A machine learning approach for phenotype name recognition

Maryam Khordad[1] *Robert E. Mercer*[1] *Peter Rogan*[1,2]
(1)Department of Computer Science
(2)Department of Biochemistry
The University of Western Ontario, London, ON, Canada
mkhordad@csd.uwo.ca, mercer@csd.uwo.ca, progan@uwo.ca

ABSTRACT

Extracting biomedical named entities is one of the major challenges in automatic processing of biomedical literature. This paper proposes a machine learning approach for finding phenotype names in text. Features are included in a machine learning infrastructure to implement the rules found in our previously developed rule-based system. The system also uses two available resources: MetaMap and HPO. As we are not aware of any available corpus for phenotype names, a corpus has been constructed. Since manual tagging of the corpus was not possible for us, we started tagging only HPO phenotypes in the corpus and then using a semi-supervised learning method, the tagging process improved. The evaluation results (F-Score 92.25) suggest that the system achieved good performance and it outperforms the rule-based system.

KEYWORDS: Phenotype, Named Entity Recognition, MetaMap, Human Phenotype Ontology.

Proceedings of COLING 2012: Technical Papers, pages 1425–1440,
COLING 2012, Mumbai, December 2012.

1 Introduction

A large amount of biomedical knowledge is available in the biomedical literature. So, automatic processing of biomedical literature to capture and formalize this embedded information is very demanding. Current Natural Language Processing (NLP) systems try to extract from the biomedical literature different types of knowledge such as, protein–protein interactions (Leroy et al., 2003) (He and DiMarco, 2005) (Fundel et al., 2007) (kiong Ng and Wong, 1999) (Yu et al., 2005) (Bui et al., 2011), new hypotheses(Swanson, 1986) (Hristovski et al., 2005) (Hristovski et al., 2006), relations between drugs, genes, and cells (Rindflesch et al., 2000) (Friedman et al., 2001) (Tanabe et al., 1999), relations between genes and diseases (Rindflesch et al., 2003) (Coulet et al., 2010), protein structure (Humphreys et al., 2000) (Gaizauskas et al., 2003), and protein function (Andrade and Valencia, 1998) (Valencia, 2005).

Fundamental to each of these applications is the Named Entity Recognition and Classification (NERC) task. Research over the past years has shown that recognizing the names of biomedical objects is not a simple task. Factors that preclude a straightforward procedure include the existence of an ever-increasing large (millions) set of entity names, a penchant for the use of abbreviations, the use of synonyms, and the fact that some biological entities have complex names consisting of many words, like "increased erythrocyte adenosine deaminase activity", and lacking agreement on the name boundaries, even among biologists (Leser and Hakenberg, 2005).

Named Entity Recognition and Classification in the biomedical domain has been extensively studied. As a consequence, many methods have been proposed. Generic methods, like MetaMap (Aronson, 2001) and mgrep (Dai et al., 2008) recognize and classify many kinds of entities in text. Some methods, however, are specialized to recognize particular types of entities like gene or protein names (Krauthammer et al., 2000) (Gaizauskas et al., 2003), diseases and drugs (Rindflesch et al., 2000) (Xu et al., 2008) (Segura-Bedmar et al., 2008), mutations (Horn et al., 2004), and properties of protein structures (Gaizauskas et al., 2003). Each method employs one or more of the following techniques (Leser and Hakenberg, 2005): (1) dictionary-based techniques (like (Krauthammer et al., 2000)) which match phrases from the text against existing dictionary entries, (2) rule-based techniques ((Fukuda et al., 1998), for instance) which make use of lexical and linguistic rules to find entity names in the text, and (3) machine learning techniques (for example, (Nobata et al., 1999)) which treat the NER task as a classification problem. Some methods use hybrid approaches to find named entities: ChemSpot (Rocktäschel et al., 2012) blends Conditional Random Fields with dictionary matching to identify chemicals in texts, a biomedical name entity recognizer (Gong et al., 2009) uses POS tagging, rules and a dictionary, and a protein name recognizer (Seki and Mostafa, 2005) uses rules, a probabilistic model and a dictionary to find protein names in biomedical text.

Every day many research experiments are performed to discover the role of DNA sequence variants in human health and disease and the results of these experiments are published in the biomedical literature. Because of the large quantity of information, a reliable automatic system to extract this information for future organization is desirable. Human phenotypes comprise a very important part of this knowledge. Phenotypes are the observable characteristics of a cell or organism, including its appearance, its morphology, physiology and ways of life. A phenotype of an organism is determined by the interaction of its genetic constitution and the environment. Skin colour, height and behaviour are some examples of phenotypes.

Currently, many systems which use phenotypes to find information like phenotype-genotype relations (for instance, (Coulet et al., 2010)) use only dictionary-based techniques to recognize the phenotypes in the text. Although many biomedical-term-specialized dictionaries are available, we are not aware of a dictionary which is both comprehensive and ideally suited for phenotype name recognition. For example, the Unified Medical Language System (UMLS) Metathesaurus (Humphreys et al., 1998) is a very large, multi-purpose, and multi-lingual vocabulary database that contains more than 1.8 million concepts. All concepts in the Metathesaurus are assigned to at least one semantic type, but *Phenotype* is not one of them. So, it alone is not adequate to distinguish between phenotypes and other objects in text. In addition, some phenotype names do not exist in the UMLS MetaThesaurus. The Pharmacogenetics Knowledge Base (PharmGKB) (Klein et al., 2001) attempts to collect all knowledge of how human genetic variation impacts drug–response phenotypes. It is a high quality database queried by clinicians and bioinformaticians. It is a manually curated database that summarizes published gene–drug–phenotype relationships. Nevertheless this manual curation process is not sustainable considering the growth of the scientific literature in this domain. The Online Mendelian Inheritance in Man (OMIM) (McKusick, 2007) is the most important single information source about human genes and genetic phenotypes (Robinson and Mundlos, 2010). Nonetheless, OMIM does not use a controlled vocabulary to describe the phenotypic features in its clinical synopsis section making it inappropriate for data mining purposes (Robinson and Mundlos, 2010). And, it is manually curated. The Human Phenotype Ontology (HPO) (Robinson and Mundlos, 2010) has been developed using information from OMIM. Although it contains approximately 10,000 terms, it is incomplete. Also, new phenotypes are being constantly introduced to the biomedical world and HPO currently is being refined, corrected, and expanded manually, meaning that it will have difficulty keeping pace with all the new phenotypes.

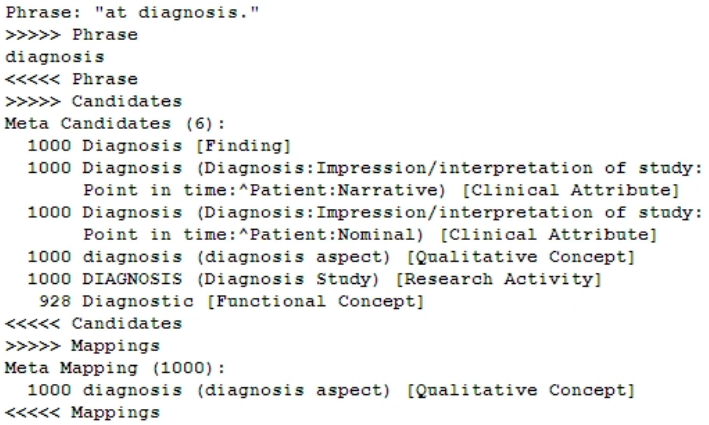
To our knowledge the only method to extract phenotype names automatically from text is the rule-based system we proposed in (Khordad et al., 2011). In the current paper we discuss a machine-learning-and-dictionary-based NER system that recognizes the human phenotype names in molecular biology literature which has been inspired by the rule-based method mentioned above. We have integrated existing databases (UMLS Metathesaurus(Humphreys et al., 1998) and the Human Phenotype Ontology (HPO) (Robinson and Mundlos, 2010)) and tools (MetaMap (Aronson, 2001) and BANNER (Leaman and Gonzalez, 2008)) to achieve our goal.

2 Background

2.1 MetaMap

MetaMap, a program developed by the National Library of Medicine (NLM) (Aronson, 2001), provides a link between biomedical text and the structured knowledge in the Unified Medical Language System (UMLS) Metathesaurus. To map phrases in the text to concepts in the UMLS Metathesaurus, MetaMap analyzes the input text lexically and semantically. First, MetaMap tokenizes the input text. In the tokenization process the input text is broken into meaningful elements, like words. After part of speech tagging and shallow parsing using the Specialist Lexicon, the text has been broken into phrases. Phrases then undergo further analysis: Each phrase is mapped to a set of candidate UMLS concepts, each candidate being given a score that represents how well the phrase matches the candidates. An optional last step is word sense disambiguation (WSD) which chooses the best candidate with respect to the surrounding text (Aronson, 2001).

MetaMap is configurable with options for vocabularies and data models in use, output format



```
Phrase: "at diagnosis."
>>>>> Phrase
diagnosis
<<<<< Phrase
>>>>> Candidates
Meta Candidates (6):
  1000 Diagnosis [Finding]
  1000 Diagnosis (Diagnosis:Impression/interpretation of study:
       Point in time:^Patient:Narrative) [Clinical Attribute]
  1000 Diagnosis (Diagnosis:Impression/interpretation of study:
       Point in time:^Patient:Nominal) [Clinical Attribute]
  1000 diagnosis (diagnosis aspect) [Qualitative Concept]
  1000 DIAGNOSIS (Diagnosis Study) [Research Activity]
   928 Diagnostic [Functional Concept]
<<<<< Candidates
>>>>> Mappings
Meta Mapping (1000):
  1000 diagnosis (diagnosis aspect) [Qualitative Concept]
<<<<< Mappings
```

Figure 1: MetaMap output for “at diagnosis.”

and algorithmic computations. An example of the human-readable output format for the text “at diagnosis.” is shown in Figure 1. MetaMap finds 6 candidates for this phrase and after WSD it maps the phrase to the “diagnosis aspect” concept. In UMLS each Metathesaurus concept is assigned to at least one semantic type. In Figure 1 the semantic type of each concept is given in the square brackets. Semantic types are categorized into groups, called Semantic Groups (SG), that are subdomains of biomedicine such as Anatomy, Living Beings and Disorders (McCray et al., 2001). Each semantic type belongs to exactly one SG.

2.2 The Human Phenotype Ontology (HPO)

The Human Phenotype Ontology (HPO) (Robinson and Mundlos, 2010) provides a standardized vocabulary of phenotypic abnormalities encountered in human disease. The HPO was constructed using information initially obtained from the Online Mendelian Inheritance in Man (OMIM) (McKusick, 2007) expanded with synonym terms. The hierarchical structure in the HPO represents the subclass relationship. The HPO currently contains over 9500 terms.

2.3 The Rule-Based Method

In (Khordad et al., 2011) we proposed a rule-enhanced dictionary-based named entity recognition system based on MetaMap. The block diagram of the system is shown in Figure 2. In this system the input text is processed and each Noun Phrase is tagged by MetaMap. The UMLS semantic types are categorized into 15 more general and comprehensive categories called Semantic Groups (SG) (McCray et al., 2001). The definition of the SG *Disorders* is close to the meaning of Phenotype. SG *Disorders* contains 12 semantic types. (Khordad et al., 2011) categorized these semantic types into two categories : *Phenotypes* and *Phenotype Candidates*. Using the MetaMap output, Disorder Recognizer considers noun phrases in the Phenotype category as phenotypes and searches for the Phenotype Candidates in HPO to see whether they are phenotypes or not. OBO-Edit (an open source Java program) (Day-Richter et al., 2007)

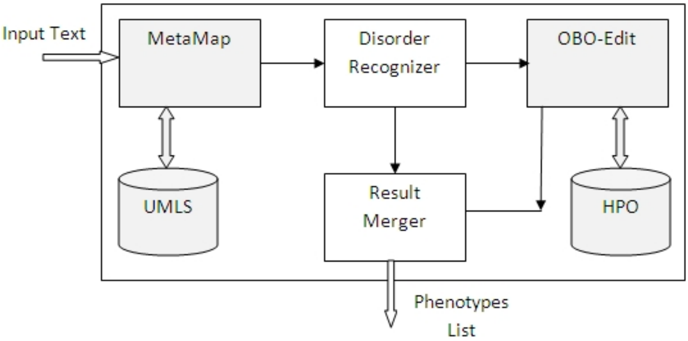


Figure 2: Rule-Based System Block Diagram.

has been used for searching in HPO. To improve the results 5 stylistic and linguistic rules are employed. These rules include:

- Rule 1: Resolve the acronym referencing problem by making and using a list of acronyms occurring in a paper.
  Often the names of phenotypes are used in acronym form. This makes it difficult to recognise them. However we can find the full form of them in their first usage.
- Rule 2: The semantic type of a noun phrase is the semantic type assigned by Metamap to its head.
  Sometimes MetaMap breaks a noun phrase to different parts with different Semantic Types. Using this rule it is possible to assign one semantic type to the whole noun phrase.
- Rule 3: If a phrase is "modifier (from the list of special modifiers (Burgun et al., 2009)) + [Anatomy] or [Physiology]" it is a phenotype.
  Some phenotypes follow a special pattern. There is a list of special modifiers (Burgun et al., 2009) which if come before a Noun Phrase in the SG *Anatomy* or *Physiology* make a phenotype.
- Rule 4: If the single form of a phrase is a phenotype, the plural form is a phenotype, too.
  This rule applies while searching in HPO, when the phenotype number in HPO is not in agreement with the noun phrase number in text.
- Rule 5: If the head of a phenotype candidate phrase is a phenotype, the whole phrase is a phenotype.
  This rule is also useful for searching in HPO. Sometimes a noun phrase in text contains a word in HPO but it is surronding with adjective and adverbs.

2.4 BANNER

BANNER (Leaman and Gonzalez, 2008) is an open-source biomedical named entity recognition system implemented using second order conditional random fields (CRF), a machine learning technique. The BANNER architecture is illustrated in Figure 3. A BANNER input file contains a text which has been separated into sentences. Each sentence is taken individually and is tokenized. The tokenization process in BANNER breaks tokens into either a contiguous block of

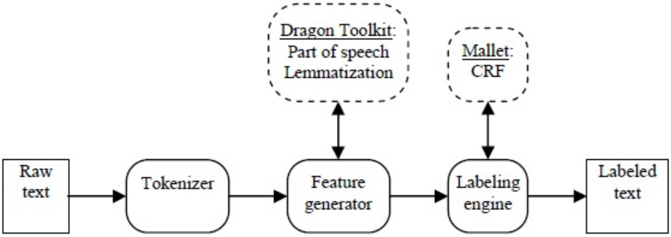


Figure 3: BANNER Architecture (Leaman and Gonzalez, 2008)

letters and/or digits or a single punctuation mark. As an example, the string "Bub2p-dependent" is broken into 3 tokens: "Bub2p", "-", and "dependent". In the next step features are assigned to each individual token. Each feature is a name/value pair for use by the machine learning algorithm. And finally in the labeling process, each feature gets exactly one label. BANNER makes use of the Mallet CRF (McCallum, 2002) in both feature generation and labeling. The set of machine learning features used in BANNNER is listed in Table 1.

Feature set definition	Description
The part of speech which the token plays in the sentence	Provided by the Dragon toolkit implementation of the Hepple tagger.
The lemma for the word represented by the token, if any	Provided by the Dragon toolkit.
A set of regular expression features	Includes variations on capitalization and letter/digit combinations.
2, 3 and 4-character prefixes and suffixes	
2 and 3 character n-grams	Including start-of-token and end-of-token Indicators
Word class Convert	upper-case letters to "A", lowercase letters to "a", digits to "0" and other characters to "x"
Numeric normalization	Convert digits to "0"
Roman numerals	
The names of the Greek letters	

Table 1: Set of features in BANNER (Leaman and Gonzalez, 2008)

BANNER considers a token window of 2 to make features, meaning that the features of each token contains the features of the two previous and the two following tokens.

BANNER has been used for NER in the Gene names and the disease names domains, and it has achieved results comparable with other NER systems in these domains.

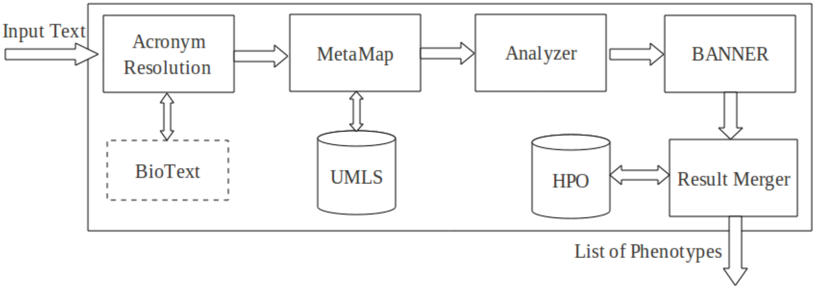


Figure 4: System Block Diagram

3 Proposed Method

The rule-based system described in Section 2.3 achieves good results. However, it, like other rule-based methods, has some shortcomings. As the rules are extracted from a small corpus, they may be overfitted to that corpus and cannot cover new phenotypes in other texts. And if we wish to use it on a larger corpus, we may need to add extra rules. It is not easy to analyse the errors manually to generate the new rules. So, we decided to extend the capabilities of our rule-based system using machine learning methods.

BANNER is open-source and it provides a good infrastructure for NER. Its results are convincing and features can be easily added. Therefore starting with BANNER and its CRF method, the rule-based method was incorporated to it.

The block diagram of the system is illustrated in Figure 4. The first phase in this system is finding acronyms in the input text and resolving them. Usually, papers indicate the local unambiguous reference for each acronym used at its first usage. So acronym resolution is done, making a list of local full forms for acronyms (according to rule 1 of the rule-based method). Then the output which does not contain any acronym is processed by MetaMap. According to our settings, MetaMap finds composite noun phrases with up to 3 simple phrases and prints out the syntax of each noun phrase. Each noun phrase is also tagged with a semantic type. The Analyzer analyses the MetaMap output and changes it to BANNER input format. Our feature-enhanced BANNER takes each sentence separately and using the features, finds some phenotypes. In the last step the system searches for the HPO phenotypes in the text and found phenotypes are added to our list of phenotypes. The details of how we made use of rules and features in our machine learning method are explained in the following sections.

3.1 Incorporating the rules

Rule 1 *Acronym Resolution* implements Rule 1. It finds the full forms of acronyms using their local unambiguous reference in the text and replaces acronyms with their unabbreviated forms. BioText (Schwartz and Hearst, 2003) has been used to make a list of acronyms and their full forms in the text.

Rule 2 This rule is implemented in *Analyzer*. Analyzer finds the noun phrases and their heads from the MetaMap output. If Metamap breaks a noun phrase into different parts and assigns different semantic types to them, the Analyzer assigns the semantic type of the head to the whole phrase. An example of such a phrase and the Analyzer output is shown in Figure 5. The phrase "of Diamond-Blackfan anemia patients" is broken into two noun phrases "Diamond-Blackfan anemia" and "patients", each with a different semantic type. MetaMap output shows that the head of the whole phrase is "patients" so analyzer assigns its semantic type, i.e. *Patient or Disabled Group*, to the whole phrase.

Rule 3 To help our machine learning method learn this rule, three binary features were added to the system: *Special Modifier*, *Anatomy* and *Physiology* to indicate whether a noun phrase is in the Special Modifier, Anatomy or Physiology classes. Furthermore, some sentences containing this class of phenotypes (*Special Modifier* + [Anatomy] Or [Physiology]) were added to our training set.

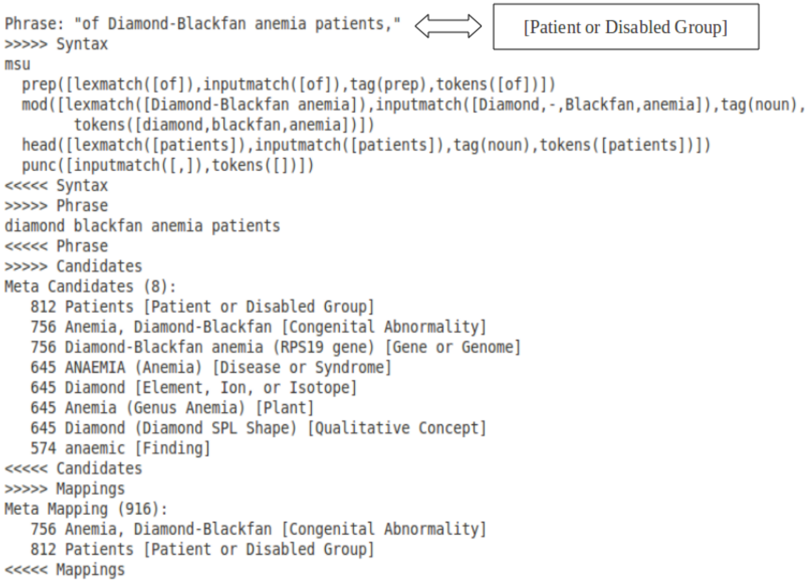
Rule 4 When the Result Merger searches for HPO phenotypes in the text, it searches for their singular and plural forms.

Rule 5 This rule is considered in Result Merger: if the head of a noun phrase is found in HPO the whole noun phrase is tagged as a phenotype.

3.2 Adding features

To implement the rule-based system completely, Rule 3 was added as three features to the CRF, in addition to the features already possessed by BANNER. Finally, some other features which seemed to be helpful were added to the system. These features were tested several times and finally the best set of features were selected. These features include:

1. Phenotype: This feature comes from the rule-based method (see Section 2.3). In this method semantic types in SG *Disorders* are categorized into two categories. This feature is a binary feature that indicates whether a noun phrase is in the *Phenotypes* category.
2. Phenotype Candidates: This feature is a binary feature that indicates whether a noun phrase is in the *Phenotype Candidates* category of SG *Disorders*.
3. Special Modifier: A binary feature that indicates whether a noun phrase is in the list of special modifiers (Burgun et al., 2009).
4. Anatomy: A binary feature which means a noun phrase is in SG *Anatomy* or not.
5. Physiology: A binary feature which means a noun phrase is in SG *Physiology* or not.
6. List Separator: We found out that usually in biomedical literature when a number of elements in a list are phenotypes, there is a good chance that the other elements are phenotypes too. The List Separator feature was added to designate the availability of list indicators (*and* or *comma*) in the sentence.
7. Semantic type: The semantic type which is assigned by MetaMap to noun phrases. This feature is null for other phrases.
8. NP: A binary feature which indicates whether a token is a part of a noun phrase.
9. POS: The part of speech of a token. This feature is available in BANNER.
10. Lemma: The lemma of each token. This is available in BANNER.
11. NPstart: A binary feature to indicate whether a token is the first token in a noun phrase.
12. NPend: A binary feature to designate whether a token is the last token in a noun phrase.
13. 2, 3 and 4-character prefixes and suffixes: This is a part of BANNER.
14. 2 and 3 character n-grams: This is a part of BANNER.



```
Phrase: "of Diamond-Blackfan anemia patients,"
>>>>> Syntax
msu
  prep([lexmatch([of]),inputmatch([of]),tag(prep),tokens([of])])
  mod([lexmatch([Diamond-Blackfan anemia]),inputmatch([Diamond,-,Blackfan,anemia]),tag(noun),
        tokens([diamond,blackfan,anemia])])
  head([lexmatch([patients]),inputmatch([patients]),tag(noun),tokens([patients])])
  punc([inputmatch([,]),tokens([])])
<<<<< Syntax
>>>>> Phrase
diamond blackfan anemia patients
<<<<< Phrase
>>>>> Candidates
Meta Candidates (8):
   812 Patients [Patient or Disabled Group]
   756 Anemia, Diamond-Blackfan [Congenital Abnormality]
   756 Diamond-Blackfan anemia (RPS19 gene) [Gene or Genome]
   645 ANAEMIA (Anemia) [Disease or Syndrome]
   645 Diamond [Element, Ion, or Isotope]
   645 Anemia (Genus Anemia) [Plant]
   645 Diamond (Diamond SPL Shape) [Qualitative Concept]
   574 anaemic [Finding]
<<<<< Candidates
>>>>> Mappings
Meta Mapping (916):
   756 Anemia, Diamond-Blackfan [Congenital Abnormality]
   812 Patients [Patient or Disabled Group]
<<<<< Mappings
```

Figure 5: Analyzer Example

Also, it should be remembered that BANNER makes a window of size 2 for each token, i.e. features of each token contain features of the 2 tokens before and the 2 tokens after it. We used the default configuration of BANNER. The NP feature comes from the MetaMap results and the POS feature is provided by BANNER. They do not align perfectly, but it is the task of the CRF to find a solution using these conflicting features.

3.3 Corpus

The most significant problem for us in using the machine learning method was the lack of an annotated specialized corpus of sufficient size. As no one before has used machine learning on phenotype name recognition, no corpus was available for us to train and test our system with. Therefore we had to make our own corpus with a sufficient number of sentences. However, making a large corpus is a very difficult and time consuming task. So we decided to use semi-supervised learning to make our corpus.

3.3.1 Collecting the papers

To find the papers which are relevant to phenotypes, we started with two available databases: PubMed (2009) and BioMedCentral (2004). All HPO phenotypes were searched for in these databases and every paper which contained at least three different phenotypes was added to our collection. In this way we found 100 papers which were used to train the system. We had another 13 papers which had been collected for the development of the rule-based method (see Section 2.3) and were annotated manually. These 13 papers were used to test the system.

3.3.2 Annotating the corpus

As it was not possible for us to annotate the 100 papers manually, we used a semi-supervised learning method starting with the information provided by HPO. First, HPO phenotypes were searched for in the set of papers and were tagged as phenotypes. These papers along with their tags made our initial training corpus. When annotating the corpus, it should be noted that the last phase of the system (the Result Merger) was omitted because all HPO-annotated phenotypes were already annotated.

The trained model was used to annotate the training set again. The newly found phenotypes were analysed manually and the correct ones were added as annotations to the training set. Also, on each iteration, the system was tested using the test set to find out how many iterations would be sufficient for training the system. This process was repeated several times until we reached the results that we were satisfied with when testing the last model on the test set.

One important point to mention is that we only included positive sentences (sentences with at least one phenotype) in our training set, because the number of negative sentences was far greater than positive sentences and it prevented the system from training efficiently. Therefore, whenever new phenotype names were found, the number of sentences in the training set increased for the next iteration.

4 Evaluation

We compared the performance of our system against the rule-based system (Khordad et al., 2011), which is the only specialized system for phenotype name recognition that we are aware of. The final training corpus which is made from 100 papers contains 2755 sentences and 4233 annotated phenotypes. All sentences in this corpus include at least one phenotype. A test set is collected from 13 papers and includes 216 sentences and 373 phenotypes.

To evaluate the system, 10-fold cross validation has been used. Also the system has been tested using a separate test set. Table 2 gives the details of the results. The base system is our machine learning method ignoring the Result Merger. Result Merger finds HPO phenotypes in text and adds them to the list of phenotypes. The results after using Result Merger are mentioned in the column labelled *HPO added*. The rule-based system has been tested using the test set and the results are displayed in Table 3. To calculate these results, a returned phrase is considered to be a true phenotype if its head contains a phenotype. For example, in the phrase "acute myloid leukemia" the head is "leukemia", a phenotype that is confirmed by its inclusion in HPO. However, in the phrase"Diamond-Blackfan anemia patients" the correct phenotype is: "Diamond-Blackfan anemia". If the system returns "Diamond-Blackfan anemia patients" as the phenotype, it is deemed false.

As this table demonstrates, the results are comparable with other named entity recognition systems which are specialized for finding other biomedical entities even though they may have larger training corpora. For example BANNER has been trained and tested for finding gene names, using BioCreative 2 GM corpus containing 15,000 sentences (7500 for training and 7500 for testing) which is much larger than our current corpus. However our results are really better than Banner's (Precision 85.09, Recall 79.06 and F-measure 81.96)(Leaman and Gonzalez, 2008). Although our task is different these results mean that our system is performing well.

In addition the machine learning method outperformed the rule-based method, even though the corpus is not fully annotated. We believe that if we had a fully annotated corpus the system would achieve an even better performance.

	Base system			HPO Added		
	Precision	Recall	F-measure	Precision	Recall	F-measure
10-Fold	82.83	68.13	74.35	86.89	98.33	92.25
Test Set	93.44	57.37	71.09	95.76	90.88	93.25

Table 2: System Evaluation Results

	Precision	Recall	F-measure
Machine Learning Method	95.76	90.88	93.25
Rule-Based Method	88.34	73.19	80.05

Table 3: Comparing the system with the Rule-Based Method

To have an idea of how well our annotation process works, we selected 100 random sentences from our corpus. Then, we tagged these sentences manually. There were 157 phenotypes in these 100 sentences. Our semi-supervised machine learning method found 142 phenotypes and 1 of its phenotypes was incorrect, i.e. it missed 16 phenotypes. So, the annotation process misses about 10 percent of the phenotypes.

5 Discussion

Finding the best set of features is one of the most important parts of developing the system. Tables 4 and 5 show the role and importance of each feature. To illustrate the contribution of each feature in Table 4 we considered a small set of features as the basic set of features for our system. These features came from the rule-based system and are very important in signifying phenotype names in the text. Then, in each line a feature is added to the basic feature set until we have the complete feature set in the last line. Adding some features (Phenotype candidates, List separators, Semantic Types, and Lemma) drops the results slightly but the results of the last feature set is better than the previous feature sets.

Analysing Table 4, it may seem that including some features is not necessary. Table 5 illustrates the role of each feature in a different way. In each line of Table 5 only one feature is ignored from the feature set and the system is tested using the separate test set. Note that the results are calculated without adding the HPO. As one can see, removing each feature reduces the results slightly. The exception to this modest reduction in performance is the removal of the NP feature which causes a significant drop in precision and recall, because not having NP information causes errors in finding NP end and NP start. In some cases (Lemma, Phenotype candidates, NP start and NP end, and Physiology) ignoring the feature causes small improvement in precision or recall. However the F-score is always less than the F-score of final results.

Reviewing the results, it has been found that both the rule-based and machine learning methods are dependent on MetaMap. MetaMap does make mistakes. MetaMap makes some errors in finding the boundaries of NPs and in determining the semantic types. NP boundaries and semantic types are features used by both methodologies, so the errors made by MetaMap have effects on the performance of each system. However the rule-based system is more dependent on MetaMap output and errors in MetaMap output changes the results completely. But the machine learning system is more robust and it sometimes finds the correct phenotype names despite Metamap errors. For example consider the sentence *"Diamond-Blackfan anemia is a rare inherited bone marrow failure syndrome."*. The phrase *Diamond-Blackfan anemia* is a phenotype but MetaMap assigns the [Gene or Genome] semantic type to it, which is not in the SG *Disorders*.

Features	Precision	Recall	F-Score
Phenotype, Anatomy, Physiology, Special Modifier	89.89	47.72	62.34
+Phenotype candidates	90.72	47.18	62.07
+List Separator	90.41	40.48	55.92
+Semantic type	90.24	49.59	64
+NP	90.24	49.59	64
+POS	91.15	55.22	68.77
+Lemma	92.72	54.69	68.79
+NP start, NP end	93.44	57.37	71.09

Table 4: Contribution of each additional feature

Ignored Feature	Precision	Recall	F-Score
Anatomy	92.82	55.49	69.45
Lemma	93.57	54.69	69.03
List Separator	91.66	56.03	69.54
Phenotype	92.54	56.56	70.20
NP	81.81	38.6	52.45
Phenotype Candidate	92.64	57.37	70.85
NP start and NP end	93.18	54.95	69.13
Physiology	93.21	55.22	69.35
POS	92.05	52.81	67.11
Semantic Type	91.89	54.69	68.56
Special Modifier	92.44	55.76	69.56

Table 5: Contribution of each feature

So the rule-based system does not tag it as a phenotype. The phrase *missing vertebrae* is another example of MetaMap errors. MetaMap does not consider this phenotype name as one NP. It separates this phenotype name into two phrases *missing* and *vertebrae*.

In addition determining the boundary of an NP is very important in the rule-based system. MetaMap has an option to make larger NPs by merging simpler NPs. If we only use simple NPs, we cannot get larger phenotype names as one NP and the system will miss them. The phrase *Partial hypoplasia of the corpus callosum* is an example of a phenotype name with composite NPs. On the other hand if we use composite NPs, the head of the NP may change and the semantic type of the NP may change as a result. This is problematic for the rule-based system. The phrase *the associations of facial dysmorphism* is an example. The word “associations” is the head of this phrase, so the semantic type [Mental Process] is assigned to it and it is not tagged as a phenotype name by the rule-based system.

On the other hand, there are some cases in which MetaMap assigns the correct semantic type to a phrase and found a good boundary for a phenotype name but the machine learning method does not mark it as a phenotype. For example “arhinia” is not tagged as a phenotype by the machine learning method in the following sentence *“These phenotypes may resemble that of the only confirmed case of an individual with a lethal compound heterozygous PAX6 mutation and may include anophthalmia, arhinia and severe central nervous system defects”* although MetaMap assigns the semantic type [Congenital Abnormality] (which is in the category of Phenotypes) to it.

Table 6 shows how many false negatives and true positives are available in the final results for both the machine learning and the rule-based methods. And Table 7 illustrates the percent of these errors caused by MetaMap or the boundary of noun phrases.

	Rule-based	Machine Learning
True positives	273	339
False negatives	100	34

Table 6: number of TPs and FNs in each method

	Rule-based	Machine Learning
MetaMap errors	37%	11.76%
NP boundary errors	26%	8.82%

Table 7: Analysis of NPs

Comparing the precision errors gave interesting observations. There are no common errors between the two systems. False positive errors in the rule-based system were caused by the rules not being discriminating enough. For each returned phrase, one of the rules produced that phrase. But there are exceptions to each rule that can cause these false positive errors. The exceptions to the rules do not cause any problem for the machine learning system. The machine learning system was able to learn all of these exceptions. For the false phenotypes returned by the machine learning system, analysis indicates that none of these would be suggested by application of a rule in the rule-based system, explainging why there are no common errors.

Conclusion

In this paper we discussed the development of a named entity recognition system which is specialized to find phenotype names in biomedical literature. The system has been generated using machine learning. Some of its features are based on the rules found in our previous rule-based method (Khordad et al., 2011). We added some other features. The system makes use of MetaMap and HPO.

As there was no annotated corpus available for training the machine learning system, a corpus was annotated using a semi-supervised learning method and HPO. The corpus does not fully annotate all phenotype names. About 10 percent of the phenotype names are missed.

The system has been evaluated using both a 10-fold cross validation and a separate test set and the results are really promising.

The current system is extremely dependent on MetaMap output, although it is less dependent than the rule-based system. Still, when MetaMap gives erroneous output, it makes it difficult for our system to work correctly. In addition the corpus is not completely annotated and its annotation has some errors.

Despite these problems, the system achieved an F-Score of 92.25 and its performance is comparable to other NER systems which are specialized for finding other entities in biomedical literature.

To improve the system performance, it is imperative to overcome the phenotype name recognition errors initiated by MetaMap parsing errors. Using a more reliable partial parser to provide the NPs for MetaMap's mapping to UMLS concepts, instead of using the Metamap embedded parser, may fix this problem. Furthermore, adding more features to the machine learning system can be considered. Finally, a machine learning method other than CRF might achieve better results.

Acknowledgments

This work was partially funded through a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant to Robert E. Mercer.

References

Andrade, M. A. and Valencia, A. (1998). Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics*, 14(7):600–607.

Aronson, A. R. (2001). Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *AMIA Symposium*, pages 17–21.

Bui, Q.-C., Katrenko, S., and Sloot, P. M. A. (2011). A hybrid approach to extract protein-protein interactions. *Bioinformatics*, 27(2):259–265.

Burgun, A., Mougin, F., and Bodenreider, O. (2009). Two approaches to integrating phenotype and clinical information. *AMIA Annual Symposium proceedings*, 2009:75–79.

Coulet, A., Shah, N. H., Garten, Y., Musen, M. A., and Altman, R. B. (2010). Using text to build semantic networks for pharmacogenomics. *Journal of Biomedical Informatics*, 43(6):1009–1019.

Dai, M., Shah, N. H., Xuan, W., Musen, M. A., Watson, S. J., Athey, B. D., and Meng, F. (2008). An efficient solution for mapping free text to ontology terms. In *AMIA Summit on Translational Bioinformatics, San Francisco, CA*.

Day-Richter, J., Harris, M. A., Haendel, M., OBO, T. G. O., and Lewis, S. (2007). OBO-Edit an ontology editor for biologists. *Bioinformatics*, 23(16):2198–2200.

Friedman, C., Kra, P., Yu, H., Krauthammer, M., and Rzhetsky, A. (2001). GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics (Oxford, England)*, 17 Suppl 1(suppl_1):S74–82.

Fukuda, K., Tamura, A., Tsunoda, T., and Takagi, T. (1998). Toward information extraction: identifying protein names from biological papers. *Pac Symp Biocomput*, pages 707–718.

Fundel, K., Küffner, R., and Zimmer, R. (2007). Relex - relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371.

Gaizauskas, R., Demetriou, G., Artymiuk, P. J., and Willett, P. (2003). Protein structures and information extraction from biological texts: the PASTA system. *Bioinformatics*, 19(1):135–143.

Gong, L.-J., Yuan, Y., Wei, Y.-B., and Sun, X. (2009). A hybrid approach for biomedical entity name recognition. In *Biomedical Engineering and Informatics, 2009. BMEI '09. 2nd International Conference on*, pages 1 –5.

He, X. and DiMarco, C. (2005). Using lexical chaining to rank protein-protein interactions in biomedical texts. In *BioLink 2005: Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics, Conference of the Association for Computational Linguistics (poster Presentation)*.

Horn, F., Lau, A. L., and Cohen, F. E. (2004). Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. *Bioinformatics*, 20(4):557–568.

Hristovski, D., Friedman, C., Rindflesch, T. C., and Peterlin, B. (2006). Exploiting semantic relations for literature-based discovery. *AMIA Annual Symposium proceedings*, pages 349–353.

Hristovski, D., Peterlin, B., Mitchell, J. A., and Humphrey, S. M. (2005). Using literature-based discovery to identify disease candidate genes. *I. J. Medical Informatics*, 74(2-4):289–298.

Humphreys, B. L., Lindberg, D. A., Schoolman, H. M., and Barnett, G. O. (1998). The Unified Medical Language System: an informatics research collaboration. *J Am Med Inform Assoc*, 5(1):1–11.

Humphreys, K., Demetriou, G., and Gaizauskas, R. (2000). Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 505–516.

Khordad, M., Mercer, R. E., and Rogan, P. (2011). Improving phenotype name recognition. *Canadian Conference on AI*, 6657:246–257.

kiong Ng, S. and Wong, M. (1999). Toward routine automatic pathway discovery from on-line scientific text abstracts. *Genome Informatics*, 10:104–112.

Klein, T. E., Chang, J. T., Cho, M. K., Easton, K. L., Fergerson, R., Hewett, M., Lin, Z., Liu, Y., Liu, S., Oliver, D. E., Rubin, D. L., Shafa, F., Stuart, J. M., and Altman, R. B. (2001). Integrating genotype and phenotype information: an overview of the PharmGKB project. *The pharmacogenomics journal*, 1(3):167–170.

Krauthammer, M., Rzhetsky, A., Morozov, P., and Friedman, C. (2000). Using BLAST for identifying gene and protein names in journal articles. *Gene*, 259(1-2):245–252.

Leaman, R. and Gonzalez, G. (2008). Banner: An executable survey of advances in biomedical named entity recognition. In *Pacific Symposium on Biocomputing*, pages 652–663.

Leroy, G., Chen, H., and Martinez, J. D. (2003). A shallow parser based on closed-class words to capture relations in biomedical text. *Journal of Biomedical Informatics*, 36(3):145–158.

Leser, U. and Hakenberg, J. (2005). What makes a gene name? named entity recognition in the biomedical literature. *Briefings in Bioinformatics*, 6(4):357–369.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

McCray, A., Burgun, A., and Bodenreider, O. (2001). Aggregating UMLS semantic types for reducing conceptual complexity. *Proceedings of Medinfo*, 10(pt 1):216–20.

McKusick, V. (2007). Mendelian Inheritance in Man and Its Online Version, OMIM. *The American Journal of Human Genetics*, 80(4):588–604.

Nobata, C., Collier, N., and ichi Tsujii, J. (1999). Automatic term identification and classification in biology texts. In *In Proc. of the 5th NLPRS*, pages 369–374.

Rindflesch, T. C., Libbus, B., Hristovski, D., Aronson, A. R., and Kilicoglu, H. (2003). Semantic relations asserting the etiology of genetic diseases. *AMIA Annual Symposium Proceedings*, pages 554–558.

Rindflesch, T. C., Tanabe, L., Weinstein, J. N., and Hunter, L. (2000). Edgar: Extraction of drugs, genes and relations from the biomedical literature. In *Pacific Symposium on Biocomputing*, volume 5, pages 514–525.

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

Rocktäschel, T., Weidlich, M., and Leser, U. (2012). Chemspot: a hybrid system for chemical named entity recognition. *Bioinformatics*, 28(12):1633–1640.

Schwartz, A. S. and Hearst, M. A. (2003). A simple algorithm for identifying abbreviation definitions in biomedical text. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 451–462.

Segura-Bedmar, I., Martinez, P., and Segura-Bedmar, M. (2008). Drug name recognition and classification in biomedical texts: A case study outlining approaches underpinning automated systems. *Drug Discovery Today*, 13(17-18):816 – 823.

Seki, K. and Mostafa, J. (2005). A hybrid approach to protein name identification in biomedical texts. *Inf. Process. Manage.*, 41(4):723–743.

Swanson, D. (1986). Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspect. Bio. Med*, 30:7–18.

Tanabe, L., Scherf, U., Smith, L. H., Lee, J. K., Hunter, L., and Weinstein, J. N. (1999). MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *BioTechniques*, 27(6):1210–1217.

Valencia, A. (2005). Automatic annotation of protein function. *Current opinion in structural biology*, 15(3):267–274.

Xu, R., Supekar, K., Morgan, A., Das, A., and Garber, A. (2008). Unsupervised method for automatic construction of a disease dictionary from a large free text collection. *AMIA Annual Symposium proceedings*, pages 820–824.

Yu, H., Zhu, X., Huang, M., and Li, M. (2005). Discovering patterns to extract protein-protein interactions from the literature: Part ii. *Bioinformatics*, 21(15):3294–3300.