluation results of the classification systems on DEVTEST (precision/recall/F1-score, in %)

eraged precision, recall and F1 scores over all species. In more detail, for each model organism that appears in the DEVTEST dataset, we collect two lists of entity mentions of that species: one from the gold-standard DEVTEST dataset, and the other from the output of a disambiguation system. Then the list of system output is compared against the gold-standard list to obtain precision, recall and F1 score. For each system, the scores obtained from all species are averaged using micro-average and macro-average. The 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. The macro-average is the mean of precision, recall, or F1 over all labels, thus attributing equal weights to each species, and measuring a system's adaptability across different model organisms.

5.2.2 Evaluation Results

First of all, Table 2 shows the results of the classification methods described in Section 4.2. The multi-classification system using a maximum entropy model (Maxent) yielded the highest overall micro-averaged F1. Among the SVM-based systems, the one using IG feature selection achieved better performance. In particular, it outperformed the Maxent model in term of macro-averages. The performance of the SVM model with BNS feature selection is disappointing, perhaps because the occurrences of a feature in each instance are not normally distributed. As the Maxent system obtained better results, it was used to compare with other disambiguation systems.

Table 3 shows the results of a number of methods described in the previous sections. The methods are categorised into 4 groups: rule-based baseline systems, a Maxent classification model, relation-classification methods, and a hybrid system. The difference between the relation classification systems is the features adopted. Rel-Context was trained on only bag-of-word and distance features, whereas each other system also used syntactic features provided by a specific parser. For example, the Rel-RASP system identifies an entity's species by finding positive relations between the entity and its neighbouring species words, using features including bag-of-word, distance, and dependency paths generated by RASP. The hybrid system (Hbrd) ran the Rel-ENJU-Genia system on top of the outcome of Maxent. When a conflict occurs, the species ID is chosen by Rel-ENJU-Genia. The idea is that the relation classification system is more accurate than Maxent when it is applicable, and hence would improve precision on disambiguating the species with few or no training instances.

Without spreading (shown in the "NO SPRD" columns of Table 3), most of the rule-based and relation classification systems only work on a subset of DEVTEST, resulting in low recall: Rule-Sp works on the small proportion of entities (5.68%) with a preceding species word, while the other systems only work on the collection of sentences containing at least one species word and one entity, which covers 4.60% sentences and 22.16% entity mentions. Rule-Majority, Maxent, and Hbrd, on the other hand, apply to all entity mentions, and therefore they are only compared against the others when spreading was applied.

The results shown in the "NO SPRD" columns can be viewed as a comparative evaluation of the usefulness of the syntactic features supplied by the parsers on this particular task. The rule-based systems set high baselines: Rule-Sp produced good precision and Rule-SpSent achieved the highest micro-averaged F1, thanks to its high coverage, which is also an upperbound of recall for the relation classification systems. Nevertheless, it is encouraging that the relation classification systems obtained higher precision than Rule-SpSent, which is important, considering the decisions will be transfered to the untagged entity mentions across the document. Indeed, as shown in the SPRD columns in Table 3, most relation classification systems outperformed the Rule-SpSent baseline when spreading was used. The scores of the systems using different parser outputs only vary slightly. Rel-Context, on the other hand, surpassed others in terms of micro-averaged precision, while sacrificing micro-averaged recall and macro-averaged scores.

Next, the SPRD columns in Table 3 show the results when the spreading rules were applied, which

METHOD	NO SPRD (micro-avg)	NO SPRD (macro-avg)	SPRD (micro-avg)	SPRD (macro-avg)
Rule-Majority	N/A	N/A	66.14 / 61.99 / 64.00	16.76 / 21.75 / 18.08
Rule-Sp	88.96 / 5.02 / 9.51	33.77 / 8.55 / 10.18	66.96 / 63.41 / 65.13	28.25 / 30.65 / 27.00
Rule-SpSent	80.82 / 16.88 / 27.93	43.16 / 28.85 / 24.73	67.34 / 63.22 / 65.21	22.65 / 26.42 / 23.10
Maxent	N/A	N/A	70.48 / 70.48 / 70.48	10.07 / 10.00 / 9.85
Rel-Context	90.04 / 3.71 / 6.13	15.23 / 4.45 / 4.90	67.34 / 63.22 / 65.21	22.65 / 26.42 / 23.10
Rel-C&C	82.79 / 16.14 / 27.02	43.97 / 29.56 / 25.60	66.59 / 63.64 / 65.08	32.29 / 33.20 / 29.14
Rel-ENJU	83.39 / 15.87 / 26.66	46.89 / 29.88 / 25.95	68.28 / 65.02 / 66.61	31.82 / 34.08 / 29.67
Rel-ENJU-Genia	83.54 / 15.74 / 26.49	44.13 / 29.93 / 25.78	68.91 / 65.45 / 67.13	32.00 / 34.87 / 30.21
Rel-Minipar	81.82 / 16.27 / 27.14	43.63 / 27.88 / 24.15	67.98 / 63.77 / 65.81	31.83 / 33.93 / 29.44
Rel-RASP	81.67 / 16.10 / 26.90	43.95 / 28.92 / 25.03	66.62 / 64.08 / 65.33	32.66 / 33.54 / 29.80
Rel-Stanford	82.75 / 16.10 / 26.95	44.05 / 29.49 / 25.92	66.81 / 63.81 / 65.28	32.67 / 33.03 / 29.45
Rel-Stanford-Genia	82.22 / 16.04 / 26.84	43.37 / 29.40 / 25.22	66.85 / 63.64 / 65.21	32.72 / 32.29 / 28.64
Hbrd	N/A	N/A	74.15 / 73.26 / 73.70	43.98 / 37.47 / 31.80

Table 3: Evaluation results of the species disambiguation systems on DEVTEST (precision/recall/F1-score, in %)

effectively improved recall (see Section 5.2.3 for discussion on statistical significance tests on the results). The Maxent system achieved very good micro-averaged precision, but low macro-averaged scores. In fact, as shown in Table 4, Maxent can only disambiguate 7 species (out of a total of 54) that have relatively large amount of training instances,[12] and failed completely on other species. This suggests that Maxent may not be able to generate good micro-averaged scores when applied to a dataset where the dominant species are different from those in the training set. On the other hand, the relation-classification approaches have a clear advantage over Maxent as measured by macro-averaged scores. As shown in Table 4, Rel-ENJU-Genia worked well on most of the species, displaying its good adaptability, while achieving comparable micro-averaged F1 to Maxent. Overall, Hbrd, which combines the strengths of relation classification and the Maxent classification model, obtained the highest points as measured by every metric.

Species Name	Pct	Mxt	Rel	Hbrd
H. sapiens	50.13%	76.25	65.33	79.51
M. musculus	13.99%	66.41	58.29	68.27
X. tropicalis	7.35%	64.80	77.72	71.39
D. melanogaster	6.34%	93.17	78.46	95.15
S. cerevisiae	4.79%	90.12	83.32	87.68
R. norvegicus	2.97%	44.04	38.69	51.77
T. aestivum	2.62%	0.00	89.68	23.35
P. americana	2.27%	0.00	98.50	7.76
C. elegans	2.08%	96.83	95.88	97.50
H. herpesvirus 5	1.58%	0.00	54.46	4.27
R. virus	1.45%	0.00	28.54	6.45
H. spumaretrovirus	1.17%	0.00	99.37	2.49
...	...	...	...	...
Macro-average		9.85	30.21	31.80
Micro-average		70.48	67.13	73.70

Table 4: The micro-averaged F1 scores (%) of Maxent (Mxt), Rel-ENJU-Genia with spreading (Rel), and Hbrd with respect to each of the most frequent 12 species in DEVTEST.

5.2.3 Statistical Significance

To see whether our methods significantly improved the baseline systems, we performed randomisation tests (Noreen, 1989; Yeh, 2000) on some of the results shown in Table 3. The intuition of randomisation test is as follows: when comparing two systems (e.g., A and B), we erase the labels "output of A" or "output of B" from all observations. The null hypothesis is that there is no difference between A and B, and thus any response produced by one of the systems could have as likely come from the other. We shuffle these responses R times, reassign each response to A or B and see how likely such a shuffle produces a difference in the metric of interest that is at least as large as the difference observed when using A and B on the test data. Let r denote the number of times that such a difference occurred, then as $R \rightarrow \infty$, $\frac{r+1}{R+1}$ approaches the significance level. In our case, the metrics tested were micro- and macro-averaged precision, recall and F1.

Following this procedure, we tested whether the improvements made by a relation classification based system (i.e., Rel-ENJU-Genia with SPRD) and the hybrid system (i.e., Hbrd) over the baseline systems were statistically significant. We carried out approximate randomisation with 10,000 shuffles and the test results are shown in Table 5. The numerical figures in the cells are differences in precision, recall and F1 between a pair of systems. The significance levels (i.e., p-values) are indicated by superscript marks, whose corresponding values are displayed in Table 6. For exam-

[12] The following 7 species occur most frequently in the training set: H. sapiens (43.25%), M. musculus (27.05%), R. norvegicus (5.35%), S. cerevisiae (3.98%), X. tropicalis (3.56%), D. melanogaster (3.33%) and C. elegans (0.94%).

		Rule-Majority	Rule-Sp	Rule-SpSent	Maxent
Rel	micro-avg	2.77*/3.46*/3.13*	1.95*/2.04*/2.00*	1.57*/2.22*/1.92*	-1.57* / -5.02* / -3.35*
	macro-avg	15.24*/13.12*/12.13*	3.75[a]/4.21[a]/3.20[a]	9.35*/8.44*/7.10*	21.92*/24.87*/20.35*
Hbrd	micro-avg	8.01*/11.27*/9.70*	7.19*/9.85*/8.57*	6.81*/10.04* /8.49*	3.67*/2.78*/2.82[b]
	macro-avg	27.22*/15.72[c]/13.72[d]	15.73*/6.82[e]/4.80[f]	21.33*/11.05[g] / 8.70[h]	33.91[i]/27.47*/21.95*

Table 5: Results of paired randomisation tests on whether Rel-ENJU-Genia with SPRD (Rel) and Hbrd significantly improved the baseline systems. The numerical figures in the cells show the differences between the two systems as measured by precision/recall/F1 in percentage. The superscript marks indicate the significance levels and are explained in Table 6.

ple, the difference in micro-averaged precision between Rel-ENJU-Genia and Rule-Majority on the test data was 2.77%, and in 10,000 approximate randomisation trials, there was zero times[13] that Rel-ENJU-Genia's micro-averaged precision is greater than Rule-Majority's by at least 2.77% ($p < 0.0001$).

MARK	VALUE	MARK	VALUE
*	$p < 0.0001$	a	$p < 0.06$
b	$p < 0.002$	c	$p < 0.0003$
d	$p < 0.0002$	e	$p < 0.03$
f	$p < 0.05$	g	$p < 0.003$
h	$p < 0.005$	i	$p < 0.07$

Table 6: p-values.

The test results confirmed that, the improvements made by Hbrd are statistically significant with at least 95% confidence as measured by all metrics except for macro-averaged precision. The relation classification approach achieved significantly lower performance than Maxent in terms of micro-averaged scores (hence the "-" sign in the corresponding cell in Table 5), but in all other cases it can reject the null hypothesis with very high confidence (i.e., $p < 0.0001$).

6 Conclusions and Future Work

This paper proposes a method that tackles a complex disambiguation problem by breaking it into two cascaded simpler tasks of cue word discovery and binary relation classification. We evaluated the method on the task of disambiguating the model organisms of biomedical named entities, along with a number of other approaches. As measured by micro-averaged F1 score, a supervised classification approach (Maxent) yielded the second best result. However, it can only disambiguate a small number of species that have abundant training instances. With spreading rules, a relation classification system (Rel-ENJU-Genia) trained on word and syntactic features from ENJU-Genia also obtained good micro-averaged F1, while surpassing Maxent significantly in terms of macro-averaged scores. Combining these two systems achieved the best overall performance. Nevertheless, we combined the two methods in a rather crude way, leaving ample room for exploring better strategies in the future.

One drawback of the relation classification systems is that they can not cover all entity mentions but only the ones with informative keywords co-occurring in the same sentence. We overcame the drawback by using spreading rules. For some applications, however, it may be sufficient to make predictions exclusively for cases where the systems are applicable. Also, the predictions with high confidence can be used as seed training material for automatically harvesting more training data.

[13]The numbers of times are not shown in Table5 for brevity.

Acknowledgments

The work reported in this paper is funded by Pfizer Ltd.. The UK National Centre for Text Mining is funded by JISC. The ITI-TXM corpus used in the experiments was developed at School of Informatics, University of Edinburgh, in the TXM project, which was funded by ITI Life Sciences, Scotland.

References

E. Agirre and D. Martinez. 2004. Unsupervised WSD based on automatically retrieved examples: The importance of bias. In *Proceedings of EMNLP*.

A. Airola, S. Pyysalo, J. Björne, T. Pahikkala, F. Ginter, and T. Salakoski. 2008. A graph kernel for protein-protein interaction extraction. In *Proceedings of BioNLP*.

B. Alex, C. Grover, B. Haddow, M. Kabadjov, E. Klein, M. Matthews, S. Roebuck, R. Tobin, and X. Wang. 2008a. Assisted curation: does text mining really help? In *Proceedings of the Pacific Symposium on Biocomputing*.

B. Alex, C. Grover, B. Haddow, M. Kabadjov, E. Klein, M. Matthews, S. Roebuck, R. Tobin, and X. Wang. 2008b. The ITI TXM corpus: Tissue expression and protein-protein interactions. In *Proceedings of the Workshop on Building and Evaluating Resources for Biomedical Text Mining at LREC*.

E. Briscoe, J. Carroll, and R. Watson. 2006. The second release of the RASP system. In *Proceedings of the COLING/ACL Interactive Presentation Sessions*.

R. Bunescu and M. Paşca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of EACL*.

L. Chen, H. Liu, and C. Friedman. 2005. Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*, 21(2):248–256.

S. Clark and J. R. Curran. 2007. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4).

M-C de Marneffe, B. MacCartney, and C. D. Manning. 2006. Generating typed dependency parses from phrase structure. In *Proceedings of LREC*.

G. Erkan, A. Ozgur, and D. R. Radev. 2007. Semi-supervised classification for extracting protein interaction sentences using dependency parsing. In *Proceedings of the Joint Conference of EMNLP and CoNLL*.

G. Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305.

C. Grover, M. Lapata, and A. Ascarides. 2003. A comparison of parsing technologies for the biomedical domain. *Natural Language Engineering*, 1(1):1–38.

J. Hakenberg, C. Plake, R. Leaman, M. Schroeder, and G. Gonzalez. 2008. Inter-species normalization of gene mentions with GNAT. *Bioinformatics*, 24(16).

T. Hara, Y. Miyao, and J. Tsujii. 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*.

V. Hatzivassiloglou, PA Duboué, and A. Rzhetsky. 2001. Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Bioinformatics*, 17(Suppl 1).

L. Hirschman, M. Krallinger, J. Wilbur, and A. Valencia, editors. 2007. *The BioCreative II - Critical Assessment for Information Extraction in Biology Challenge*, volume 9(Suppl 2). Genome Biology.

L. Hunter and K. B. Cohen. 2006. Biomedical language processing: what's beyond PubMed. *Molecular Cell*, 21(5):589–594.

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

D. Klein and C. D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of ACL*.

R. Koeling, D. McCarthy, and J. Carroll. 2005. Domain-specific sense distributions and predominant sense acquisition. In *Proceedings of HLT/EMNLP*.

D. Lin. 1998. Dependency-based evaluation of MINIPAR. In *Proceedings of Workshop on the Evaluation of Parsing Systems*.

M. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1994. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

M. Miwa, R. Satre, Y. Miyao, T. Ohta, and J. Tsujii. 2008. Combining multiple layers of syntactic information for protein-protein interaction extraction. In *Proceedings of SMBM*.

Y. Miyao and J. Tsujii. 2008. Feature forest models for probabilistic HPSG parsing. *Computational Linguistics*, 34(1).

Y. Miyao, R. Sætre, K. Sagae, T. Matsuzaki, and J. Tsujii. 2008. Task-oriented evaluation of syntactic parsers and their representations. In *Proceedings of ACL-08: HLT*.

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

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. Liu, R. Torres, M. Krauthammer, W. W. Lau, H. Liu, C. Hsu, M. Schuemie, K. B. Cohen, and L. Hirschman. 2008. Overview of BioCreAtIvE II gene normalization. *Genome Biology*, 9(Suppl 2).

A. Moschitti. 2006. Making tree kernels practical for natural language learning. In *Proceedings of EACL*.

E. W. Noreen. 1989. *Computer Intensive Methods for Testing Hypothesis*. John Wiley & Sons.

C. Pollard and I. A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago.

A. S. Schwartz and M. A. Hearst. 2003. Identifying abbreviation definitions in biomedical text. In *Proceedings of the Pacific Symposium on Biocomputing*.

B. Settles. 2005. ABNER: An open source tool for automatically tagging genes, proteins, and other entity names in text. *Bioinformatics*, 21(14):3191–3192.

M. Steedman. 2000. *The Syntactic Process*. The MIT Press, Cambridge, MA.

Y. Tateisi, A. Yakushiji, T. Ohta, and J. Tsujii. 2005. Syntax annotation for the GENIA corpus. In *Proceedings of IJCNLP*.

X. Wang and C. Grover. 2008. Learning the species of biomedical named entities from annotated corpora. In *Proceedings of LREC*.

Y. Yang and J. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of ICML*.

A. Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of COLING*.