Learning to Recognize Phenotype Candidates in the Auto-Immune Literature Using SVM Re-Ranking

Nigel Collier[1,2]*, Mai-vu Tran[2,3], Hoang-quynh Le[2,3], Quang-Thuy Ha[3], Anika Oellrich[4], Dietrich Rebholz-Schuhmann[1,5]

1 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge, United Kingdom, **2** National Institute of Informatics, Tokyo, Japan, **3** Knowledge Technology Laboratory, University of Engineering and Technology - VNU, Hanoi, Vietnam, **4** Mouse Informatics Group, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, **5** Department of Computational Linguistics, University of Zurich, Zurich, Switzerland

Abstract

The identification of phenotype descriptions in the scientific literature, case reports and patient records is a rewarding task for bio-medical text mining. Any progress will support knowledge discovery and linkage to other resources. However because of their wide variation a number of challenges still remain in terms of their identification and semantic normalisation before they can be fully exploited for research purposes. This paper presents novel techniques for identifying potential complex phenotype mentions by exploiting a hybrid model based on machine learning, rules and dictionary matching. A systematic study is made of how to combine sequence labels from these modules as well as the merits of various ontological resources. We evaluated our approach on a subset of Medline abstracts cited by the Online Mendelian Inheritance of Man database related to auto-immune diseases. Using partial matching the best micro-averaged F-score for phenotypes and five other entity classes was 79.9%. A best performance of 75.3% was achieved for phenotype candidates using all semantics resources. We observed the advantage of using SVM-based learn-to-rank for sequence label combination over maximum entropy and a priority list approach. The results indicate that the identification of simple entity types such as chemicals and genes are robustly supported by single semantic resources, whereas phenotypes require combinations. Altogether we conclude that our approach coped well with the compositional structure of phenotypes in the auto-immune domain.

Citation: Collier N, Tran M-v, Le H-q, Ha Q-T, Oellrich A, et al. (2013) Learning to Recognize Phenotype Candidates in the Auto-Immune Literature Using SVM Re-Ranking. PLoS ONE 8(10): e72965. doi:10.1371/journal.pone.0072965

Editor: Luis M. Rocha, Indiana University, United States of America

Received February 11, 2013; **Accepted** July 15, 2013; **Published** October 14, 2013

Funding: NC is supported by an EC Marie Curie International Incoming Fellowship grant (http://ec.europa.eu/research/mariecurieactions/index_en.htm) for project Phenominer (project number 301806). MT and HL are supported by a National Institute of Informatics internship grant (www.nii.ac.jp). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: collier@ebi.ac.uk

Introduction

Since the discovery of the relationship between genotype, environment and phenotype, phenotype data has been used to investigate disease–gene relations [1,2], drug repurposing [3] and in evolutionary studies [4]. A diverse landscape of resources has evolved harboring genotype–phenotype associations such as the Mouse Genome Informatics database (MGD) [5] and the Online Mendelian Inheritance of Man (OMIM) database [6]. This landscape, shown in Figure 1, ranges from narrative descriptions to ontological concepts. Only once we are able to integrate these co-existing data reprentations will be able to fully understand the biological content encoded by each.

While the integration of phenotype data on an ontological level has been demonstrated to enable the prediction of novel gene–disease associations or drug–disease associations [3], the integration of textual data, such as scientific literature, still lags behind. To achieve semantic integration on an ontological level, there was a shift from pre-composed, species-specific phenotype ontologies (e.g. Mammalian Phenotype Ontology (MP) [7]) to a post-composition of phenotype data using species-agnostic ontologies (e.g. Gene Ontology (GO) [8] and PATO [9]). A post-composed phenotype representation requires an entity that is further described based on a quality, e.g. *brown fur colour* or *decreased body weight*. Phenotype data extracted from textual content would have to facilitate both, the normalisation to pre-composed phenotype representations as well as the post-composition of a phenotype.

Furthermore, the data contained within model organism database is obtained through curation of the scientific literature. A need to support database curation work has been identified [10] and current solutions have been found to be insufficient to support the curation workflow [11]. While multiple studies have examined the automatic annotation of genes, proteins and diseases in scientific texts, there is a significant gap in our understanding of how to identify and normalise phenotype mentions. This is partially due to the complexity of the phenotype descriptions, but can also be attributed to incompleteness of phenotype data [12] and a consequent lack of comprehensive semantic resources covering their full scope. Any progress in the automatic identification of phenotypes in the scientific literature would drive the scientific progress in the above mentioned research fields.

This paper presents novel techniques for identifying potential complex phenotype mentions by exploiting a hybrid model based

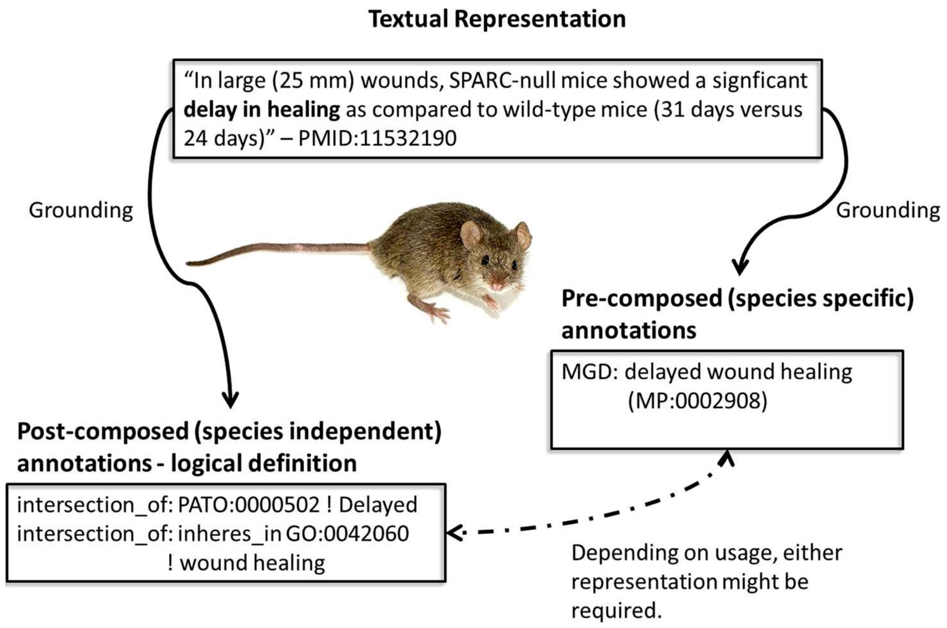

Textual Representation
"In large (25 mm) wounds, SPARC-null mice showed a signficant **delay in healing** as compared to wild-type mice (31 days versus 24 days)" – PMID:11532190
Grounding
Grounding
Pre-composed (species specific) annotations
MGD: delayed wound healing (MP:0002908)
Post-composed (species independent) annotations - logical definition
intersection_of: PATO:0000502 ! Delayed
intersection_of: inheres_in GO:0042060 ! wound healing
Depending on usage, either representation might be required.

Figure 1. Representation of phenotypes in textual narratives and as pre-composed and post-composed terms. Imagine *Mus musculus* courtesy of George Shuklin published at Wikimedia Commons.
doi:10.1371/journal.pone.0072965.g001

on machine learning, rules and dictionary matching. A systematic study is made of how to combine sequence labels from these modules as well as the merits of various ontological resources such as the Human Phenotype Ontology (HPO) [13], the Foundation Model of Anatomy (FMA) [14] and PATO. We evaluated our approach on a subset of Medline abstracts cited by OMIM for auto-immune diseases. After a review of related research we start start by outlining a conceptual analysis of phenotypes.

Background

The task of identifying and classifying phenotype mentions in text requires an understanding of the complex nature of their semantics and syntactic structure. In contrast to *simple* entities such as tissues and organs which have a clear structural and spatial basis, the definition and rightful delineation of phenotypes appears puzzling even to researchers and clinicians. This is partly due to phenotypes cutting across both physical objects and processes but also across levels of granularity from the molecular level to the organism. The class of phenotypes is also viewed differently in the clinical and biological data contributing for example to more frequent disease terms in the HPO than in the MP. Phenotypes may be defined experimentally or clinically according to a model reference so that concepts include a notion of *difference to reference model*, leading to a notion of *abnormality* [15]. In the approach that we are taking, we argue that it is vital to annotate surface mentions of phenotypes in a machine readable form that can then be linked to pre-composed phenotype ontologies, and, at the same time, makes explicit their internal dependencies and links their substructures to species-agnostic ontologies to support logical reasoning and hypothesis exploration through post-composition.

The automated recognition of biomedical terms in text has been a highly active area for over two decades and is referred to variously as terminology extraction', term recognition', entity extraction' and named entity recognition' (NER). Most previous NER research has focused on single rather than joint semantic classes such as genes, proteins, cells, anatomical entities and organisms in the experimental biology domain, e.g. [16], and medication, dosage and symptom in the clinical domain, e.g. [17]. Common approaches include supervised machine learning [18–21], active learning [22], semi-supervised learning [23], dictionary-based approaches [24,25], rule-based approaches [26] and hybrid approaches [27–30]. Open-source tools for NER include BANNER [21], ABNER [20], LINGPIPE [31], the GENIA tagger [32] and NERSuite, a named entity recognition toolkit based on CRFSuite [33]. Recent community evaluations of state-of-the-art tools for common entity types reported in the BioCreative II [34] and CALBC [35] challenges show quite widely varying F-score performance (see Matching metrics) when trained and tested on the same corpus with the highest scoring approaches generally achieving performance for entity detection and classification of about 80% for genes/gene products, chemicals and diseases and about 90% for organisms. For anatomical entities a granular approach based on 11 levels such as cell, organ and tissue achieved performance of about 71% F-score [36]. In a recent evaluation [37], performance for state-of-the-art NER taggers such as Banner [21], Abner [20] and Lingpipe [31] have been found to offer between 41% and 61% for genes when trained and tested on different corpora. The evaluation in this study was carried out using the partially overlapping annotation method; training was done on standardly available corpora of abstracts such as BioCreative II, JNLPBA [38], GENIA [39] and GeneTag [40] and testing on a newly released full text corpus called CRAFT. We refer readers to the overviews for BioCreative II and CALBC for further background information.

Compared to other entity classes there are very few studies that focus on capturing phenotypes [30,41–43]. Chen and Friedman [41] adapted a rule-based system called BioMedLEE by writing specialised grammatical rules and importing vocabulary from the Unified Medical Language System (UMLS) and the Mammalian Ontology [5]. In a recent study, Khordad *et al.* [42] applied a staged rule-based system on the UMLS, HPO and MetaMap. In our earlier study [30] we provided a comparison of Conditional

Random Fields (CRFs), Hidden Markov Models (HMMs) and a hybrid approach against Khordad's method in the domain of human auto-immune diseases. On a two class corpus, performance for phenotypes was 77% F-score for the hybrid system, 65% for the next best performing model CRF, 61% for Khordad's approach and 36% for the HMM. The results indicated the importance of applying a range of resources that can capture phenotypes in experimental papers. [43], Groza *et al.* [44] took a different approach by trying to explicitly model the internal term structure according to qualities and the anatomical entities to which they apply. This is aimed at reducing problems associated with disjoint mentions such as *irregular flared metaphyses… with streaky sclerosis* by normalising to *irregular flared streaky sclerosis metaphyses*. They tested their technique on a corpus of HPO terms under *Abnormality of the skeletal system* (HP:000924).

From these studies we consider the following conclusions to be important: (a) Intuitions about phenotypes are highly variable among experts and therefore good annotation guidelines are necessary for consistency [41], (b) Rule based approaches bootstrapped with ontologies and tools such as the UMLS, HPO and MetaMap are all valuable [41,42] but their combination with corpus-based approaches can lead to improvements [30], (c) Performance is considered to vary depending on whether phenotypes include both objects and processes [30,41], (d) Surface term variation remains a key issue [43].

In our approach, rather than solve the problem of identifying free-text phenotypes in one stage, we have divided the task into two stages. (Stage 1) is the identification of candidate terms and, (Stage 2) is candidate confirmation by compositional analysis through grounding to ontologies such as PATO and the FMA, used for the post-composition of phenotype data. The study we report here contributes to both stages of our task, even though Stage 2 is not finished yet. With the work presented in this manuscript, we highlight future directions to be taken in order to enable the identification of the internal structure of phenotypes and their relation to species-agnostic ontologies.

Our previously reported study [30] showed the benefits of a hybrid approach to phenotype candidate recognition. This model combined a state-of-the-art sequence labelling model (Conditional Random Fields) trained on lexical features, with a rule-based MetaMap module and dictionary matching. The target classes were phenotypes and gene/gene products. Hypothesis resolution used a small set of heuristic rules. However, it seemed unlikely that we had reached optimal performance since the domain resources employed and the method we used to combine alternative sequence labeling hypotheses were limited in scope. The study we present here seeks to extend this in a number of important ways:

- We explore additional semantic resources including 320,000 chemical terms from the Joint Chemical Dictionary (Jochem), 9,000,000 gene terms from the National Library of Medicine gene list, 120,000 human anatomy terms from the FMA, 275,000 terms from the UMLS related to diseases and abnormalities, 9,900 phenotype terms from the HPO with 15,800 synonyms, 8,800 phenotype terms from the Mammalian Phenotype Ontology (MP) with 23,700 synonyms, 1400 quality terms from PATO with 2,200 synonyms, species terms from the Linnaeus tool [45] and 5,400 anatomy terms from the Brenda Tissue Ontology [46] with 9,600 synonyms. This is exemplified in Figure 2.
- We evaluate several alternative approaches for hypothesis selection in the merge module by comparing our previous priority list approach to a Maximum Entropy model with beam search (ME+BS) and a Support Vector Machine with learn to rank (SVM+LTR). The full experimental system is illustrated in Figure 3 highlighting the modules where we make our contribution.
- We incorporated four new entity types in our evaluation.

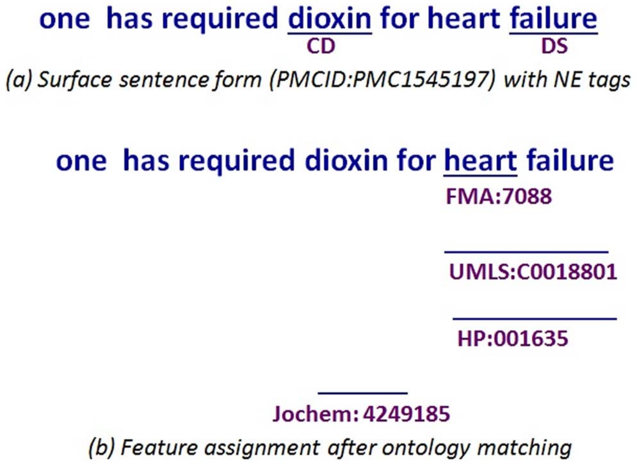

one has required dioxin for heart failure
CD
DS
(a) Surface sentence form (PMCID:PMC1545197) with NE tags
one has required dioxin for heart failure
FMA:7088
UMLS:C0018801
HP:001635
Jochem: 4249185
(b) Feature assignment after ontology matching

Figure 2. Example tagging of phenotypes along with features from external vocabularies and ontologies.
doi:10.1371/journal.pone.0072965.g002

We base our results on the previous study's 122 abstract corpus in order to show a comparison against our earlier methods using phenotype entities.

Materials and Methods

Concept analysis

Given the complexity of phenotypes, one important factor we see for achieving automated annotation accuracy is to avoid conceptual inconsistencies in the coding scheme. In this respect principles from formal ontology might be beneficial [47] such as rigorous definition of markable classes as well as semantic linkage to extant standards within the biomedical community. The *de facto* quality assurance standard in NER has been to empirically validate annotation schemes through Cohen's kappa coefficient (κ) score (e.g. see [48] for a broad discussion). Properly applied this can provide valuable evidence about expert intuition. However if the corpus is not balanced across entity classes then any inferences drawn from agreement on the whole coding scheme becomes weakened. Since it is in practice often difficult to create balanced corpora for NER, if κ is applied in this way, any changes to systems that improve agreement with the unbalanced corpus may actually move models further away from part of their actual goal which should be to maximise agreement across all classes. Whilst we do not neglect the fact that κ is an important tool for schema development, we also note that empirical studies have pointed to the benefits of formal conceptual analysis techniques such as OntoClean [49]. This is based on an understanding that a failure to clearly define the entities is at least partly responsible for inconsistencies in annotating mentions leading to modeling error.

Here we base our named entities on a formal analysis of biological concepts related to disease by Scheuermann et al. [50] and Beisswanger's BioTop [47]. The entity types we annotate are given in abbreviated upper case form, i.e. GG, CD, AN, PH, DS and OR which we now define.

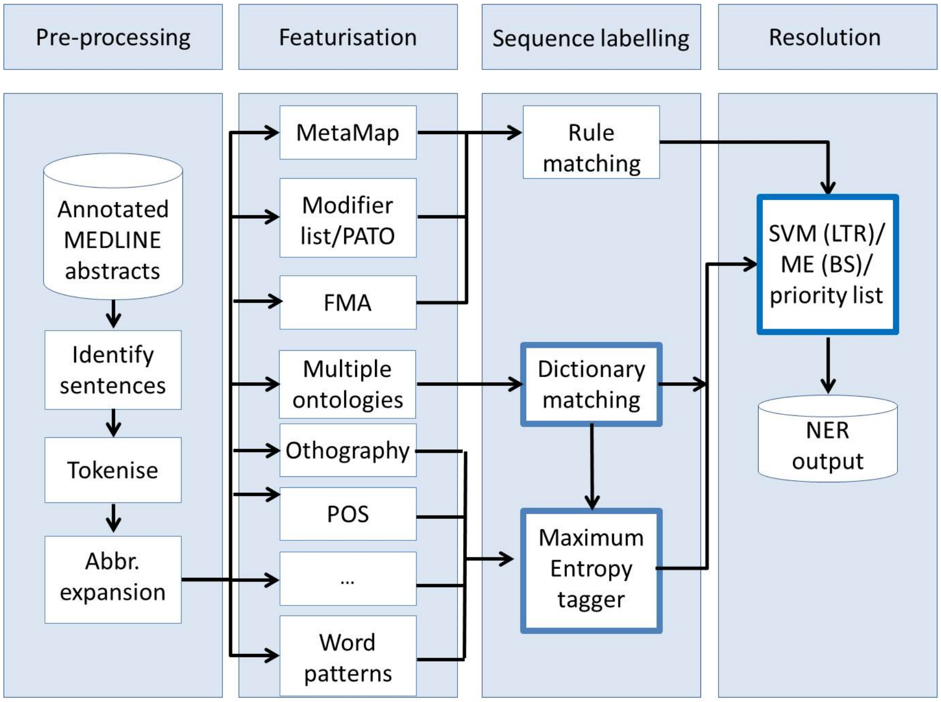


Figure 3. The stages of our experimental phenotype candidate system.
doi:10.1371/journal.pone.0072965.g003

Definition: A a gene/gene product (abbreviated as GG) entity is a mention of one of the three major macromolecules DNA, RNA or protein.

Examples include: [cryoglobulins], [anticariolipin antibodies], [AFM044xg3], [chromosome 17q], [CC16 protein].

Definition: A chemical or drug entity (abbreviated as CD) entity is a mention of a chemical part or family other than genes and gene products (DNA, RNA and protein).

Kim *et al.* [51] indicate in the GENIA encoding manual that chemical entities contain element chemicals and compound chemicals, where the later can be either organic or inorganic. Here we apply a granular cut off to organic chemicals, considering that proteins and nucleic acid compounds are a separate entity class called GG. Small biomolecules are included within the scope of CD.

Following Corbett et al. [52] and the CALBC challenge guidelines [53] we include chemical compounds, molecular formulas, IUPAC nomenclature and drug names within scope.

Examples include: [Panadol], [antibiotic], [calcium], [3-ethyl-2-methylhexane] and [$C_6H_{12}O_6$].

Definition: An anatomy entity (abbreviated as AN) is a mention of an anatomical structure or other physical component within or on the surface of the human or mouse body, including organs, cells, portions of bodily substances such as blood, body fluids, tissues and their combinations.

The definition here follows on from that in Scheuermann *et al.* [50] except that (a) we apply a granular cut off at the level of cell (but include cell internal structures such as nucleus). Units smaller than a cell may be included in either CD (chemical or drug) or GG (gene/gene product), and (b) we apply AN only to the morphology of human and mouse organisms.

Examples include: [endothelial cells], [liver], [nervous system], [HeLa cells], [left collar bone], [both kidneys].

Definition: A phenotype entity (abbreviated as PH) is a textual mention that describes an observable and measurable characteristic of an organism. Phenotype entities can be further broken down into an affected entity and a describing quality for that entity.

Examples include: [differences in the levels of the protein], [airway inflammation], [absent ankle reflexes].

Our definition of phenotype require two caveats (a) in contrast to Khordad *et al.* [42] we did not apply a granular cut off at the level of cell, and (b) because of the diversity of phenotypes across organisms we took a decision to focus our definition of this entity on mouse as a model organism and human as the most important species. Following the discussion of phenotypes as processes in physiology [3] we include some mentions of processes within the scope of our annotation schema.

Definition: A disease entity (abbreviated as DS) is a mention of a disposition to undergo pathological processes in an organism because of one or more disorders in that organism.

Examples include: [Felty's syndrome], [rheumatoid arthritis], [heterozygous C2 deficiency], [Paget's disease], [inherited skeletal dysplasia].

Definition: An organism entity (abbreviated as OR) is mention of a type of living biological system which functions as a stable whole.

This definition is adapted from Beisswanger *et al.*'s [47] concept for *living organism* (BioTop ID LivingOrganism). In common with both BioTop and the GENIA ontology [54] we include both multi-cellular and mono-cellular organisms within this definition. For simplicity we also include viruses within this definition.

Our definition of an organism entity encompasses both mentions of names of species as well as individuals of those species. Individuals can be named or in some cases described.

Examples include: [Hepatitis type B virus], [food sanitation inspectors], [cholera cases], [hypergammaglobulinaemic patients], [45-year-old male], [asthmatics].

We should not however ignore the important lexical and syntactic considerations about how to annotate mentions in text. Within the annotation guidelines we developed we further describe whether specific, generic, underspecified and negatively quantified mentions qualify. This is summarised in Table 1. We follow [55] in differentiating between (a) *specific mentions* with specific reference to objects or group of objects, (b) *generic mentions* which refer to the kind of entity, (c) *underspecified mentions* which have non-generic non-specific reference, e.g. everyone, and (d) *negative mentions* which refer to the empty set of the kind of entity.

Data preparation

The Phenominer A corpus (available as Data S1 or on request from the first author) contains 122 abstracts selected from Medline. 19 auto-immune diseases were selected from OMIM and from these records citations were then chosen. Citations were only selected for the corpus if they contained the auto-immune disease term and at least one term from either OMIM's clinical synopsis field, the HPO [56] or the MP [57]. This strategy is designed to ensure that the abstracts have some association to phenotypes or anatomical entities in addition to the disease itself. Table 2 shows the 19 diseases and the corresponding affected organism. Descriptive statistics are shown in Table 3. Despite being small, the number of annotated entities is consistent with several previous specialised studies, e.g. [18,42,58].

Corpus annotation was carried out by a single experienced annotator who had previously worked on the GENIA corpus and the BioNLP shared task corpus. The annotator is not one of the authors and is independent from the experiments. Tool support was provided by the BRAT annotation tool (http://brat.nlplab.org). Entities were annotated using the commonly used **B**egin **I**n **O**ut annotation scheme, so for example *between airway responsiveness* would be annotated with the sequence *O B-PH I-PH* where 'O' denotes a word outside an entity, 'B' a word at the beginning of an entity, and 'I' as a word inside an entity.

Table 2. Auto-immune diseases from OMIM represented in the Phenominer A corpus.

Disease	Organism
Auto immune thyroid disease	human
Auto immune skin diseases	human
Immune mediated diseases	human
Immuno-mediated gastrointestinal diseases	human
Celiac's disease/Caliac disease	human
Grave's disease/Grave disease	human
Hashimoto's disease/Hashimoto disease	human
Crohn's disease/Crohn disease	human
Addison's disease/Addison disease	human
Type 1 diabetes	human
Rhematoid arthritis	human
Multiple sclerosis	human
Systemic lupus erythematosus	human
Asthma	human
Familial psoriasis	human
Auto immune encephalomyeliti	mouse
Inflammatory arthritis	mouse
Histamine sensitization	mouse
Mouse lupus	mouse

doi:10.1371/journal.pone.0072965.t002

Experimental system

Our experiments were divided into two stages. In the first stage we wanted to find the optimal combination of external resources for the range of entity types described above. The hypothesis resolution approach used in these experiments was the same as our previous method in [30], i.e. a priority list. After this we froze the

Table 1. Referential semantics and scoping of mentions by entity type.

	specific	generic	underspecified	modifiers	conjunctions	processes	negation
	reference	reference	reference		disjunction		
GG	Yes	Yes	No	No	Yes[1]	No	No
DS	Yes	Yes	No	No[4]	Yes[1]	No	No
CD	Yes	Yes	No	No	Yes[1]	No	No
OR	Yes[2]	Yes	No	No	Yes[1]	No	No
AN	Yes	Yes	No	Yes[3]	Yes[1]	No	No
PH	Yes	Yes	No	Yes[3,7]	Yes[1]	Yes[5]	Yes[6]

Notes on annotation:

[1] Where there is elision of the head, e.g. [IA/H5 virus], then annotate the whole expression. Otherwise annotate each expression separately, e.g. [IA virus] and [H5 virus].

[2] Markable expressions include specific people, e.g. [Jane] as well as definite noun phrases such as, the [24-year-old man].

[3] Quantitive modifers are included, e.g. [both kidneys] as well as spatial modifiers, e.g. [left collar bone].

[4] When modifiers are considered to be part of the disease name they are included, e.g. [highly pathogenic avian influenza], [end-stage renal disease].

[5] We exclude however finite verb forms, infinite verb forms with to', verbs in a progressive or perfect aspect, verb phrases, clauses or sentences and any phrase with a relative clause or complement clause.

[6] If the negation appears in a noun phrase with an anatomical entity then we generally allow it, e.g. [absent ankle reflexes], [no left kidney].

[7] Qualitative modifiers are included. For example, physical components: [black hair], underspecified ranges: [normal height], locational modifers: [low set ears], and level modifiers: [quite small fingers].

doi:10.1371/journal.pone.0072965.t001

Table 3. Descriptive statistics for entities in the Phenominer A corpus.

Entity	# Entities	# Unique Entities	Average length of entity
PH	472	393	3.0
OR	764	402	1.8
DS	875	270	1.9
GG	1611	885	1.7
AN	188	132	2.2
CD	48	31	1.4

doi:10.1371/journal.pone.0072965.t003

external resource features and proceeded to compare hypothesis resolution strategies. Three approaches were evaluated.

Figure 3 shows the complete system. The pre-processing stage collects the abstracts from the source provider (PubMed), splits the text into sentences and tokenises using the OpenNLP library with a Maximum Entropy model. This is then followed by abbreviation expansion using BioText [59]. Abbreviations are replaced using their full forms if they are given in the abstracts.

Three distinctive classification modules are applied within the NER system. The first of these Rule matching' follows a similar approach to Khordad's use of MetaMap (UMLS) with staged rules for post-processing [42] and a modifier list derived from HPO (85 terms) and PATO. We also added the Gene dictionary from NCBI to this module in line with our original experimental system. The second module is Dictionary matching'. This uses a longest string matching approach to identify term candidates for each entity class in the relevant ontology. For example, FMA and the Brenda tissue ontology for AN entities, Jochem for CD entities, PATO/MP/HPO for PH entities and so on. A precise list of the resources and term counts is given in the Introduction. Finally the third module is a Maximum Entropy with Beam Search (ME+BS) supervised sequence labeler using multiple linguistic features associated with the training corpus. Features include the focus word, surrounding context words, part of speech labels. Additionally we added semantics tags from a ME+BS model trained on the JNLPBA corpus [38] and Linnaeus [45]. The JNLPBA corpus contains 2000 Medline abstracts selected by a search using terms *human, blood cell, transcription factor* and then hand annotated for 5 NE classes including RNA, DNA and protein which we merge to form our GG class.

Experiment 1: Rule-based hypothesis resolution with multiple ontologies. Based on our best performing approach from [30] we applied a hybrid method to entity recognition across the six classes. For the variable component we wanted to test the influence of each standard ontology and so used ablation to knock out' each resource in turn, thereby measuring its contribution to the accuracy for each class.

A Maximum Entropy model with Beam search (ME+BS) [60] was selected as the machine learning method using the Java-based OpenNLP toolkit (http://opennlp.apache.org/) with default parameters. At this stage we treat NER assignment of tags as a sequence labeling problem. This is implemented through a sliding window of features around the target word being classified and by optimisation of the sequence of tag assignments during the decoding phase, i.e. through the beam search algorithm.

The Resolution module for deciding the final class of the entity based on competing hypotheses used a ranked priority list of hand built rules as described in our previous experiments. In summary this gives priority to labels in the following order: $DS > PH > GG > CD > AN > OR$. This judgment was based on introspective analysis of terms, e.g. that phenotypes usually contain an anatomy or a gene component (*pannus formation*, *elevated serum levels of cartilage oliomeric matrix protein*), and that genes sometimes contain a organism name (*mouse H19 gene*, *mouse ABcg2/Breast cancer resistance protein (BCRP) gene*). However organism names never contain a gene name and anatomy names.

In contrast to straightforward supervised learning our system combines the traditional machine learning based approach to NER, with its advantage of context sensitivity and compensation for lexical variation, with other approaches that bootstrap extant domain vocabularies. For example, the Mammalian Phenotype Ontology contains the term *skull anomaly, congenital* but in the text this may appear as the more general mention *congenital anomalies*. A number of string matching algorithms have been adapted for identifying synonyms and related terms such as [25] whilst others have tried to normalise external resources to a standard format [61]. As we might expect, performance has been found to vary considerably across resources and entity types. Here we use a simple longest string matching strategy between the text and the term in the external resource but normalising for plurals. As noted previously, hypothesis resolution is conducted sentence by sentence using a staged set of rules, given here from [30]:

1. We combine the putative entity labels by collecting any entity-specific result that has been proposed by at least one module. This is intended to maximise recall. The O tag (non-entity label) has the least priority.
2. Based on our ontological analysis of PH and GG it is often possible for a GG to form a fully embedded part of a PH mention. For example, [high [IgE]$_{GG}$ levels]$_{PH}$. We therefore apply a longest span rule and give priority to PH over GG giving [high IgE levels]$_{PH}$.
3. If there is a boundary conflict, we merge neighbouring entity mentions that share parts of their token sequence. For example, if we have [AB]$_{GG}$ and [BC]$_{PH}$ then we merge them into one phrase [ABC] and label it with the highest priority tag, i.e. PH. Although this appears rare in GG and PH we included this rule for expandability when we want to introduce further entity classes.

The testing framework was 10-fold cross validation using the Phenominer A corpus described in Data Preparation, i.e. the corpus is partitioned in 10 rounds so that 9 equal parts are used for training the models and the remaining 1 equal part of unseen data is used for testing. Results are collected from each of the 10 testing partitions and the accuracy is calculated against the reference standard.

Our primary purpose in these experiments is to focus on the contribution made for phenotype candidate recognition but at the same time to take into consideration the effects that resources have on the recognition performance of other entity types.

Experiment 2: Alternative hypothesis resolution strategies. The baseline method we chose used the priority list approach used in Experiment 1. This is shown as a flow diagram in Figure 4. We have outputs from 7 labelers: Rule matching, PH dictionary matching, DS dictionary matching, CD dictionary matching, AN dictionary matching and GG dictionary matching and a ME+BS tagger. Outputs from these labelers were screened using an Unambiguous/Ambiguous case detection module. Where we detected a labeling conflict, i.e. an ambiguous case, we used the priority list approach to resolve this and chose only one output, otherwise, the agreed output was considered as the final output.

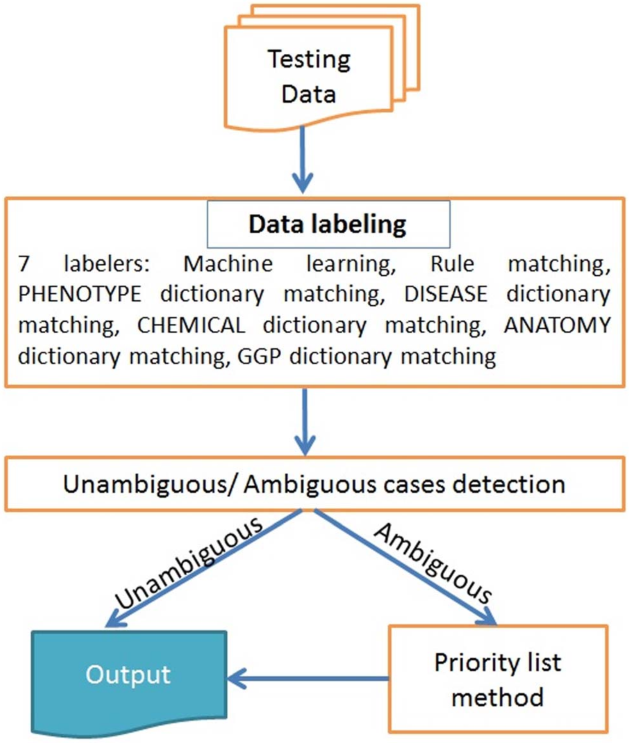

Testing Data
Data labeling
7 labelers: Machine learning, Rule matching, PHENOTYPE dictionary matching, DISEASE dictionary matching, CHEMICAL dictionary matching, ANATOMY dictionary matching, GGP dictionary matching
Unambiguous/ Ambiguous cases detection
Unambiguous
Ambiguous
Output
Priority list method

Figure 4. Hypothesis resolution using a priority list.
doi:10.1371/journal.pone.0072965.g004

Figure 5 illustrates a possible scenario for an unambiguous and an ambiguous case. Labelers 1 to 5 represent the different modules providing alternative hypotheses. In the unambiguous case two label sequences are proposed as PH for *X Y* and GG for *W Z* but there is no conflict and the final labelling will be *B-BF I-BF O B-GG I-GG* under the BIO scheme. In the ambiguous case there are multiple alternative hypotheses suggested for the first token *A* with the labelers suggesting PH, GG, O and AN.

Whilst our priority list approach seemed to perform adequately we wanted to investigate other hypothesis resolution strategies based on machine learning using the 10-fold validation framework we employed in Experiment 1.

Maximum entropy model with beam search. The first alternative that we explored was a Maximum Entropy Model [60,62] with beam search (ME+BS) as shown in Figure 6. The maximum entropy estimate is the least biased estimate possible on the given information, i.e. it is maximally noncommittal with regard to missing information.

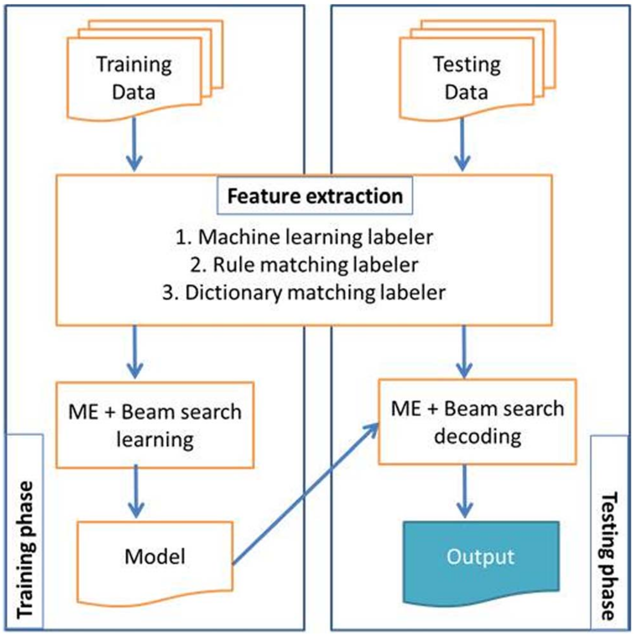
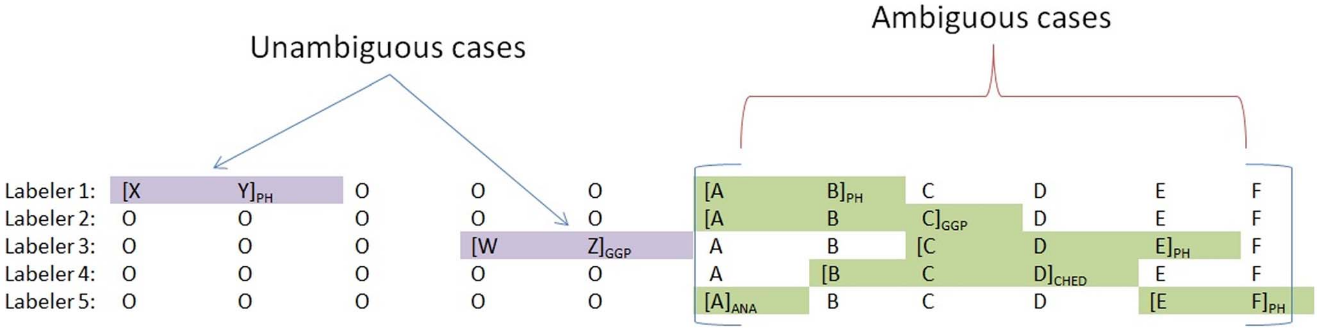

Training Data
Testing Data
Feature extraction
1. Machine learning labeler
2. Rule matching labeler
3. Dictionary matching labeler
ME + Beam search learning
ME + Beam search decoding
Model
Output
Training phase
Testing phase

Figure 6. Hypothesis resolution using maximum entropy with beam search (MS+BS).
doi:10.1371/journal.pone.0072965.g006

The original Maximum Entropy model for named entity labeling used the Viterbi algorithm for decoding, a dynamic programming technique. Instead of Viterbi we used beam search decoding. Beam search is a variant of breadth first search using a parameter k to decrease the search space (in our model, we set $k=3$). The advantage of using beam search is that it allows the tractable use of maximum entropy for each labeling decision but forgoes the ability to find the optimal label sequence using dynamic programming techniques. The computational complexity of beam search decoding is $O(kT)$ compare to $O(N^T)$ for Viterbi decoding (in which, T is the number of words, N is the number of labels). To implement ME+BS, we used the Java-based OpenNLP toolkit (http://opennlp.apache.org/) with default parameters.

The outputs from the machine learning (ME+BS) labeler, rule based labeler and dictionary based labelers were used as features to train the ME+BS resolution model, then, we used this model to choose the final output. Note that in contrast to the other two hypothesis resolution methods, this approach did not apply



Labeler 1:	[X	Y]$_{PH}$	O	O	O	[A	B]$_{PH}$	C	D	E	F
Labeler 2:	O	O	O	O	O	[A	B	C]$_{GGP}$	D	E	F
Labeler 3:	O	O	O	[W	Z]$_{GGP}$	A	B	[C	D	E]$_{PH}$	F
Labeler 4:	O	O	O	O	O	A	[B	C	D]$_{CHED}$	E	F
Labeler 5:	O	O	O	O	O	[A]$_{ANA}$	B	C	D	[E	F]$_{PH}$

Figure 5. Handling ambiguous versus unambiguous cases.
doi:10.1371/journal.pone.0072965.g005

Table 4. Features used by the Maximum Entropy model for hypothesis resolution.

No.	Feature	Example
1	Current word	w_i
2	Context words	$w_{i-2},w_{i-1},w_{i+1},w_{i+2}$
3	ME+BS labels	$ml_i : B-GG,ml_{i-1} : B-PH,ml_{i+1} : I-PH,ml_{i+2} : O$
4	Rule matching labels	$rule_i : B-PH,rule_{i-1} : I-PH,rule_{i+1} : O$
5	PH dictionary labels	$dm_i^1 : B-PH,dm_{i-1}^1 : I-PH,dm_{i+1}^1 : O$
6	DS dictionary labels	$dm_i^2 : B-DS,dm_{i-1}^2 : I-DS,dm_{i+1}^2 : O$
7	CD dictionary labels	$dm_i^3 : B-CD,dm_{i-1}^3 : I-CD,dm_{i+1}^3 : O$
8	AN dictionary labels	$dm_i^4 : O,dm_{i-1}^4 : B-AN,dm_{i+1}^4 : O$
9	GG dictionary labels	$dm_i^5 : B-GG,dm_{i-1}^5 : I-GG,dm_{i+1}^5 : O$

doi:10.1371/journal.pone.0072965.t004

screening for unambiguous or ambiguous cases since it resolved the conflict with the sequence labeling technique. The features we employed are shown in Table 4.

Support vector machine with learn-to-rank. With an appropriate scoring function it is possible to consider the choice of alternative named entity labels from the various modules and dictionaries as a ranking problem. This means that each source is scored against certain criteria and the scores are then compared with the highest one being chosen. We implemented this using the SVM^{rank} software from Thorsten Joachims at Cornell University (http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html). The experimental system is shown in Figure 7. Essentially processing proceeds token by token through the sentence. When an ambiguous token is discovered - one in which there is more than one alternative label being proposed by the labelers - SVM^{rank} is used to decide on the named entity tag.

In the first stage we applied the same screening technique as the priority list approach for unambiguous/ambiguous case detection. Unambiguous cases are considered as the final output labels with no further processing. For ambiguous cases, three rules were used to create the ranked lists. Through the feature extraction module, these ranked lists were used to trained an SVM learn-to-rank model. Then we used this model to choose the final output if conflict appeared in the test set.

In training ranking was decided by the following three heuristic rules:

1. Candidates having the same label with the training annotation receive the highest rank. Among these, candidates matching closer to the left hand side of the annotated sequence have a higher rank than candidates which match further to the right since we process the sequence in a left to right order.
2. Candidates having a partial overlap in tag assignment with the training annotation receive the second rank. Among these, candidates matching closer to the left hand side of the sequence have a higher rank than candidates which match further to the right. Again this is because we process the sequence in a left to right order.
3. Candidates that have no overlap in tag assignment with the training annotation receive the lowest rank.

SVM^{rank} is trained using these heuristics and compared against the ME+BS and priority list methods.

Matching metrics

We follow standard metrics of evaluation for the task using F1, i.e. the harmonic mean of recall and precision. This is calculated as follows:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad (1)$$

where precision indicates the percentage of system positives that are true instances, and recall indicates the percentage of true instances that the system has retrieved. More formally this is shown by the following two equations and Table 5.

$$Precision = \frac{TP}{TP + FP} \qquad (2)$$

$$Recall = \frac{TP}{TP + FN} \qquad (3)$$

Different applications require a different approach to defining a true positive. In these experiments we consider a correct match to be recorded when a *partial matching* occurs, i.e. when the span of text that is manually annotated in the gold standard corpus and the span of text output as an entity by the NER tagger partially overlap. For example a system annotation of *[median cleft lip]/palate* would be judged correct for a gold standard annotation of *median [cleft lip/palate]*. Various authors in the biomedical NER domain such as [63] have offered a reason for why this or other methods such as sloppy left boundary matching might be preferred to strict matching for genes and proteins. In summary it is thought that with partial matching, for the entity types examined so far, the core part of the entity was in most cases correctly found. In contrast, strict matching places too much faith on possibly arbitrary annotation choices as well as corpus selection, meaning that system performance might not be repeated on new texts outside the narrow domain of the gold standard. However whilst our focus is on partial matching we have included results for exact matching for comparison purposes.

Significance tests

Based on [64,65], we compared performance across different systems using an approximate randomization approach for testing significance. In order to calculate significance for two different systems (system A and system B) on the Phenominer corpus (with i sentences), we performed the following steps:

- (1) Compute micro-average F-scores using 10 fold cross validation from each system and note the difference in performance $f = f_A - f_B$;
- (2) Generate set S (with $2 \times i$ sentences) by taking the outputs from the 10 fold validations on the two systems;
- (3) Obtain i sentences randomly from set S to create set A_j, the remainder of S is set B_j (A_j is used for system A and B_j is used for system B);
- (4) Calculate $f_j = f_{A_j} - f_{B_j}$ (in which, f_{A_j} and f_{B_j} are micro-average F-scores using 10 fold cross validation for set A_j and B_j respectively).

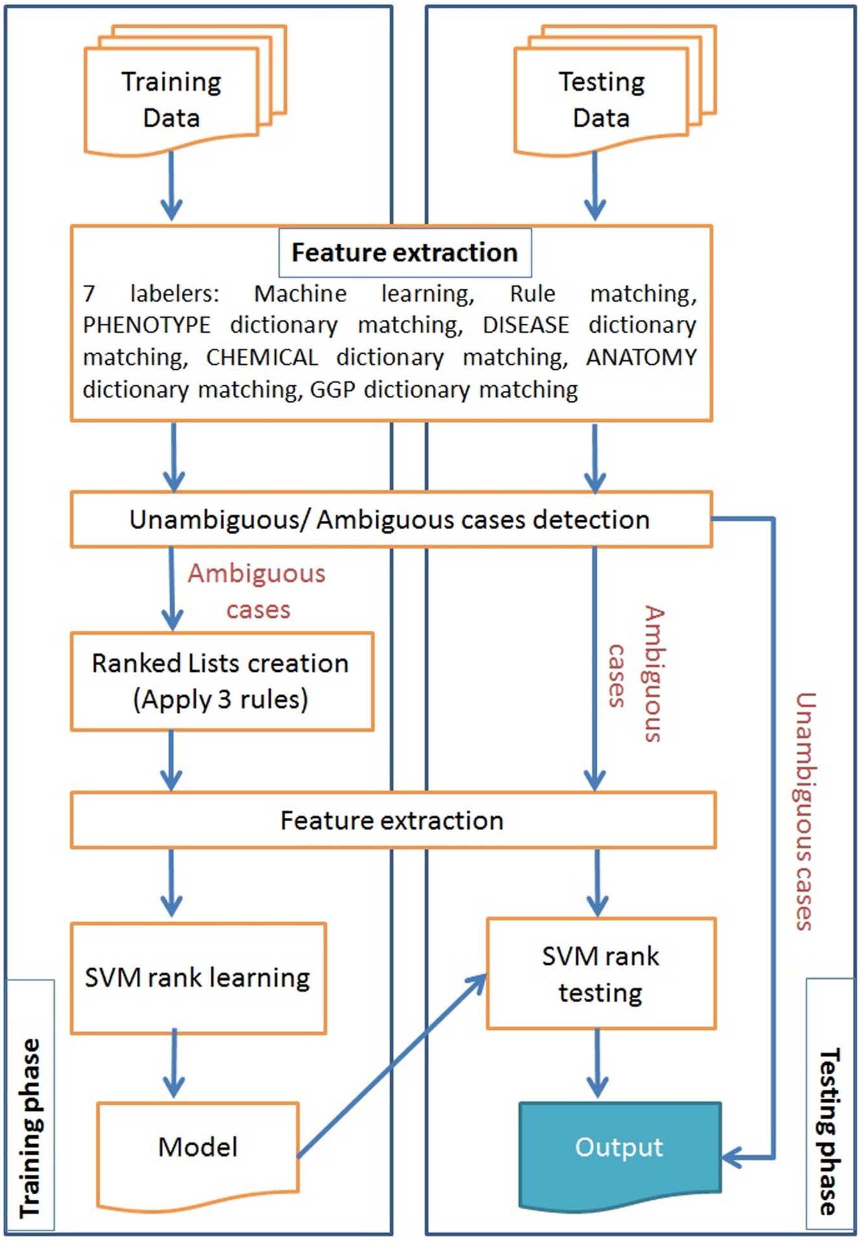

Training Data
Testing Data
Feature extraction
7 labelers: Machine learning, Rule matching, PHENOTYPE dictionary matching, DISEASE dictionary matching, CHEMICAL dictionary matching, ANATOMY dictionary matching, GGP dictionary matching
Unambiguous/ Ambiguous cases detection
Ambiguous cases
Ambiguous cases
Unambiguous cases
Ranked Lists creation (Apply 3 rules)
Feature extraction
SVM rank learning
SVM rank testing
Model
Output
Training phase
Testing phase

Figure 7. Hypothesis resolution using support vector machine and learn to rank (SVM+LTR).
doi:10.1371/journal.pone.0072965.g007

Table 5. Defining the test metrics.

		Gold standard class	
		True	**False**
System	True	TP	FP
			(Type 1 error)
class	False	FN	TN
		(Type 2 error)	

doi:10.1371/journal.pone.0072965.t005

Steps 2–4 were repeated n times (we set $n = 1000$ as in [64]). The number of times that $f_j - f \leq 0$ in n loops divided by n is the p-value between system A and system B.

Results and Discussion

Results

Resources contribution. Table 6 shows the contribution by each external resource by comparing F-scores for each NE class when it is removed from the system. As noted above, a partial matching metric was used. For comparison we include the same evaluation using exact matching in Table 7. Performance for PH is notably lower using exact matching, indicating the challenge caused by their high variability and length (see Table 3). The last row is the result when applying all resources; the hypothesis resolution module used the priority list method. All external resources help to increase the F1, but the contribution varies among them. Some resources help to increase the result greatly whereas others just bring minor improvements; some resources seem to be important for only one NE class but others affect many entities.

Using the ME+BS model trained on the JNLPBA corpus brings much better results for GG (85.2% compared to 71.0%) whereas using Gene Dictionary from NCBI helped GG to gain from 82.7% to 85.2%. Both the HPO as well as the MP help PH to increase from 61.8% and 54.4% to 74.9% respectively. The use of PATO allows the PH score to increase only slightly (from 74.7% to 74.9%). Linnaeus seems to play an important role in recognizing OR; when removing Linnaeus, OR's result is down significantly from 75.4% to 49.9%. Similarly, removing the FMA results in a drop in performance for AN from 77.1% to 59.0%, but removing the Brenda Tissue Ontology just makes AN's result drop slightly to 76.0%. Jochem's dictionary focuses on CD, resulting in a very large increase of 38.8% (from 41.6% to 80.4%). Using UMLS and MetaMap helps increase results for both PH (from 68.3% to 74.9%) and DS (from 61.4% to 74.3%).

Using the approximate randomization approach we calculated significance scores for these results. These are shown in Figure 8 and highlight resource contributions with the rows and columns showing which resource was not used in the system (e.g. *J* means the system did not use JNLPBA trained ME+BS model feature, *AR* means all resources are used). The corresponding cell shows entities which have a significance test value for difference in performance between two systems with p< = 0.05. For example, the cell in row *AR* and column *H* marked with *PH*, means there was a significant test value for *PH* for difference in performance when a system without HPO (H) was compared to a system with All Resources (AR) with p< = 0.05. Hyphen (-) stands for No significant difference', meaning that there is no entity which has significant test value with p< = 0.05. The significance scores highlight the contribution of UMLS to three NE classes (BG,GG and DS), the MP to phenotype candidates (PH) and GG, as well as the ineffectiveness of PATO for our corpus.

Resolution methods. In the resolution module we used three separate method for resolving conflict: a rule-based method (priority list), Maximum Entropy with beam search decoding and SVM learn-to-rank. The results are shown on Table 8. Maximum Entropy has the worst results with F-score of 74.9. F-score for the Priority list approach is 79.2% and SVM learn-to-rank has the best result with 79.9%. SVM learn-to-rank shows its advantage compared to the Priority List approach across almost all entity classes, included PH, GG, CD and AN with the exception of OR and DS. Table 9 shows the significance test results for the resolution module.

Because the difference between results of SVM learn-to-rank and Priority List is quite small (0.7%), we try to investigate the results in more detail in the Discussion section below to get an understanding behind the complex contributing factors.

In order to obtain and understanding about how the model performed on unique mentions, i.e. those that did not appear in the training set, we provide a side by side comparison in Table 10. The table shows a relatively large fall in performance for phenotypes from 75.3% to 62.8%. The drop in performance for each class appears proportional to the rate of unique entities.

Discussion

Our first impression was that the use of all resources had contributed to increasing the results. Examples of mentions in the corpus where we noticed a gain in recall with each of the resources are given in Table 11.

The greatest contributions we observed came from Jochem's dictionary for CD (+38.8%) and Linnaeus for OR (+25.5%). We interpret this result as reasonable because of the referential semantics and scoping of our entity mentions as well as the completeness of these resources: OR contains many generic references which are very hard to recognize for the machine learning labeler or the rule-based labeler (such as [family], [case], [cohort], etc.), Linnaeus helped to resolve these cases; Jochem's dictionary is a very large and comprehensive resource which combines information from UMLS, MeSH, ChEBI, DrugBank, KEGG, HMDB, and ChemIDplus.

Both HPO and MP affect PH's results in a positive way. However although the two resources both look at phenotypes, what they contribute is quite different because of their structures.

Table 6. Performance of named entity recognition using using partial matching for ME+BS in machine learning labeler and priority list in resolution module.

External resources										*Named entity classes*						
J	**U**	**H**	**M**	**G**	**L**	**F**	**P**	**C**	**B**	***PH***	***OR***	***AN***	***GG***	***CD***	***DS***	***ALL***
−	+	+	+	+	+	+	+	+	+	73.7	75.6	76.2	**71.0**	78.9	74.2	68.8
+	−	+	+	+	+	+	+	+	+	**68.3**	72.1	76.8	83.2	78.7	**61.4**	73.1
+	+	−	+	+	+	+	+	+	+	**61.8**	74.0	77.1	84.8	80.4	73.6	73.7
+	+	+	−	+	+	+	+	+	+	**54.4**	75.2	75.6	85.0	80.4	73.2	72.1
+	+	+	+	−	+	+	+	+	+	74.6	75.4	77.1	**82.7**	80.4	74.3	78.9
+	+	+	+	+	−	+	+	+	+	73.2	**49.9**	76.7	85.2	79.3	73.8	77.4
+	+	+	+	+	+	−	+	+	+	74.9	75.4	**59.0**	85.2	80.4	74.3	77.1
+	+	+	+	+	+	+	−	+	+	**74.7**	75.4	77.1	85.2	80.4	74.3	79.1
+	+	+	+	+	+	+	+	−	+	74.9	75.4	77.1	85.2	**41.6**	74.3	75.2
+	+	+	+	+	+	+	+	+	−	74.9	75.4	**76.0**	85.2	80.4	74.3	79.1
+	+	+	+	+	+	+	+	+	+	74.9	75.4	77.1	85.2	80.4	74.3	79.2

Each horizontal row shows a combination of features and the associated F-scores for each class on test data. *ALL* shows micro-averaged F-score. Key to external resources: J: JNLPBA model, U: UMLS and MetaMap, H: Human Phenotype Ontology, M: Mammalian Phenotype Ontology, G: Gene Dictionary from NCBI, L: Linnaeus, F: Foundation Model of Anatomy, P: Phenotypic Trait Ontology, C: Jochem's dictionary, B: Brenda Tissue Ontology.

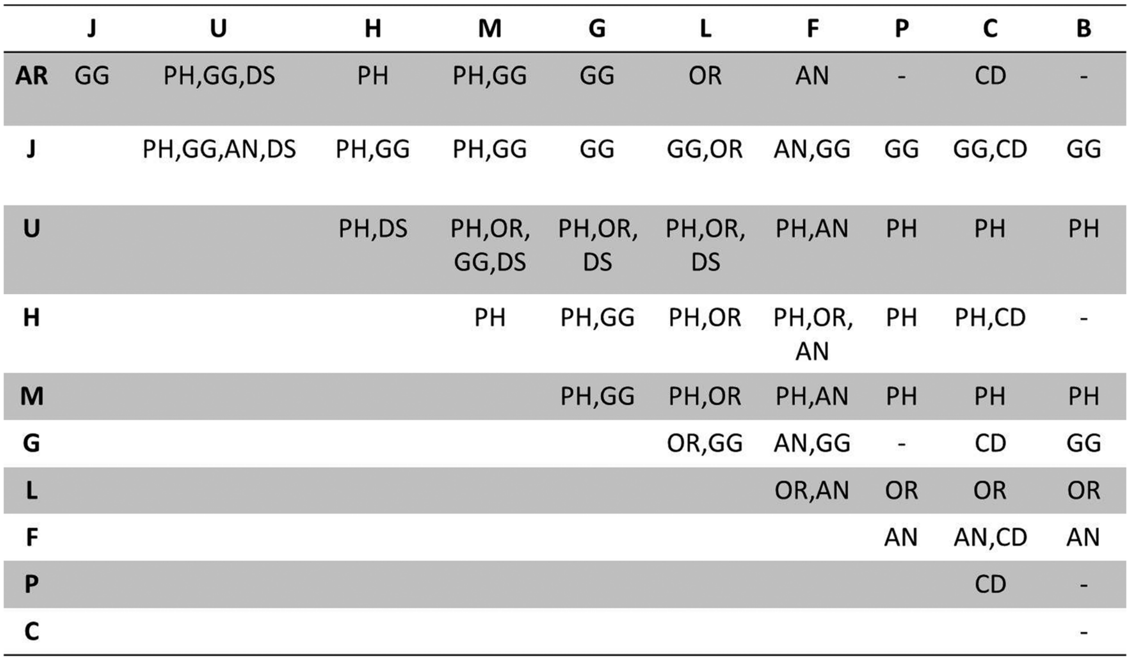
doi:10.1371/journal.pone.0072965.t006

Table 7. Performance of named entity recognition using exact matching for ME+BS in machine learning labeler and priority list in resolution module.

External resources										*Named entity classes*						
J	**U**	**H**	**M**	**G**	**L**	**F**	**P**	**C**	**B**	*PH*	*OR*	*AN*	*GG*	*CD*	*DS*	*ALL*
−	+	+	+	+	+	+	+	+	+	36.0	61.3	58.0	48.5	71.3	55.3	50.1
+	−	+	+	+	+	+	+	+	+	35.3	60.4	58.0	57.1	71.2	49.4	52.2
+	+	−	+	+	+	+	+	+	+	33.5	58.2	58.0	56.4	71.3	54.3	53.0
+	+	+	−	+	+	+	+	+	+	30.0	57.4	57.4	58.7	71.3	53.7	52.7
+	+	+	+	−	+	+	+	+	+	36.0	61.3	58.0	58.2	71.3	55.3	54.4
+	+	+	+	+	−	+	+	+	+	35.4	35.6	57.6	59.2	70.8	55.0	53.2
+	+	+	+	+	+	−	+	+	+	36.3	61.3	39.2	59.2	71.3	55.3	54.5
+	+	+	+	+	+	+	−	+	+	35.5	61.3	58.0	59.2	71.3	55.3	55.4
+	+	+	+	+	+	+	+	−	+	36.3	61.3	58.0	59.2	38.4	55.3	55.3
+	+	+	+	+	+	+	+	+	−	36.3	61.3	56.9	59.2	71.3	55.3	55.3
+	+	+	+	+	+	+	+	+	+	36.3	61.3	58.0	59.2	71.3	55.3	55.4

Each horizontal row shows a combination of features and the associated F-scores for each class on test data. *ALL* shows micro-averaged F-score. Key to external resources: J: JNLPBA model, U: UMLS and MetaMap, H: Human Phenotype Ontology, M: Mammalian Phenotype Ontology, G: Gene Dictionary from NCBI, L: Linnaeus, F: Foundation Model of Anatomy, P: Phenotypic Trait Ontology, C: Jochem's dictionary, B: Brenda Tissue Ontology.
doi:10.1371/journal.pone.0072965.t007

Note that we estimated the overlap between HPO and MP using approximate string matching giving an estimate for overlap of about 481 root terms, or 4.9% of the HPO root terms and 5.5% of the MP root terms. The phenotype mentions in our corpus appear to be more similar to MP than HPO (MP increase PH's results by +20.5% while HP increased PH's results by +13.1%). It is worth noting that some PH mentions are not recognisable directly in either resource although with transformation and the application of semantic functions such as generality matching this should improve. For example, *serum total immunoglobin* as a PH would match to the MP entry *abnormal serum level of immunoglobin/increased serum level of immunoglobin G*. To avoid an unacceptable increase in false negatives this requires deeper semantic analysis than we have provided here, to decompose the term into entity and quality parts. We will focus more on this in future work.



	J	**U**	**H**	**M**	**G**	**L**	**F**	**P**	**C**	**B**
AR	GG	PH,GG,DS	PH	PH,GG	GG	OR	AN	-	CD	-
J		PH,GG,AN,DS	PH,GG	PH,GG	GG	GG,OR	AN,GG	GG	GG,CD	GG
U			PH,DS	PH,OR, GG,DS	PH,OR, DS	PH,OR, DS	PH,AN	PH	PH	PH
H				PH	PH,GG	PH,OR	PH,OR, AN	PH	PH,CD	-
M					PH,GG	PH,OR	PH,AN	PH	PH	PH
G						OR,GG	AN,GG	-	CD	GG
L							OR,AN	OR	OR	OR
F								AN	AN,CD	AN
P									CD	-
C										-

Figure 8. Statistical significance tests for differences in performance using approximate randomization on resources contributions. The entries in cells indicate that the two systems are significantly different in F-scores. AR: All resources, J: JNLPBA model, U: UMLS and MetaMap, H:Human Phenotype Ontology, M: Mammalian Phenotype Ontology, G: Gene Dictionary from NCBI, L: Linnaeus, F: Foundation Model of Anatomy, P: Phenotypic Trait Ontology, C: Jochem's dictionary, B: Brenda Tissue Ontology, -: No significant difference. Significance is decided at p< = 0.05.
doi:10.1371/journal.pone.0072965.g008

Table 8. Performance of named entity recognition using Priority List (PL), ME plus beam search (ME+BS)and SVM learn-to-rank (SVM+LTR).

	PL			***ME+BS***			***SVM+LTR***		
NE class	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
PH	73.7	76.1	74.9	73.3	68.2	70.7	74.3	76.4	**75.3**
GG	87.0	83.5	85.2	84.7	84.0	84.4	86.8	85.0	**85.9**
OR	72.8	78.1	**75.4**	62.1	65.9	63.9	70.2	77.2	73.5
CD	79.6	81.3	80.4	74.2	71.6	72.9	80.5	81.4	**80.9**
AN	72.4	82.5	77.1	69.4	71.6	70.5	75.6	80.1	**77.8**
DS	75.8	72.9	**74.3**	71.9	70.4	71.1	73.2	71.6	72.4
ALL	-	-	79.2	-	-	74.9	-	-	**79.9**

Each horizontal row shows Precision, Recall and F-score performance for a class using alternative methods. *ALL* shows micro-averaged F-score.
doi:10.1371/journal.pone.0072965.t008

Table 9. Statistical significance tests for differences in performance using approximate randomization on Resolution methods.

	Priority list	**ME+BS**
SVM LRT	GG, OR	PH, GG, OR, AN, DS
Priority list		PH, GG, OR, AN, DS

The entries in cells indicate that the two systems are significantly different in F-scores. CD has no significant difference for all tests. Significance is decided at p< = 0.05.
doi:10.1371/journal.pone.0072965.t009

With regard to anatomical entities it is clear that the FMA has greater coverage on the Phenominer A corpus than the Brenda Tissue Ontology which focuses on tissue. This results in the FMA gaining AN +18.1% whereas using the Brenda Tissue Ontology only gave +1.1%. For genes and proteins, using a sequence labeler trained on the JNLPBA corpus resulted in GG's result increasing by +14.2% but using the NCBI Gene Dictionary only gave an increase of +2.5%.

Table 10. Performance of named entity recognition using SVM learn-to-rank (SVM+LTR) for all entities in the cross-validation test and unique entities only.

	All mentions			***Unique only***			**Unique**
NE class	**P**	**R**	**F**	**P**	**R**	**F**	**Rate**
PH	74.3	76.4	**75.3**	65.4	60.3	62.3	26.2
GG	86.8	85.0	**85.9**	80.2	79.4	79.8	14.6
OR	70.2	77.2	73.5	67.3	69.3	68.3	22.9
CD	80.5	81.4	**80.9**	74.3	71.0	72.6	41.3
AN	75.6	80.1	**77.8**	71.3	72.6	72.0	19.2
DS	73.2	71.6	72.4	70.1	69.2	69.7	12.3
ALL	-	-	**79.9**	-	-	73.2	-

Each horizontal row shows Precision, Recall and F-score performance for a class using alternative methods. *Unique Rate* shows the percentage of unique entity mentions seen in the cross-validation test for each class. *ALL* shows micro-averaged F-score.
doi:10.1371/journal.pone.0072965.t010

Table 11. Examples of mentions in the corpus where we noticed a gain in recall with each of the resources.

No.	**Resource**	**Entity example**	**Named entity class**
1	JNLPBA ME+BS corpus	[human gammaglobulin]	PH
		[eukaryotic elongation factor 1A-1]	PH
		[high-affinity human mAb]	PH
2	UMLS & MetaMap	[disorder of the Steroidogenic Acute Regulatory Protein]	PH
		[Dermatitis Herpetiformis]	DS
		[uveitis]	DS
3	HPO	[immunoglobulin abnormality]	PH
		[asthma phenotype]	PH
		[autoimmunity]	PH
4	MP	[oxidative stress pathway]	PH
		[intestinal inflammation]	PH
		[insulitis]	PH
5	Gene dictionary	[CEACAM6]	GG
		[COL29A1]	GG
		[Slc30A8]	GG
6	Linnaeus	[adenoviruses]	OR
		[murine]	OR
		[adherent-invasive E. coli]	OR
7	FMA Ontology	[lung]	AN
		[multiple organ systems]	AN
		[central nervous system]	AN
8	PATO	[high IgE levels]	PH
9	Jochem dictionary	[S €nitrosoglutathione]	CD
		[histamine]	CD
		[dapsone]	CD
10	Brenda Tissue ontology	[ileal mucosa]	AN

Named entity class is the correct results.
doi:10.1371/journal.pone.0072965.t011

Table 12. Errors by resolution module using Priority List (PL) and SVM learn-to-rank (SVM LTR).

No.	Entity	CA	ML	RB	DB					*Merge module*	
					PH	GG	DS	CD	AN	PL	LTR
1	[susceptibilities to autoimmune disease]	PH	PH	-	-	-	DS[a]	-	-	DS	**PH**
2	[asthma and atopy phenotypes]	PH	PH	-	PH[b]	-	DS[c]	-	-	DS	**PH**
3	[IgE levels]	PH	GG	-	PH[d]	-	-	-	-	**PH**	GG
4	[Toll-like receptor/ IL-1R pathways]	PH	GG	-	-	GG[e]	-	-	-	GG	GG
5	[MyD88-deficiency]	PH	GG	-	-	-	-	-	-	GG	GG
6	[allergen-induced bronchial inflammation]	PH	DS	-	-	-	-	-	-	DS	DS

CA: Corpus annotation. Key to labeler: ML: Machine Learning labeler, RB: Rule-based labeler, DB: Dictionary-based labeler. PL: Priority list, LTR: SVM- Learn to rank. The resources which the dictionary-based labelers used to recognize the entity are as follows:

[a] UMLS C0004364,

[b] HP 0002099,

[c] UMLS C0004096,

[d] MP 0002492 and HP 0003212,

[e] NCBI Gene dictionary.

doi:10.1371/journal.pone.0072965.t012

Finally, the UMLS and MetaMap have been shown to be effective cross-class resources, using them increased results for both PH by +6.6% and DS by +12.9%.

In Table 12, we show several examples of errors by the Priority List and SVM learn-to-rank. Examples 1 and 2 show where the Priority List disagreed with the gold standard annotation about a mistaken disease mention but SVM learn-to-rank agreed. In example 3, the Priority List is correct but SVM learn-to-rank is incorrect.

The Priority List method appears in a minority of cases to be too strict where there is ambiguity in making a choice. These include systematic ambiguities between DS and PH, OR and DS, PH and OR, etc. For example, the Priority List gives a higher priority to DS over PH. This rule is correct in the case of diseases included in the HPO (e.g. *[asthma]*$_{DS}$, *[allergy]*$_{DS}$) but it is incorrect if entities have the form: phenotype of disease' (e.g. *[addison disease only (ADO) phenotype]*$_{PH}$, *[asthma-related phenotypes]*$_{PH}$, *[pathogenesis of early-onset persistent asthma]*$_{PH}$). Similarly, the rule giving DS priority over OR is correct if a disease appears in human or mouse (*[human autoimmune disease]*$_{DS}$) but incorrect if a particular individual has a disease (e.g. *[lupus patients]*$_{OR}$, *[non-obese diabetic (NOD) mouse]*$_{OR}$). For these ambiguities, SVM learn-to-rank shows its advantage, as it is more flexible than the Priority List and can choose the final label based on many factors.

However, in many cases the Priority List is still a strong choice of resolution method. For example, based on our ontological analysis of PH and GG it is often possible for a GG to form a fully embedded part of a PH mention. Non-conforming examples seem to be very rare. Thus, the rule that PH takes priority over GG may bring correct results in the majority of cases while SVM learn-to-rank's flexibility is unneeded.

Finally, it is important to mention that our resolution module only affects the final output if ambiguity is detected. Rows 4–6 in Table 12 show examples of where both the Priority List and SVM learn-to-rank disagreed with the Phenominer A annotation. Because there isn't any labeler output conflict, the incorrect final results come from the incorrect results of input modules.

Conclusions

In this article we have presented a systematic study of how to combine sequence labels from various ontological resources and methods in an attempt to address the task of phenotype candidate recognition. The study is the first we believe to evaluate such a rich set of features for the complex class of phenotypes. Our system achieved the best micro-averaged F-score for the six entity classes of 79.93 with 75.31 for phenotype candidates in the auto-immune domain. We observed the advantage of using SVM learn-to-rank for hypothesis resolution and using all resources. We conclude that selected semantic types such as chemicals and genes are well covered by single semantic resources whereas phenotype candidates require combinations. In this respect key roles were observed for the Mammalian Phenotype Ontology, the Human Phenotype Ontology and the UMLS.

Our approach has coped well with the compositional structure of phenotype representations. We note though that so far we have used these ontologies as terminology resources and there will undoubtedly be potential to exploit the structures within their hierarchies in ways that can extend performance further. Beyond this, the next step is to take the phenotype candidates and decompose them according to domain concepts, i.e. to ground them. This will allow free text articles to be linked through community vocabularies, streamlining phenotype vocabulary and enabling the systematic investigation of disease-gene relationships through textual data integration.

Supporting Information

Data S1 Annotated data for the auto-immune corpus of PubMed abstracts.
(ZIP)

Author Contributions

Conceived and designed the experiments: NC DR MT AO HL. Performed the experiments: MT HL. Analyzed the data: NC MT HL. Wrote the paper: NC MT HL QH AO DS.

References

1. Hoehndorf R, Schofield P, Gkoutos G (2011) Phenomenet: a whole-phenome approach to disease gene discovery. Nucleic Acids Research.
2. Smedley D, Oellrich A, Kohler S, Ruef B, Westerfield M, et al. (2013) PhenoDigm: analyznig curated annotations to associate animal models with human diseases. Nucleic Acids Research 2013: bat025.
3. Hoehndorf R, Harris MA, Herre H, Rustici G, Gkoutos GV (2012) Semantic integration of physiology phenotypes with an application to the cellular phenotype ontology. Bioinformatics 28: 1783–1789.
4. Mabee P, Ashburner M, Cronk Q, Gkoutos G, Haendel M, et al. (2007) Phenotype ontologies: the bridge between genomics and evolution. Trends in Ecology and Evolution 22: 345–350.
5. Blake A, Sinclair MT, Sugiyarto G (2003) Quantifying the impact of foot and mouth disease on tourism and the UK economy. Tourism Economics 9: 449–465.
6. McKusick V (2007) Mendelian Inheritance of Man and Its Online Version, OMIM. American Journal of Human Genetics 80: 588–604.
7. Smith CL, Goldsmith CW, Eppig JT (2004) The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biology 6: R7.
8. Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. Nature Genetics 25: 19–29.
9. Gkoutos GV, Green ECJ, Mallon AM, Hancock JM, Davidson D (2004) Pacific symposium on biocomputing (PSB 2004), Hawai'i, USA. In: Building mouse phenotype ontologies. pp. 178–189.
10. Hirschman L, Burns G, Krallinger M, Arighi C, Bretonnel Cohen K, et al. (2012) Text mining for the biocuration workflow. Database 2012.
11. Dowell K, McAndrews-Hill M, Hill D, Drabkin D, Blake J (2009) Integrating text mining into the MGI biocuration workflow. Database bap019.
12. Freimer N, Sabatti C (2003) The human phenome project. Nature genetics 34: 15–21.
13. Robinson PN, Kohler S, Bauer S, Seelow D, Horn D, et al. (2008) The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. The American Journal of Human Genetics 83: 610–615.
14. Rosse C, Mejino JLV (2003) A reference ontology for bioinformatics: the Foundational Model of Anatomy. Journal of Biomedical Informatics 36: 478–500.
15. Hoehndorf R, Oellrich A, Rebholz-Schuhmann R (2010) Interoperability between phenotype and anatomy ontologies. Bioinformatics 24: 3112–3118.
16. Campos D, Matos S, Oliveira J (2013) Gimli:open source and high-performance biomedical name recognition. BMC Bioinformatics 14.
17. Cui L, Bozorgi A, Lhatoo S, Zhang G, Sahoo S (2012) Epidea: Extracting structured epilepsy and seizure information from patient discharge summaries for cohort identification. In: AMIA Annual Symposium Proceedings. American Medical Informatics Association, volume 2012, p. 1191.
18. Collier N, Nobata C, Tsujii J (2000) Extracting the names of genes and gene products with a hidden Markov model. In: Proceedings of the 18th International Conference on Computational Linguistics (COLING'2000), Saarbrucken, Germany. pp. 201–207.
19. Kazama J, Makino T, Ohta Y, Tsujii J (2002) Tuning support vector machines for biomedical named entity recognition. In: Workshop on Natural Language Processing in the Biomedical Domain at the Association for Computational Linguistics (ACL) 2002. pp. 1–8.
20. Settles B (2005) ABNER: an open source tool for automatically tagging genes, proteins, and other entity names in text. Bioinformatics 21: 3191–3192.
21. Leaman R, Gonzalez G (2008) BANNER: an executable survey of advances in biomedical named entity recognition. In: Proceedings of the Pacific Symposium on Biocomputing (PSB 2008), Hawai'i, USA. pp. 652–663.
22. Shen D, Zhang J, Su J, Zhou G, Tan C (2004) Multi-criteria-based active learning for named entity recognition. In: Proceedings of the 42nd Annual Meeting of hte Association for Computational Linguistics, Barcelonia, Spain. pp. 589–596.
23. de Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X (2011) Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. Journal of the American Medical Informatics Association 18: 557–562.
24. Krauthammer M, Rzhetsky A, Morozov P, Friedman C (2000) Using BLAST for identifying geneand protein names in journal articles. Gene 259: 245–252.
25. Hanisch D, Fundel K, Mevissen HT, Zimmer R, Fluck J (2005) Prominer: rule-based protein and gene entity recognition. BMC Bioinformatics 6: S14.
26. Fukuda K, Tsunoda T, Tamura A, Takagi T (1998) Toward information extraction: identifying protein names from biological papers. In: Proceedings of the Pacific Symposium on Biocomputing'98 (PSB'98). pp. 707–718.
27. Tsuruoka Y, Tsujii J (2003) Boosting precision and recall of dictionary-based protein name recognition. In: Ananiadou S, Tsujii J, editors, Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine. pp. 41–48.
28. Zhou G, Su J (2004) Exploring deep knowledge resources in biomedical named entity recognition. In: Collier N, Ruch P, Nazarenko A, editors, Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA), Geneva, Switzerland. pp. 96–99.
29. Vlachos A, Gasperin C (2006) Bootstrapping and evaluating named entity recognition in the biomedical domain. In: Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language Processing and Biology. pp. 138–145.
30. Collier N, Tran M, Le H, Oellrich A, Hall-May M, et al. (2012) A hybrid approach to finding phenotype candidates in genetic texts. In: Proceedings of the 24th International Conference on Computational Linguistics (COLING'2012), Mumbai, India.
31. Carpenter B (2007) Lingpipe for 99.99% recall of gene mentions. In: Proceedings of the Second BioCreative Challenge Evaluation Workshop. volume 23, pp. 307–309.
32. Tsuruoka Y, Tateisi Y, Kim JD, Ohta T, McNaught J, et al. (2005) Developing a robust part-ofspeech tagger for biomedical texts. In: Bozanis P, Houstis E, editors, Advances in Informatics: 10th Panhellenic Conference on Informatics, Volos, Greece, Proceedings. LNCS, Springer. pp. 382–392.
33. Okazaki N (2007) CFRsuite: a fast implementation of conditional random fields (CRFs). Available from http://www.chokkan.org/software/crfsuite (last accessed 27th July 2013).
34. Smith L, Tanabe L, Ando R, Kuo C, Chung I, et al. (2008) Overview of BioCreative II gene mention recognition. Genome Biology 9 (Suppl 2): s2.
35. Rebholz-Schuhmann D, Yepes A, Li C, Kafkas S, Lewin I, et al. (2011) Assessment of ner solutions against the first and second CALBC silver standard corpus. Journal of Biomedical Semantics 2: S11.
36. Ohta T, Pyysalo S, Tsujii J, Ananiadou S (2012) Open-domain anatomical entity mention detection. In: Proceedings of the Workshop on Detecting Structure in Scholarly Discourse. Association for Computational Linguistics, pp. 27–36.
37. Verspoor K, Cohen K, Lanfranchi A, Warner C, Johnson H, et al. (2012) A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. BMC Bioinformatics 13: 207.
38. Kim J, Ohta T, Tsuruoka Y, Tateisi Y, Collier N (2004) Introduction to the bio-entity recognition task at JNLPBA. In: Collier N, Ruch P, Nazarenko A, editors, Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA), Geneva, Switzerland. . pp. 70–75. Held in conjunction with COLING'2004.
39. Kim JD, Ohta T, Tateishi Y, Tsujii J (2003) GENIA corpus - a semantically annotated corpus for bio-textmining. Bioinformatics 19: 180–182.
40. Tanabe L, Xie N, Thom LH, Matten W, Wilbur WJ (2005) GENETAG: a tagged corpus for gene/protein named entity recognition. BMC Bioinformatics 6: S3.
41. Chen L, Friedman C (2004) Extracting phenotypic information from the literature via natural language processing. Studies in health technology and informatics 107: 758–762.
42. Khordad M, Mercer RE, Rogan P (2011) Improving phenotype name recognition. In: Advances in Artificial Intelligence, Lecture Notes in Computer Science, volume 6657/2011. pp. 246–257.
43. Groza T, Hunter J, Zanki A (2012) Supervised segmentation of phenotype descriptions for the human skeletal phenome using hybrid methods. BMC Bioinformatics 13: 265.
44. Groza T, Zankl A, Hunter A (2012) Experiences with modeling composite phenotypes in the skeletome project. In: et al CM, editor, The Semantic Web ISWC 2012, Springer Berlin/Heidelberg, Lecture Notes in Computer Science. pp. 82–97.
45. Gerner M, Nenadic G, Bergman C (2010) Linnaeus: A species name identification system for biomedical literature. BMC Bioinformatics 11: 85.
46. Gremse M, Chang A, Schomburg I, Grote A, Scheer M, et al. (2011) The BRENDA tissue ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. Nucleic Acids Research 39: D507–D513.
47. Beisswanger E, Schulz S, Stenzhorn H, Hanh U (2008) BioTop: an upper domain ontology for the life sciences. International Journal of Applied Ontology 3: 205–212.
48. Di Eugenio B, Glass M (2004) The kappa statistic: A second look. Computational Linguistics 30: 95–101.
49. Kawazoe A, Jin L, Shigematsu M, Barerro R, Taniguchi K, et al. (2006) The development of a schema for the annotation of terms in the BioCaster disease detection/tracking system. In: KRMED 2006: Proc. Int. Workshop on Biomedical Ontology in Action, Baltimore, USA. pp. 77–85.
50. Scheuermann R, Ceusters W, Smith B (2009) Toward an ontological treatment of disease and diagnosis. In: AMIA Summit on Translational Bioinformatics, San Francisco, CA. pp. 116–120.
51. Kim JD, Ohta T, Tateisi Y, Tsujii J (2006) GENIA corpus manual - encoding schemes for the corpus and annotation. Technical report TR-NLP-UT-2006-1, School of Information Science, University of Tokyo.
52. Corbett P, Batchelor C, Teufel S (2007) Annotation of chemical named entities. In: Proceedings of the Association of Computational Linguistics Workshop on Biomedical Natural Language Processing, Prague, Czech Republic. pp. 57–64.
53. Rebholz-Schuhmann D, Jimeno-Yepes AJ, van Mulligen EM, Kang N, Kors J, et al. (2010) CALBC silver standard corpus. Journal of Bioinformatics and Computational Biology 8: 163–179.
54. Tateisi Y, Ohta T, Collier NH, Nobata C, Tsujii J (2000) Building an annotated corpus from biology research papers. In: Proc. COLING 2000 Workshop on

Semantically Annotated Corpora and Intelligent Content, Saarbrucken, Germany. pp. 28–34.

55. Magnini B, Pianta E, Popescu O, Speranza M (2006) Ontology population from textual mentions: task definition and benchmark. In: Proc. ACL/COLING Workshop on Ontology Population and Learning (OLP2), Sidney, Australia. pp. 26–32.
56. Robinson PN, Mundlos S (2010) The human phenotype ontology. Clinical Genetics 77: 525–534.
57. Smith CL, Eppig JT (2009) The mammalian phenotype ontology: enabling robust annotation and comparative analysis. Wiley Interdisciplinary Reviews: Systems Biology and Medicine 1: 390–399.
58. Suakkaphon N, Zhang Z, Chen H (2011) Disease named entity recognition using semisupervised learning and conditional random fields. Journal of the American Society for Information Science and Technology 62: 727–737.
59. Schwartz A, Hearst M (2003) A simple algorithm for identifying abbreviations in biomedical text. In: Pacific Symposium on BioComputing, Hawai'i, USA. pp. 451–462.
60. Ratnaparkhi A (1996) A maximum entropy model for part-of-speech tagging. In: Conference on Empirical Methods in Natural Language Processing. University of Pennsylvania, pp. 133–142.
61. Tuason O, Chen L, Liu H, Blake J, Friedman C (2004) Biological nomenclatures: a source of lexical knowledge and ambiguity. In: Proc. of the Pacific Symposium on Biocomputing, Big Island, Hawai'i, USA. pp. 238–249.
62. Borthwick A, Sterling J, Agichtein E, Grishman R (1998) Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In: Charniak E, editor, Proceedings of the Sixth Workshop on Very Large Corpora (WVLC'98) at COLING-ACL'98, Montreal, Canada. pp. 152–160.
63. Kabiljo R, Clegg A, Shepherd A (2009) A realistic assessment of methods for extracting gene/protein interactions from free text. BMC Bioinformatics 10: 233.
64. Uzuner, Solti I, Cadag E (2010) Extracting medication information from clinical text. Journal of the American Medical Informatics Association 17: 514–518.
65. Doan S, Collier N, Xu H, Duy P, Phuong T (2012) Recognition of medication information from discharge summaries using ensembles of classifiers. BMC Medical Informatics and Decision Making 12: 36.