A hybrid approach to finding phenotype candidates in genetic texts

Nigel Collier[1,2] *Mai – Vu Tran*[1,3] *Hoang – Quynh Le*[1,3]
Anika Oellrich[4] *Ai Kawazoe*[1] *Martin Hall – May*[5]
Dietrich Rebholz – Schuhmann[2,6]
(1) National Institute of Informatics, Tokyo, Japan
(2) European Bioinformatics Institute, Cambridge, UK
(3) University of Engineering and Technology - VNU, Hanoi, Vietnam
(4) Wellcome Trust Sanger Institute, Cambridge, UK
(5) University of Southampton, Southampton, UK
(6) The University of Zurich, Zurich, Switzerland
collier@nii.ac.jp,vutm@vnu.edu.vn,lhquynh@vnu.edu.vn
ao5@sanger.ac.uk,zoeai@nii.ac.jp
mhm@ecs.soton.ac.uk,rebholz@ebi.ac.uk

ABSTRACT
Named entity recognition (NER) has been extensively studied for the names of genes and gene products but there are few proposed solutions for phenotypes. Phenotype terms are expected to play a key role in inferring gene function in complex heritable diseases but are intrinsically difficult to analyse due to their complex semantics and scale. In contrast to previous approaches we evaluate state-of-the-art techniques involving the fusion of machine learning on a rich feature set with evidence from extant domain knowledge-sources. The techniques are validated on two gold standard collections including a novel annotated collection of 112 abstracts derived from a systematic search of the Online Mendelian Inheritance of Man database for auto-immune diseases. Encouragingly the hybrid model outperforms a HMM, a CRF and a pure knowledge-based method to achieve an F1 of 77.07. Disagreement analysis points to further improvements on this emerging NE task. The annotated corpus and guidelines are available on request.

KEYWORDS: conditional random fields, biomedicine, machine learning, genetic disorders, text mining.

Proceedings of COLING 2012: Technical Papers, pages 647–662,
COLING 2012, Mumbai, December 2012.

1 Introduction

Biomedical named entity recognition (NER) is a computational technique used to identify and classify strings of text (*mentions*) that designate important concepts in biomedicine. Over the last fourteen years there has been considerable interest in this problem with a variety of generic and entity-specific algorithms applied to extract the names of genes, gene products, cells, chemical compounds and diseases (Fukuda et al., 1998; Rindflesch et al., 1999; Collier et al., 2000; Kazama et al., 2002; Zhou et al., 2003; Settles, 2004; Kim et al., 2004; Leaman and Gonzalez, 2008). As the first stage in the integrated semantic linking of knowledge between literature and structured databases it is critically important to maximise the effectiveness of this step.

Despite significant progress in NER there is still no one size fits all solution. Barriers arise because of ambiguity in the text and coding schema. Ambiguity in the text comes in various forms according to the semantic type of the entity but can be caused by a lack of standard nomenclatures, extensive and growing nomenclatures for proteins/genes across multiple organisms or the widespread use of abbreviations and descriptive names. For example, (Krauthammer and Nenadic, 2004) illustrate uncontrolled naming in genes with *bridge of sevenless (boss)* (FlyBase ID FBgn0000206) and Hunter and Bretonnel Cohen (2006) discuss term class ambiguity (e.g. is *group* a chemical entity or an assemblage of organisms?). Such challenges have led to a variety of proposed solutions involving a wide range of resources. Among these, linguistically annotated corpora such as GENIA (Tateisi et al., 2000; Kim et al., 2003) have proven to be central to the NER solution. However due to the size of the vocabularies involved, annotated corpora by themselves do not provide a complete solution. Researchers have therefore also looked at the rich availability of formally structured biomedical knowledge (ontologies) such as the Unified Medical Language System (UMLS) (Bodenreider et al., 2002) and the Gene Ontology (Gene Ontology Consortium, 2000). Nevertheless corpora remain a key part of the solution as they provide the contextual evidence that link mentions to terms through the author's intentions. Creating such resources though is time consuming and expensive, especially when annotating new semantic types and relations.

In this paper we focus on the analysis and identification of a new class of entity: phenotypes. Two thoughts motivate this: (1) The database curation community has expressed a wish for full text entity indexing and the inclusion of phenotypes (Dowell et al., 2009; Hirschman et al., 2012), and (2) Biomedicine is rapidly moving towards full-scale integration of data, opening up the possibility to understand complex heritable diseases caused by genes. Association studies involving phenotypes are considered important to making progress (Lage et al., 2007; Wu et al., 2008). The ultimate goal of the work we present here is to allow relations mined from sentences such as the one we annotated below to feed into novel hypothesis generation procedures. From Ex 1. the reader can easily infer a relation between *IgG1 disorder* and three genes/gene products marked as GGP.

Ex 1. *Among* $[\text{patients}]_{ORG}$ *with* $[\text{systemic lupus erythematosus}]_{DIS}$ *(*$[\text{SLE}]_{DIS}$*), those with the* $[\text{IgG1 disorder}]_{PHEN}$ *have a higher prevalence of high titre* $[\text{rheumatoid factor}]_{GGP}$ *and* $[\text{antinuclear antibody}]_{GGP}$*, but a lower prevalence of* $[\text{anti-double-stranded DNA (anti-dsDNA) antibodies}]_{GGP}$ *above 30 U/ml.* (Source PMCID: PMC1003566).

Whilst other authors have tried similar approaches for other entity types, none have tried both machine learning and external resource lookup for a class as rich and semantically complex as phenotypes. The key contributions of this paper are: (1) To provide an operational semantics for identifying phenotype candidates in text, (2) To introduce a set of guidelines and an anno-

tated corpus based on a selection of 19 clinically significant auto-immune diseases from The Online Mendelian Inheritance of Man (OMIM) (Hamosh et al., 2005), one of the most widely used gene-disease databases, and (3) To mitigate linguistic variation whilst still meeting the conceptual expectations of biologists we propose a new named entity solution that uses statistical inference and external manually crafted resources. This method is tested on the new corpus and one extant corpus (Khordad et al., 2011) that has been used in previously reported experiments.

2 The challenge of phenotypes

Freimer and Sabatti (2003) describe phenotypes as referring to '*any morphologic, biochemical, physiological or behavioral characteristic of an organism. ...All phenotypic characteristics represent the expression of particular genotypes combined with the effects of specific environmental influences.*' Despite recent data integration efforts for phenotypes such as (Robinson and Mundlos, 2010), phenotypic descriptions still tend to be author/study specific and biological results may go undiscovered if the terms used lie outside an author's immediate research area (Bard and Rhee, 2004). Again, unlike genes or anatomic structures, phenotypes and their traits are complex concepts and do not constitute a homogeneous class of objects (i.e. a natural kind).

Traits such as 'eye colour', 'blood group', 'hemoglobin concentration' or 'facial grimacing' describe morphological structures, physiological processes and behaviours. When qualities or quantities of traits are used to describe a specific organism then we have phenotypic descriptions, e.g. 'blue eyes', 'blood group AB', 'not having between 13 and 18 gm/dl hemoglobin concentration'.

Traits and phenotypes can apply at all levels of anatomical granularity from chemical structures to cells and organs making it difficult to know where to draw a boundary. Phenotypes can include quantifications that are either specific (e.g. '18 gm/dl') or relative (e.g. 'normal' or 'increased'). Accordingly the first part of this paper deals with specifying exactly what we mean by the concepts of 'phenotype' with reference to current ontological research.

3 Methods

3.1 Schema

We employed two types of entity in our study: gene/gene product (GGP) and bodily feature (BF). GGP is proposed because (a) a subset of these entities are useful for applications that explore gene-phenotype relations, and (b) it allows us to compare our results against the many biomedical NER studies of the past, e.g. (Kim et al., 2004; Rebholz-Schuhmann et al., 2010). Because of space limitations we will not provide a rigidly formal definition or a taxonomic analysis (Beisswanger et al., 2008). Future work will explore the relationships between these and other entity types.

In line with BioTop (Beisswanger et al., 2008), GGP is relatively straightforward to define by the conjunction of (BioTop ID Nucleic Acid Structure) and (BioTop ID Peptide Structure).

Definition: A GGP (gene/gene product) entity is a mention of one of three major macromolecules DNA, RNA or protein. DNA and RNA are nucleic acid sequences containing the genetic instructions used in the development and function of an organism. Proteins are polypeptide sequences, or parts of polypeptide sequences, folded into structures that facilitate biological function.

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

Our definition of phenotype was taken from the formal analysis in Scheuermann *et al.* (Scheuermann et al., 2009) who define phenotype as '*A (combination of) bodily features(s) of an organism determined by the interaction of its genetic make-up and environment*'. It is important to recognise that this definition requires us to know the underlying cause. Since causality is often difficult to establish using narrow contextual evidence of the sort used in NER it seems reasonable that we focus here on identifying bodily features themselves, i.e. *phenotype candidates*, and then determine causality in another stage of processing.

Definition: A BF (bodily feature) entity is a mention of a bodily quality in an organism.

Examples include: [lack of kidney], [abnormal cell migration],[absent ankle reflexes] as well as more complex cases such as [no abnormality in his heart], [unfavorable serum lipid levels] and [suceptibility to ulcerative colitis].

Our definition of bodily features require two caveats (a) in contrast to Khordad et al. (2011) we did not apply a granular cut off at the level of cell, and (b) because of the diversity of bodily features 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 (Hoehndorf et al., 2012) we include some mentions of processes within the scope of our annotation schema.

Linguistic forms of entities require a number of policy decisions to be made about how to annotate mentions in text. For a class as complex as phenotypes this is a particular consideration. Although more complex approaches exist, for simplicity we make the common assumption here that named entities are 'continuous, non-nested and non-overlapping' (Alex et al., 2007). As a basic policy we do not allow embedding of entities within our corpus so annotators have to make a choice of entity class based on the longest matching span even though one entity may contain another entity of the same or a different type. We leave to future work consideration of other approaches, e.g. for handling discontinuous entity mentions. Within our guidelines we describe whether specific, generic, underspecified and negatively quantified mentions qualify. A summary of the rule set (available from the first author) is shown in Table 1. We follow (Magnini et al., 2006) in differentiating between specific, generic and underspecified mentions.

3.2 Annotated data sources

3.2.1 Phenominer

The Phenominer version 1 corpus contains 112 abstracts we selected from PubMed Central (PMC). 19 auto-immune diseases were selected from OMIM and from these records citations were then chosen. Diseases include Type 1 diabetes, Grave's disease, Crohn's disease, autoimmune thyroid disease, multiple sclerosis and inflammatory arthritis. In order to ground the article in discussion about both a disease and a phenotype, citations needed to contain the auto-immune disease term and at least one term from either OMIM's free form clinical synopsis field, the Human Phenotype Ontology (HPO) (Robinson and Mundlos, 2010) or the Mammalian Phenotype Ontology (MP) (Smith and Eppig, 2009).

Despite being small, the number of annotated abstracts is consistent with several previous specialised studies, e.g. (Suakkaphon et al., 2011; Collier et al., 2000). Annotation was carried

	BF	GGP
specific reference	Yes	Yes
generic reference	Yes	Yes
underspecified reference	No	No
modifiers	Yes[1,2]	No
conjunctions	Yes[3]	Yes[3]
processes	Yes[4]	No
negation	Yes[5]	No

Table 1: Referential semantics and scoping of mentions by entity type. Notes on annotation: [1] Quantitative modifiers are included, e.g. [having five fingers] as well as spatial modifiers, e.g. [abnormality in his left hand]. [2] Qualitative modifiers are included such as physical components: [black hair], underspecified ranges: [normal height], locational modifiers: [low set ears], and level modifiers: [quite small fingers].[3] Where there is elision of the head, e.g. [IA/H5 virus], then we annotate the whole expression. Otherwise we annotate each expression separately, e.g. [IA virus] and [H5 virus]. [4] We exclude 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. [5] 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].

out by the same highly experienced biomedical annotator who had annotated the GENIA corpus. The total number of tokens (sentences) in the corpus is 26,026 (1976) from which there were 1611 GGP entities and 472 BF entities.

3.2.2 KMR

As a basis of comparison we test our methods on the same corpus and tagging model as Khordad et al. (2011) who used a collection of 3784 tokens (120 sentences) with 110 annotated phenotype mentions. This is designated as the *KMR* corpus. In contrast to the Phenominer corpus sentences in KMR were taken from 4 PubMed papers from the year 2009 in the area of human genetics. Annotation was conducted with reference to the HPO so that a term was tagged as phenotype if it was in the HPO or if it was not in the HPO but its definition showed that it was caused by a genotype (Khordad, 2012). Finally we found that there was no cross over of sentences between the Phenominer and KMR corpora.

3.3 Models

The full system we developed (designated in the Results as *Hybrid*) employs machine learning and knowledge-based approaches, combined together with a rule-based Merge module. This is illustrated in Figure 1. Below we briefly describe its component modules and resources. As a baseline comparison we use Khordad's approach, designated in the Results as *Khordad* and in Figure 1 as the *Rule matching* module - see Khordad et al. (2011).

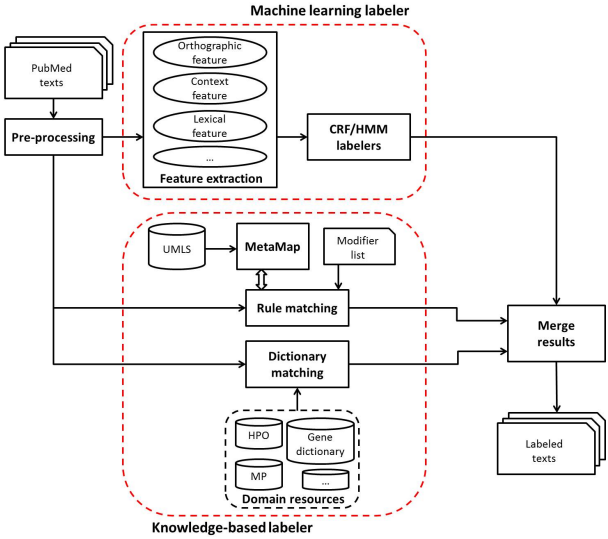

Figure 1: Phenotype tagging architecture

3.3.1 Pre-processing

The text collections are passed into a module which splits texts into sentences and tokens. This was done using the OpenNLP library with a Maximum Entropy model[1]. Abbreviation expansion is then done using BioText (Schwartz and Hearst, 2003) to make a list of local abbreviation occurring in each paper which we then replace with their full form. A similar approach is adopted by Khordad in their staged rules.

3.3.2 Machine learning labeler

Within the machine learning module we compare two widely used sequence labeling models: a second order Hidden Markov Models (HMM) (Rabiner and Juang, 1986; Bikel et al., 1997) with Viterbi decoding and a linear chain Conditional Random Fields (CRF) (Lafferty et al., 2001; McDonald and Pereira, 2005). Both are run as fully supervised models. Class labels for tokens follow the standard BIO system, i.e. each token receives the label O if it is not an NE, B plus the entity name when it starts an entity, and I plus the entity name when it is inside an entity. The main advantage of the CRF over the vanilla HMM is that it estimates the conditional probability distribution over labeled sequences. Both use the freely available Java-based MALLET implementation[2] with default parameters.

Previous research has found that utilizing various features for both the focus word (designated

[1]OpenNLP library: `http://opennlp.apache.org`
[2]Mallet: A machine learning for language toolkit: `http://mallet.cs.umass.edu/`

as w_i) and the surrounding words is crucial to obtain high performance. We take our feature set for BF and GGP labeling from the most typical and effective features used for biomedical NER (Kim et al., 2004). This is summarised in Table 2. Our experiments tested a variety of unigram, bigram and conjoined base features. These were taken from a ±2 window around the focus word for parts of speech, orthography and surface word forms. POS tagging was done using the OpenNLP library with Maximum Entropy model and Genia Corpus + WSJ Corpus (F-score 98.4%), there are 44 Penn Treebank POS tags and all of them are used. The HMM did not use conjoined features due to model memory limitations.

Feature	Description	Example
LX	Current word token	w_i
MM	MetaMap tag of the token[2]	cgab,fndg,neop
OR	Orthography of the token	initCap, isDate, allCap, isDigit
CT	Word token context	
	History context of the token	w_{i-2} , w_{i-1}
	Future context of the token	w_{i+1} , w_{i+2}
	Conjoined context	w_{i-2} . w_{i-1}
POS[1]	Part of speech tag of the token	RB, CD, NN, JJ, NNP

Table 2: Feature sets used in the machine learning labeler. [1]Part of speech tags are assigned by training the GENIA tagger (Tsuruoka et al., 2005). [2] The MetaMap semantic tags are chosen from the same group of 15 semantic types chosen by Khordad which are relevant for phenotypes.

In addition to the Phenominer and Khordad corpora outlined earlier we also make use of the JNLPBA04 corpus (Kim et al., 2004) for training the GGP labeler. The 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 GGP class. Table 3 summarises the features exploited by the two learner models.

Model	Target class	Phenominer corpus[1]	JNLPBA04 corpus	LX	MM	OR	CT	POS
HMM	BF	+	-	+	-	-	+	-
CRF	BF	+	-	+	+	+	+	+
HMM	GGP	+	+	+	-	-	+	-
CRF	GGP	+	+	+	+	+	+	+

Table 3: Resource combinations compared in our experiments. [1]This is applied within the 10-fold cross validation framework.

3.3.3 Knowledge-based labeler

The knowledge-based labeler is divided into *Rule matching* and *Dictionary matching* modules. Rule matching is an implementation of Khordad's approach using MetaMap, a subset of the UMLS, the HPO as well as 5 staged heuristics to identify phenotypes. For example, if a phrase has the form: "modifier (from the list of selected modifiers[3]) +[Anatomy] or [Physiology]" it is a phenotype name.

[3]The list of 85 high frequency modifiers from the HPO is available from the first author

Dictionary matching uses a longest string matching approach to recognise entities from the following resources: BF entities from the HPO (9500 terms describing human phenotypes) and the MP (9162 terms describing mouse phenotypes); GGP entities from the National Center for Biotechnology Information's gene list (9 million gene names). These sources were chosen because of their high standing within the biomedical community.

3.3.4 Merge results

Merge results assigns the final entity label to each token in the corpus by applying the following rules to each source module output. Processing proceeds sentence by sentence.

1. Following Jimeno et al. (2008) we combine the putative entity labels by collecting any entity-specific result that has been proposed by at least one method. This is intended to maximise recall. Thus, the O tag (non-entity label) has the least priority.
2. Based on our ontological analysis of BF and GGP it is often possible for a GGP to form a fully embedded part of a BF mention. For example, $[[\text{HLA-DQ}]_{GGP}\ \text{expression}]_{BF}$. We therefore apply a longest span rule and give priority to BF over GGP giving $[\text{HLA-DQ expression}]_{BF}$.
3. If there is a boundary conflict, we merge neighbouring entity mentions that share parts of their token sequence. For example, if we have $[\text{AB}]_{GGP}$ and $[\text{BC}]_{BF}$ then we merge them into one phrase [ABC] and label it with the highest priority tag, i.e. BF. Although this appears rare in GGP and BF we included this rule for expandability when we want to introduce further entity classes.

4 Evaluation

4.1 Metrics

We follow standard metrics of evaluation for the task using F1 as our primary method of comparison[4]. In these experiments matching is calculated using partial matching, i.e. a correct match is recorded 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 Kabiljo et al. (2009) 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 in arbitrary choices in annotation guidelines.

4.2 Experiments on the KMR corpus

Our initial test run is conducted on the KMR corpus with micro-averaged F1 scores shown in Table 4. Since the corpus only contains phenotype tags no GGP results are shown.

We noted that curiously the results we observed for Khordad's method are slightly down by approximately 3 points of F1 on those given in their article. This appears mostly to affect

[4] F1 is the harmonic mean of recall (R) and precision (P) and is calculated as $F1 = 2PR/P + R$.

precision rather than recall and is possibly due to implementation differences such as changes to MetaMap/UMLS. Given the small amount of training data it is not surprising to see the pure machine learning based methods perform relatively poorly (F1:34.07,68.29) against the pure knowledge based approach (F1:83.29). Khordad's staged rule based system with dictionary lookup performs the best (F1:89.58) with the new Hybrid approach a reasonably close second (F1:85.27). It is encouraging though to see that the rule-based combination of the learner and KB outputs add value to the KB-only result.

		Model				
Class	**Metric**	**Khordad**	**HMM**	**CRF**	**KB**[1]	**Hybrid**[2]
BF	P	90.74	37.54	65.09	87.64	86.37
	R	88.44	31.18	71.83	79.36	84.19
	F	89.58	34.07	68.29	83.29	85.27

Table 4: (F)-scores, (R)ecall and (P)recision for each entity type on the KMR corpus using models with partial matching. [1]The KB method uses the Merge module to resolve conflicts. [2]Hybrid refers to the jointly applied system.

4.3 Experiments on the Phenominer corpus

For the Phenominer data set we chose to add the GENIA NER tagger trained on the JNLPBA04 corpus as a baseline for GGP. Note that we also combined this corpus data with Phenominer for training the CRF and HMM GGP recogniser. Khordad's method remains as our baseline for phenotypes. Micro-averaged F1 scores are shown in Table 5. With regard to GGP entities we observed that whilst the GENIA tagger performed robustly (F1:80.89), the Hybrid model appears to significantly outperform this (F1:85.48). The surprising result is that Khordad's method performs relatively poorly (F1:61.38) on BF. Again we try to dig down into the results in the Discussion to get an understanding behind the complex contributing factors.

			Model				
Class	**Metric**	**Khordad**	**GENIA**[3]	**HMM**[1,2]	**CRF**[1]	**KB**[4]	**Hybrid**[5]
BF	P	65.89	-	34.67	66.32	61.24	78.21
	R	57.44	-	38.11	64.17	60.91	75.96
	F	61.38	-	36.31	65.23	61.07	77.07
GGP	P	-	78.35	64.03	76.84	92.74	86.67
	R	-	83.61	65.80	80.07	61.31	84.32
	F	-	80.89	64.90	78.42	73.82	85.48
Total	Micro avg-F	-	-	56.46	75.19	71.62	85.04
	Macro avg-F	-	-	50.61	71.83	67.45	81.28

Table 5: (F)-scores, (R)ecall and (P)recision for each entity type on the Phenominer corpus using models with partial matching. [1]HMM and CRF are trained separately on each entity class and resolved in the Merge module. [2]Training included the JNLPBA04 corpus data for GGPs. [3]The GENIA method is the GENIA NER tagger trained on the GGP entities in the JNLPBA04 corpus. [4]The KB method uses the Merge module to resolve conflicts. [5]Hybrid refers to the jointly applied system.

5 Discussion

The results on the Phenominer corpus for Hybrid (F1:77.07) on BF are very encouraging and as we hoped demonstrate the strength of combining a mildly context sensitive ML approach with knowledge base lookup. Current NE methods based on a state-of-the-art learning approach such as CRF seem well suited to non-complex NE types such as GGP but maybe less effective for complex entities such as BF. Given the small size of the corpora we must be cautious in this conclusion. With regard to the KB approach for BF, our first impression was that the phenotype resources (HPO and MP) may to some extent lack coverage on the Phenominer corpus but we discuss below why this conclusion maybe too simplistic.

We start our analysis with the necessary observation that the Phenominer and KMR corpora do not offer a strict like-for-like comparison and are therefore most useful to highlight areas of difficulty. Importantly as we noted in Section 2, there is the issue of causality which is implicitly encoded into Khordad's schema and absent from ours. This means that our bodily features may not have a genetic or environmental cause. There is also the issue of granularity: our schema is more complex as it encodes bodily features from the genetic level upwards whereas Khordad's operates on the cellular level upwards. A statistical analysis points to further differences. We found that the average phenotype mention length in the KMR corpus was 1.72 tokens with the longest term being 5 tokens: [*hypoplasia of the corpus callosum*]. In contrast the average bodily feature mention in Phenominer is 2.89 tokens with the longest being [*susceptibility to psoriasis (PS) and psoriatic arthritis (PSA)*]. The longest GGP in Phenominer is 16 tokens: [*chromosomes 1 (D1S235), 4 (D4S1647), 12 (D12S373), 16 (D16S403), and 17 (D17S1301))*]. Both of these examples from Phenominer indicate structural term issues related to coordination and elipsis which are not easily handled by the simple longest term match approach that we have adopted.

Table 6 shows examples of where the Hybrid method disagreed with the KMR corpus. Whilst we have not conducted an in-depth analysis the examples seem reasonable and indicative of differences between the two coding schemas regarding causality of a bodily feature, algorithmic differences in how we prioritize UMLS semantic types related to *Disorder* and gaps in the knowledge resources.

No.	**Standard annotation**	**System annotation**	**Issue**[1]	**Cause of error**
1	eversion of the lateral eyelid	-	FN	Cannot be found in HPO or by rule matching
2	cervical rachischisis	-	FN	Hybrid system does not include default assignment for UMLS semantic types
3	absent nervi olphactorii	-	FN	
4	-	pregnancy	FP	Bodily feature does not differentiate between normal and abnormal
5	-	female	FP	
6	-	height	FP	

Table 6: Sources of error by the Hybrid system on the KMR corpus. [1] FN: False Negative; FP: False Positive.

Table 7 looks now at examples in the Phenominer corpus where the Hybrid approach disagreed

with Khordad's model. In the table the Hybrid model output agrees with the annotated corpus and the Issue column refers to the Khordad annotation. We see in particular that differences in the schema semantics account for many of the errors. The Phenominer schema for bodily features does not include disease mentions and simple anatomical entities but these may both be considered as phenotypes by the HPO. Clearly a notion of the compositional semantic relationships between types within terms is important to fully resolve the score differences.

Since Khordad's method relies to a greater extent than Hybrid on the HPO, we tested a number of terms from the Phenominer corpus by searching for them in the HPO. Using the exact match facility in OBO-Edit[5] we found several gaps. The following terms could not be found: complex terms such as [*perivascular distribution and granular deposits of immunoglobins*] as well as some gene specific terms such as [*IGG1 disorder*]. Surprisingly several seemingly common terms such as [*kidney impairment*] and [*abnormal thyroid function*] could also not be identified from a simple exact match. In the case of [*kidney impairment*] a suitable match might be found in *Abnormality of renal physiology* (HPO ID 0000082) by replacing the organ name with its anatomical adjective. Of 12 BF mentions in the Phenominer corpus that were not in the HPO our analysis revealed that 9 of them could be found by Hybrid. The ones that were not found tended to be very long and involved either coordination or a preposition phrase.

No.	**Hybrid annotation**[2]	**Khordad annotation**	**Issue**[1]	**Cause of error**
1	pathogenic process	-	FN	These entries do not belong to the UMLS's 15 target types, and are not in the HPO, and cannot be recognised by the pattern rules.
2	gene expression	-	FN	
3	RA susceptibility	- -	FN	
4	-	Inflammatory bowel disease	FP	Although this is present in HPO it is considered as a disease in our guidelines
5	-	enteropathy bowel disease	FP	Although this is present in HPO it is considered as an anatomical entity in our guidelines
6	-	asthma susceptibility gene	FP	Although this is present in HPO it is considered as GGP in our guidelines

Table 7: Sources of error by Khordad's system on the Phenominer corpus. [1] FN: False Negative; FP: False Positive. [2] We show here Hybrid system outputs that are correctly annotated.

Finally we show examples of disagreement for the Hybrid method on the Phenominer corpus in Table 8. As is common the biomedical literature we noticed a high proportion of coordination issues as well as ambiguity caused by generic terms.

[5]OBO-Edit: the OBO ontology editor: `http://oboedit.org/`

No.	Standard annotation[2]	Hybrid annotation	Issue[1]	Cause of error
1	FEV 1	-	FN	Because of orthographic similarity to genes this is tagged as GGP
2	[asthma]$_{BF}$ and [atopy phenotypes]$_{BF}$	[asthma and atopy phenotypes]$_{BF}$	FP	Coordination creates a boundary error
3	emotion	-	FN	This generic term is context sensitive
4	Diabetes Mellitus	[Diabetes Mellitus]$_{BF}$	FP	Entity class error
5	[citrullination]$_{BF}$ of the [endogenous antigen]$_{GGP}$	[citrullination of the endogenous antigen]$_{GGP}$	FP	Boundary error due preposition phrase

Table 8: Sources of error by the Hybrid system on the Phenominer corpus. [1] FN: False Negative; FP: False Positive. [2] We show here Hybrid system outputs that are correctly annotated.

6 Conclusions and future work

We have presented new results and analysis that add evidence to how phenotype candidates can be identified using named entity technology. The methods we have employed are aimed at making tractable the annotation of a critical semantics in the scientific literature. To do this we have matched surface forms to their attested forms in domain resources, balanced against contextual evidence from annotations in the scientific literature. The benchmark tests have demonstrated that the Hybrid method performs strongly on both the KMR corpus as well as the new Phenominer corpus. The evidence points towards complementarities between the existing phenotype resources and contextual evidence from annotated corpora.

Our methods have been formulated to be simple, effective and extensible with a focus on providing input to more knowledge intensive techniques downstream that can identify causality. Simplicity though may have sacrified both precision and recall in some cases, e.g. in the issue of coordination, in including generic and underspecified references and in adopting a longest matching approach to annotation.

There is considerable scope for further investigation. F1 might be increased using a machine learning framework such as integer linear programming (Koomen et al., 2005) to resolve hypotheses against multiple constraints much as we have tried to do manually in the Merge module. Coverage might be extended by including disjoint entities and a deeper analysis of embedded entity semantics such as that employed by Alex et al. (2007). In line with Hoehndorf et al. (2010) future solutions may need to focus on decomposing phenotypes in terms of their internal relations such as qualities [6].

[6] e.g. The Phenotypic Attribute and Trait Ontology `http://obofoundry.org/cgi-bin/detail.cgi?quality`

Acknowledgments

We gratefully acknowledge partial funding from NII and from an EC Marie Curie International Incoming Fellowship award (Project: Phenominer).

References

Alex, B., Grover, C., and Haddow, B. (2007). BioNLP 2007 Workshop at ACL2007, Prague, Czech Republic. In *Recognising Nested Named Entities in Biomedical Text*, pages 65–72.

Bard, J. B. L. and Rhee, S. Y. (2004). Ontologies in biology: design, applications and future challenges. *Nature Reviews Genetics*, 5(3):213–222.

Beisswanger, E., Schulz, S., Stenzhorn, H., and Hanh, U. (2008). BioTop: an upper domain ontology for the life sciences. *International Journal of Applied Ontology*, 3:205–212.

Bikel, D., Miller, S., Schwartz, R., and Wesichedel, R. (1997). Nymble: a high-performance learning name-finder. In Grishman, R., editor, *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP'97), Washington Marriot Hotel, Washington D.C., USA.*, pages 194–201.

Bodenreider, O., Mitchell, J. A., and McCray, A. T. (2002). Evaluation of the UMLS as a terminology and knowledge resource. In *Proc. Americal Medical Informatics Association (AMIA) Annual Symposium, San Antonio, TX*, pages 61–65. AMIA.

Collier, N., Nobata, C., and 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*, pages 201–207.

Dowell, K., McAndrews-Hill, M., Hill, D., Drabkin, D., and Blake, J. (2009). Integrating text mining into the MGI biocuration workflow. *Database*, bap019.

Freimer, N. and Sabatti, C. (2003). The human phenome project. *Nature Genetics*, 34(1):15–21.

Fukuda, K., Tsunoda, T., Tamura, A., and Takagi, T. (1998). Toward information extraction: identifying protein names from biological papers. In *Proceedings of the Pacific Symposium on Biocomputing'98 (PSB'98)*, pages 707–718.

Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:19–29.

Hamosh, A., Scott, A. F., Amberger, J. S., and Bocchini, C. A. (2005). Online mendelian inheritance of man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(suppl 1):D514–D517.

Hirschman, L., Burns, G., Krallinger, M., Arighi, C., Bretonnel Cohen, K., Valencia, A., Wu, C., Chatr-Aryamontri, A., Dowell, K., Huala, E., Lourenco, A., Nash, R., Veuthey, A., Wiegers, T., and Winter, A. (2012). Text mining for the biocuration workflow. *Database*, 2012(bas020). doi:10.1093/database/base020.

Hoehndorf, R., Harris, M. A., Herre, H., Rustici, G., and Gkoutos, G. V. (2012). Semantic integration of physiology phenotypes with an application to the cellular phenotype ontology. *Bioinformatics*, 28(13):1783–1789.

Hoehndorf, R., Oellrich, A., and Rebholz-Schuhmann, R. (2010). Interoperability between phenotype and anatomy ontologies. *Bioinformatics*, 24(24):3112–3118.

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

Jimeno, A., Jimenez-Ruiz, E., Lee, V., Gaudan, S., Berlanga, R., and Rebholz-Schuhmann, D. (2008). Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 9(Suppl 3):S3.

Kabiljo, R., Clegg, A., and Shepherd, A. (2009). A realistic assessment of methods for extracting gene/protein interactions from free text. *BMC Bioinformatics*, 10(1):233.

Kazama, J., Makino, T., Ohta, Y., and 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*, pages 1–8.

Khordad, M. (2012). Personal communication by email, August 31st 2012.

Khordad, M., Mercer, R. E., and Rogan, P. (2011). Improving phenotype name recognition. In *Advances in Artificial Intelligence*, volume 6657/2011, pages 246–257. Lecture Notes in Computer Science.

Kim, J., Ohta, T., Tsuruoka, Y., Tateisi, Y., and Collier, N. (2004). Introduction to the bio-entity recognition task at JNLPBA. In Collier, N., Ruch, P., and Nazarenko, A., editors, *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA), Geneva, Switzerland*, pages 70–75. held in conjunction with COLING'2004.

Kim, J. D., Ohta, T., Tateishi, Y., and Tsujii, J. (2003). GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(Suppl.1):180–182.

Koomen, P., Punyakanok, V., Roth, D., and Yih, W. (2005). Generalized inference with multiple semantic role labeling system. In *Ninth Conference on Computational Natural Language Learning (CoNLL '05), Michigan, USA*, pages 181–184.

Krauthammer, M. and Nenadic, G. (2004). Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 37(6):512 – 526.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning, Massachusetts, USA*, pages 282–289.

Lage, K., Karlberg, E. O., Storling, Z. M., Olason, P. I., Pederson, A. G., Rigina, O., Hinsby, A. M., Tumer, Z., Pociot, F., Tommerup, N., Moreau, Y., and Brunak, S. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology*, 25:309–316.

Leaman, R. and Gonzalez, G. (2008). BANNER: an executable survey of advances in biomedical named entity recognition. In *Proceedings of the Pacific Symposium on Biocomputing, Hawai'i, USA*, pages 652–663.

Magnini, B., Pianta, E., Popescu, O., and 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*, pages 26–32.

McDonald, R. and Pereira, F. (2005). Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics*, 6(Suppl 1:S6).

Rabiner, L. and Juang, B. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, pages 4–16.

Rebholz-Schuhmann, D., Jimeno-Yepes, A. J., van Mulligen, E. M., Kang, N., Kors, J., Milward, D., Corbett, P., Bukyo, E., Beisswanger, E., and Hanh, U. (2010). CALBC silver standard corpus. *Journal of Bioinformatics and Computational Biology*, 8(1):163–179.

Rindflesch, T. C., Hunter, L., and Aronson, A. R. (1999). Mining molecular binding terminology from biomedical text. In *American Medical Informatics Association (AMIA)'99 annual symposium, Washington DC, USA*, pages 127–131.

Robinson, P. N. and Mundlos, S. (2010). The human phenotype ontology. *Clinical Genetics*, 77(6):525–534.

Scheuermann, R., Ceusters, W., and Smith, B. (2009). Toward an ontological treatment of disease and diagnosis. In *AMIA Summit on Translational Bioinformatics, San Francisco, CA*, pages 116–120.

Schwartz, A. and Hearst, M. (2003). A simple algorithm for identifying abbreviations in biomedical text. In *Pacific Symposium on BioComputing, Hawai'i, USA*, pages 451–462.

Settles, B. (2004). Biomedical named entity recognition using conditional random fields. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA) at COLING'2004, Geneva, Switzerland*, pages 104–107.

Smith, C. L. and Eppig, J. T. (2009). The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 1(3):390–399.

Suakkaphon, N., Zhang, Z., and 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(4):727–737.

Tateisi, Y., Ohta, T., Collier, N. H., Nobata, C., and 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*, pages 28–34.

Tsuruoka, Y., Tateisi, Y., Kim, J. D., Ohta, T., McNaught, J., Ananiadou, S., and Tsujii, J. (2005). Developing a robust part-of-speech tagger for biomedical texts. In Bozanis, P. and Houstis, E., editors, *Advances in Informatics: 10th Panhellenic Conference on Informatics, Volos, Greece, Proceedings. LNCS*, pages 382–392. Springer.

Wu, X., Jiang, R., Zhang, M. Q., and Li, S. (2008). Network-based global inference of human disease genes. *Systems Biology*, 4(189).

Zhou, G., Zhang, J., Su, J., Shen, D., and Tan, C. (2003). Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, 20(7):1178–1190.