OPEN ACCESS

Using rule-based natural language processing to improve disease normalization in biomedical text

Ning Kang, Bharat Singh, Zubair Afzal, Erik M van Mulligen, Jan A Kors

▸ Additional data is published online only. To view this file please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2012-001173)

Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands

Correspondence to
Ning Kang, Department of Medical Informatics, Erasmus University Medical Center, PO Box 2040, Rotterdam 3000 CA, The Netherlands; n.kang@erasmusmc.nl

Received 20 June 2012
Accepted 14 September 2012
Published Online First
6 October 2012

To cite: Kang N, Singh B, Afzal Z, *et al*. *J Am Med Inform Assoc* 2013;**20**:876–881.

ABSTRACT

Background and objective In order for computers to extract useful information from unstructured text, a concept normalization system is needed to link relevant concepts in a text to sources that contain further information about the concept. Popular concept normalization tools in the biomedical field are dictionary-based. In this study we investigate the usefulness of natural language processing (NLP) as an adjunct to dictionary-based concept normalization.

Methods We compared the performance of two biomedical concept normalization systems, MetaMap and Peregrine, on the Arizona Disease Corpus, with and without the use of a rule-based NLP module. Performance was assessed for exact and inexact boundary matching of the system annotations with those of the gold standard and for concept identifier matching.

Results Without the NLP module, MetaMap and Peregrine attained F-scores of 61.0% and 63.9%, respectively, for exact boundary matching, and 55.1% and 56.9% for concept identifier matching. With the aid of the NLP module, the F-scores of MetaMap and Peregrine improved to 73.3% and 78.0% for boundary matching, and to 66.2% and 69.8% for concept identifier matching. For inexact boundary matching, performances further increased to 85.5% and 85.4%, and to 73.6% and 73.3% for concept identifier matching.

Conclusions We have shown the added value of NLP for the recognition and normalization of diseases with MetaMap and Peregrine. The NLP module is general and can be applied in combination with any concept normalization system. Whether its use for concept types other than disease is equally advantageous remains to be investigated.

INTRODUCTION

Most biomedical knowledge comes only in unstructured form, such as in scientific articles and reports. The sheer volume of these textual sources requires computer processing to extract usable information. An important step in the information extraction task is the recognition and normalization of relevant concepts in a text.[1] Concept or named-entity recognition aims at finding text strings that refer to entities, and marking each entity with a semantic type, like ‘gene’, ‘drug’, or ‘disease’. Concept normalization goes beyond entity recognition. It assigns a unique identifier to the recognized concept, which links it to a source that contains further information about the concept, such as its definition, its preferred name and synonyms, and its relationships with other concepts.

Much research has been done in concept recognition, but fewer studies addressed the more difficult task of concept normalization. Concept normalization systems are often dictionary-based, that is, they try to find concept occurrences in a text by matching text strings with concept names and their corresponding identifiers in a dictionary. The dictionary is composed of entries from one or more knowledge sources, such as Gene Ontology,[2] Entrez Gene,[3] or the Unified Medical Language System (UMLS).[4] Typically, dictionary-based systems use little or no linguistic information to find concepts, and the potential added value of such information is largely unknown.

In this study, we investigate the usefulness of natural language processing (NLP) techniques to improve biomedical concept normalization. We present a set of rules that utilize NLP information, and show that these rules substantially improve the performance of two concept normalization systems, MetaMap[5] and Peregrine,[6] in recognizing and normalizing diseases in biomedical text.

BACKGROUND

Compared to the number of named-entity recognition systems, the number of concept normalization systems in the biomedical field is small. Reported systems include MetaMap,[5] Mgrep,[7] Negfinder,[8] Peregrine,[6] and Whatizit.[9] All of these systems use a dictionary to find concepts in text and map them to concept identifiers. Several systems, such as MetaMap, perform some lexical analysis in the normalization process, but part-of-speech (POS) and chunking information are mostly not considered.

Concept normalization is generally considered a more difficult task than concept recognition. This is reflected in the variety of named-entity recognition challenges in the biomedical domain, for example, BioCreative[10] and BioNLP[11] (recognition of proteins and genes in scientific literature), and TREC[12] and i2b2[13] (drugs, diseases, and treatments in electronic patient records), whereas normalization challenges have been few. Substantial work on gene normalization has been done in a series of gene normalization tasks that were part of the BioCreative competitions.[14–16] In BioCreative I and II, the gene normalization task consisted of finding the identifiers of genes and gene products mentioned in sets of abstracts from four model organisms: yeast, fly, mouse (BioCreative I), and human (BioCreative II). In the gene normalization task in BioCreative III, systems had to assign identifiers to all named genes in full-text articles without being limited to a particular organism. For all tasks, unique gene identifiers had to be provided at the document level, rather than for individual gene mentions. The different systems participating in these challenges used a wide variety of methods, including pattern matching, machine learning, and lexical resources lookup.[14–16] Heuristic rules were mostly developed and implemented in an ad-hoc and custom manner. Hybrid use of rule-based and machine learning methods was observed in system descriptions.[15 16] For example, GNAT,[17] the top-performing gene normalization system in BioCreative II, combines dictionary matching and machine learning to recognize gene mentions; the machine-learning component of such programs requires a subsequent step to match predicted mentions to identifiers. In the next steps, recognized gene names are validated by means of several dedicated filters to remove false positives, while ambiguous mentions are disambiguated by comparing the current text with several sources of background information existing for each candidate gene. Another example of a high-performance gene normalization system is GeNo.[18] This system also combines dictionary-based and machine-learning based gene name detection, using approximate string matching to link gene mentions with dictionary identifiers. It employs automatic term variant generation and false positive filtering. The system fully relies on publicly available data and resources. GeNo did not participate in BioCreative, but was shown to have a performance on par with GNAT. Another top-ranking system in BioCreative I and II was the proprietary ProMiner gene normalization system.[19] ProMiner employs a strictly dictionary-based approach, relying on well-curated dictionaries and approximate string matching.

Species name recognition is an important subtask in many systems participating in the BioCreative III challenge. Some of the systems employed LINNAEUS, an open source species normalization system.[20] LINNAEUS follows a dictionary-based approach, using a time-efficient implementation of regular expressions for document tagging. Post-processing includes acronym detection, filtering of common words, and disambiguation.

Many algorithms and methods have been proposed to solve common problems encountered in concept recognition and concept normalization tasks.[21] For instance, to solve the problem of gene mention coordination, multiple conditional random fields (CRFs) and n-gram language models were used.[22] CRFs were also used in another study to decompose complex coordinated entity expressions into constituent conjuncts, to determine the missing elements, and thus to reconstruct explicitly all the single named entity mentions.[23] Rule-based procedures for resolving simple conjunctions of gene mentions have also been used.[17 24 25] Many papers addressed the problem of abbreviation detection and expansion.[26–29] Proposed approaches range from simple rule-based algorithms to sophisticated machine-learning methods. Term variation is another common problem that concept normalization systems have to deal with. Methods that have been proposed include approximate string matching, heuristic pattern matching rules, enhanced dictionaries, etc.[30–32] Finally, a common problem addressed in many systems is the removal of false-positive mentions that result from the recognition stage. One often used approach is to filter out terms that have an ambiguous meaning in common English,[33] but more sophisticated methods (eg, scoring the similarity between the semantic profile of a concept and the document in which it occurs[18]) have also been proposed.

There are only a few corpora in the biomedical domain that incorporate concept annotations, notably the Arizona Disease Corpus (AZDC),[34] the BioCreative gene normalization corpora,[14–16] the Colorado Richly Annotated Full-Text Corpus,[35] and the Gene Regulation Event Corpus.[36] Among these, AZDC is the only one that includes information about concept boundaries and UMLS concept identifiers, and that is publicly available. Based on the AZDC corpus, very recently the larger NCBI (National Center for Biotechnology Information) corpus was developed,[37] but this corpus only contains annotations of disease mentions, not concept identifiers. Therefore, we used AZDC as the gold standard corpus (GSC) for our experiments.

The AZDC was used before by Leaman *et al*[34] to test the performance of one dictionary-based system and two statistical systems (BANNER[34] and JNET[38]). The dictionary-based system yielded an F-score of 62.2%, while BANNER and JNET achieved F-scores of 77.9% and 77.2%, respectively. Chowdhury and Faisal[39] developed another machine-learning based system, BNER, and tested it on the same corpus, achieving an F-score of 81.1%. Both these studies were targeted at concept recognition, not at concept normalization.

METHODS

Corpus

The AZDC has been developed at the Arizona State University. It was released in 2009.[34] The corpus has been annotated with disease concepts, including UMLS codes, preferred concept names, and start and end points of disease mentions inside the sentences. The whole corpus consists of 2784 sentences, taken from 793 Medline abstracts, and 3455 disease annotations. Annotations have been mapped to a concept unique identifier (CUI) in the UMLS metathesaurus. Each annotation belongs to one of the following semantic types defined in the UMLS: disease or syndrome, neoplastic process, congenital abnormality, acquired abnormality, experimental model of disease, injury or poisoning, mental or behavioral dysfunction, pathological function, sign or symptom.

We divided the corpus into two parts: one-third of the sentences were used for developing the NLP module, the other two-thirds for testing.

Concept normalization systems

We evaluated two concept normalization systems, MetaMap and Peregrine. Both systems were downloaded from their official websites with default configurations and parameters, and no attempt was made to optimize their performance.

MetaMap (http://metamap.nlm.nih.gov/) is a dictionary-based system for normalizing concepts from the UMLS metathesaurus in biomedical texts.[5] It makes use of a minimal-commitment parser, which splits texts into chunks in which concepts are identified. MetaMap also performs word-sense disambiguation (WSD). MetaMap is dictionary-based and cannot be trained. MetaMap Transfer (MMTx) is a distributable version of MetaMap written in Java. We used the 2011 version, which includes V.2011AA of the UMLS metathesaurus.

Peregrine (https://trac.nbic.nl/data-mining/) is a dictionary-based concept recognition and normalization tool, developed at the Erasmus University Medical Center (http://www.biosemantics.org). Peregrine finds concepts by dictionary look-up, and performs WSD.[6] Rewrite and suppression rules are applied to the terms in the dictionary to enhance precision and recall.[40] In our experiments, we used Peregrine with V.2011AB of the UMLS metathesaurus.

NLP module

The NLP module that we have developed consists of a number of rules that combine the annotations of a concept normalization system with POS and chunking information. We used the OpenNLP tool suit (http://opennlp.apache.org/) to obtain the necessary POS and chunking information. The OpenNLP tool suit is based on a maximum entropy model. An OpenNLP Unstructured Information Management Architecture (UIMA) wrapper has been developed by JULIE Lab (http://www.julielab.de). The wrapper divides the OpenNLP package into small modules that perform sentence detection, tokenization, POS tagging, and chunking, which makes it easy to configure the pipeline for different purposes.[41] The rules in the NLP module are divided into five submodules, which address specific tasks and are described in the following. A detailed description of the rules is available as online supplementary material.

1. *Coordination*. This submodule performs coordination resolution. The approach is straightforward and extends the one described by Baumgartner *et al*.[24] For instance, in the following sentence from the AZDC: 'We calculated age related risks of all, colorectal, endometrial, and ovarian cancers in nt943+3 A–T MSH2 mutation carriers (…)', MetaMap and Peregrine both recognize 'ovarian cancers' as a concept, but miss 'colorectal cancers' and 'endometrial cancers'. Using POS and chunking information, this module reformats the coordination phrase and feeds the reformatted text into the concept normalization systems for proper annotation of the concepts.
2. *Abbreviation*. This submodule combines the abbreviation expansion algorithm of Schwartz and Hearst,[26] with POS and chunking information to improve the recognition of abbreviations. We chose the Schwartz and Hearst algorithm because it is very easy to implement and to combine with other rules, and has shown consistently good performance in different studies.[24 25] For an instance of abbreviation errors, in the sentence 'Deficiency of aspartylglucosaminidase AGA causes a lysosomal storage disorder Aspartylglucosaminuria AGU', the concept normalization systems annotated 'Deficiency of aspartylglucosaminidase' and 'Aspartylglucosaminuria' as disease concepts, in agreement with the gold standard annotations, but they did not recognize the abbreviations. Since 'AGA' is used as the abbreviation of 'aspartylglucosaminidase', an enzyme, it should not be annotated as a disease concept, but 'AGU' should be identified as such. This was accomplished by means of a rule that checks whether the last noun in a noun phrase is an abbreviation of all preceding tokens in the noun phrase.
3. *Term variation*. Dictionary-based systems can only find concepts if the terms by which these concepts are denoted in text are part of the dictionary. Although UMLS covers some term variation, many variations are missing. The submodule in question uses a shallow parsing based approach, similar to Ferrucci *et al*.[42] It contains a number of rules that adjust noun phrases and feed the adjusted phrase into the concept normalization system again, to check whether it refers to a concept. For instance, if a noun phrase includes a preposition, such as 'deficiency of hex A', which is not part of the UMLS, the word order is changed into 'hex A deficiency', which is contained in the UMLS.
4. *Boundary correction*. This submodule contains several rules that correct the start and end positions of concepts identified by the systems, based on POS and chunking information. For instance, if the POS of the start or end token from a concept annotation is a verb, preposition, conjunction, or interjection, it then uses POS information to adjust the concept start or end position. By applying these rules, an erroneous annotation such as 'phenylketonuria Is', contrived from 'classical phenylketonuria is an autosomal recessive disease', could be corrected to 'phenylketonuria'.
5. *Filtering*. This submodule has two rules that suppress concepts although they had been identified by the system. The first rule removes a concept if the concept annotation in the text has no overlap with a noun phrase because in our experience, most UMLS concepts in biomedical abstracts belong to a noun phrase, or at least overlap with it. The second rule removes a concept if it is part of a concept filter list, a common approach to increase precision as used by many systems in the BioCreative competitions.[14–16] Our list contains 23 generic concepts (eg, 'disease', 'abnormality') that were wrongly annotated by Peregrine in the training set.

The rules in the NLP module were developed on the basis of an error analysis of the Peregrine annotations of the training set. The annotations of MetaMap were not used for this development.

Performance evaluation

The annotations of the concept normalization systems were compared with the gold standard annotations by exact and inexact matching, both of the concept boundaries (following the same procedure as in Leaman *et al*[34] and Chowdhury and Faisal[39]) and of the concept identifiers. For exact boundary matching, an annotation was counted as true positive if it was identical to the gold standard annotation, that is, if both annotations had the same start and end location in the corpus. If a gold-standard annotation was not given, or not rendered exactly by the system, it was counted as false negative; if an annotation found by the system did not exactly match the gold standard, it was counted as false positive. For concept identifier matching, the same rules applied as for exact boundary matching, with the additional requirement that for a true positive outcome the concept identifiers had to match; if not, the annotations were counted as false positive as well as false negative. Performance was evaluated in terms of precision, recall, and F-score.

The performance of the systems was also tested by using two methods of inexact boundary matching: one-side boundary matching (ie, at least one boundary of the system annotation had to match the gold standard annotation) and overlap matching (ie, at least one word of the system annotation had to overlap with the corpus annotation). Inexact concept identifier matching followed the same rules as inexact boundary matching, but in addition required the concept identifiers to match.

An error analysis was carried out on a sample of 100 randomly selected errors that were made by each concept normalization system after applying the NLP module. Errors were grouped into five categories, following the task categorization of the five NLP submodules.

Processing pipeline

All systems and the NLP module were integrated in the UIMA framework,[43] which was easily accomplished since they either were available as a UIMA component[41] or had a web service interface. A UIMA processing pipeline was implemented, which first read the AZDC test set by the UIMA Collection Reader. Then the test set was annotated by MetaMap and Peregrine. Because the AZDC includes nine UMLS semantic types, only the annotated concepts belonging to these types were considered for evaluation. Subsequently, the NLP submodules post-processed the annotation results. Some rules, such as the coordination rules, not only processed the annotations but also modified the original input sentence. The modified text was then fed into the concept normalization systems for re-annotation.

Table 1 Performance (in %) of MetaMap and Peregrine, with and without the NLP module, on the AZDC test set for exact and inexact matching of concept boundaries and identifiers

	Exact matching						Inexact matching					
	Boundaries			Identifiers			Boundaries			Identifiers		
System	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
MetaMap	60.9	61.1	61.0	55.0	55.2	55.1	79.3	79.7	79.5	65.1	65.5	65.3
MetaMap+NLP	76.1	70.7	73.3	68.7	63.9	66.2	89.1	82.2	85.5	76.1	71.3	73.6
Peregrine	63.5	64.3	63.9	56.6	57.3	56.9	77.1	78.4	77.7	64.6	65.3	64.9
Peregrine+NLP	82.2	74.2	78.0	73.5	66.4	69.8	89.6	81.6	85.4	76.7	70.2	73.3

AZDC, Arizona Disease Corpus; NLP, natural language processing.

Finally, the system annotations before and after post-processing by the NLP submodules were evaluated separately against the gold standard annotations.

RESULTS

Performance of the concept normalization systems without and with NLP

shows the Tables 1 performance of the two concept normalization systems on the AZDC test set. Without the NLP module, MetaMap achieved an F-score of 61.0% for exact boundary matching, and 55.1% for concept identifier matching. The F-scores of Peregrine were 63.9% for exact boundary matching and 56.9% for concept identifier matching. With the aid of the NLP module, the F-scores of MetaMap and Peregrine increased by 12.3 and 14.1 percentage points, respectively, for exact boundary matching, and by 11.1 and 12.9 percentage points for concept identifier matching. For both MetaMap and Peregrine, there is a larger increase in precision than in recall (table 1).

For inexact, one-side boundary matching, MetaMap and Peregrine, without the NLP module, reached F-scores of 79.5% and 77.7%, respectively (table 1). With the NLP module, these F-scores increased to 85.5% and 85.4%. The performances for concept identifier matching increased from 65.3% to 73.6% and from 64.9% to 73.3%, respectively. The F-scores for overlap matching showed only minimal improvement (<1 percentage point, data not shown).

Performance of NLP submodules

Tables 2 shows the incremental performance improvement for the various NLP submodules, based on exact boundary matching. The baseline was the performance of MetaMap and Peregrine without any submodules. The coordination module, abbreviation module, and boundary correction module each contributed 3.0–3.5 percentage points to the betterment of performance. The smallest contribution to raising the F-score was by the filtering module, with the two rules in this module equally contributing to the performance improvement.

Table 2 Performance (in %) of MetaMap and Peregrine with incremental contributions of the NLP submodules on the AZDC test set for exact boundary matching

NLP submodules	MetaMap			Peregrine		
	Precision	Recall	F-score	Precision	Recall	F-score
Baseline	60.9	61.1	61.0	63.5	64.3	63.9
+Coordination	63.4	64.4	64.0	67.5	67.0	67.2
+Abbreviation	66.7	67.7	67.2	70.7	70.8	70.7
+Term variation	68.9	69.3	69.1	74.3	71.8	73.0
+Boundary correction	73.7	70.7	72.2	78.8	74.2	76.4
+Filtering	76.1	70.7	73.3	82.2	74.2	78.0

AZDC, Arizona Disease Corpus; NLP, natural language processing.

Error analysis

We randomly selected 100 errors that MetaMap and Peregrine, together with the NLP module, each made on the test set, and manually classified them into different error types (Tables 3). The error profiles of MetaMap and Peregrine were very similar. The majority of errors were due to term variation and boundary errors. For instance, the term ‘α-Gal A deficiency’, referring to the concept ‘α-galactosidase deficiency’, was not found as the term does not occur in UMLS. Boundary errors mainly occurred because of nested annotations in the gold standard. For example, in ‘Therefore, we screened eight familial gastric cancer kindreds of British and Irish origin (...)’, both ‘familial gastric cancer’ and ‘gastric cancer’ were annotated in the gold standard, whereas the systems only annotated ‘gastric cancer’. Filtering errors were mainly due to concepts that were inconsistently annotated by the gold standard. For example, in ‘Because of the variable expression of nm23-H1 in different tumors (...)’, ‘tumors’ was annotated by the gold standard, but it was not in ‘In neuroblastoma, higher levels of p19/nm23 (...) were observed in advanced stage tumors compared with limited stage disease’. Coordination and abbreviation errors were relatively few. For example, in the sentence ‘the majority of familial breast/ovarian cancer’, the gold standard annotated ‘familial breast cancer’ and ‘familial ovarian cancer’, whereas the coordination module recognized ‘breast cancer’ and ‘ovarian cancer’ but failed to include ‘familial’. An abbreviation error occurred in the sentence ‘One of five PWS/AS patients analyzed to date has an identifiable, rearranged HERC2 transcript derived from the deletion event.’ The systems did not annotate ‘PWS’ (Prader–Willi syndrome) and ‘AS’ (Angelman syndrome), which

Table 3 Distribution across five error types of 100 randomly selected errors of each system on the AZDC test set

System	Coordination	Abbreviation	Term variation	Boundary	Filtering
MetaMap +NLP	12	13	28	24	23
Peregrine +NLP	11	14	28	26	21

AZDC, Arizona Disease Corpus; NLP, natural language processing.

had been defined (and correctly annotated) in a preceding sentence.

DISCUSSION

We have investigated the use of NLP to improve the performance of two concept normalization systems. By applying a set of post-processing rules that utilize POS and chunking information, the F-scores of MetaMap and Peregrine on AZDC improved by 12.3 and 14.1 percentage points, respectively, for exact boundary matching, and by 11.1 and 12.9 percentage points, respectively, for concept identifier matching. For inexact matching, the improvement was smaller but still in the order of 6–8 percentage points. To our knowledge, this is the first study that assesses the performance of systems in normalizing disease concepts.

Concept recognition performed substantially better than concept normalization, even if the boundaries matched exactly. This may partially be explained by the fact that the systems assigned the wrong CUI to ambiguous terms. However, on closer inspection it turned out that the gold-standard annotators often took into account the context in which a term was used and assigned a more specific CUI than the systems. For instance, in the sentence 'A DNA-based test for the HFE gene is commercially available, but its place in the diagnosis of hemochromatosis is still being evaluated', the systems assigned the concept 'hemochromatosis' (C0018995), whereas the GSC annotated the concept 'hereditary hemochromatosis' (C0392514). It should be noted that 'hemochromatosis' is not part of the list of terms in the UMLS corresponding with the concept C0392514. Thus, this concept is not even considered by the disambiguation algorithms of the systems. Knowledge-based disambiguation approaches that can take into account the concept relationships defined in the UMLS may be able to solve these disambiguation problems.

Usage of the NLP module gives a larger increase in precision than in recall (cf. table 1), even though most rules are aimed at finding missed concepts. This can partly be explained by the filtering submodule, which by its nature can only improve precision, but also some of the other submodules improve precision more than recall. The reason is that if a rule finds a missed concept it often suppresses one or more erroneous concepts that were initially found by the system, thus improving precision.

Peregrine gave a slightly better performance than MetaMap when exact matching was used for evaluation, but for inexact matching the performances were similar, both for concept recognition and for concept identification. The two systems used different UMLS versions (2011AA and 2011AB), but the differences between these versions are very small and unlikely to be the cause of performance differences. Since the NLP module was developed on the basis of the errors made by Peregrine, one might suspect a performance bias in favor of this system in combination with the NLP module. However, when we determined the performance on the training set, the F-score turned out to be only 1.9 percentage points higher than on the test set, indicating hardly any overtraining. For MetaMap, this difference was 1.7 percentage points. With the use of the NLP module, Peregrine and MetaMap showed a comparable gain in performance.

Many of the rules in our NLP module have not been used before in their specific form, but similar rules have previously been proposed in many different studies. The combination of different types of rules in one system, showing the contribution of each submodule to total performance, and their application to disease normalization, a task which has not been addressed before, is novel in our study. The submodules are general and may be combined, as a whole or separately, with other concept normalization systems.

In developing the NLP module, we manually constructed rules and did not use machine-learning techniques, as did previous studies that used the AZDC for system development and evaluation.[34 39] These machine-learning based systems achieved comparable or slightly better performance for concept recognition as our rule-based systems, but did not address the normalization task. Moreover, we believe our approach offers several benefits. First, machine-learning based systems are often not transparent, whereas the man-made rules are comprehensible. This is likely to ease error detection and correction, incremental rule improvement, and adaptation to other domains. Also, the rules combine input from heterogeneous systems in a very flexible way, which may be more difficult to achieve by machine learning methods that have a fixed knowledge representation model. Furthermore, machine learning methods require sufficiently large GSCs for training. Although this requirement apparently was met in the case of AZDC, this may not be true for other application areas. We also need a GSC for developing our rules, but the size can be relatively small because human experts also bring in background knowledge that can compensate for scarce data. Finally, although we did not put it to the test, it is conceivable that machine-learning based concept recognition may still benefit from our NLP module because it may capture patterns not well handled by machine learning. Whether this would also translate into better concept normalization would also depend on the normalization step that needs to follow machine-learning based concept recognition.

The error analysis indicated that about half of the errors that remain after applying the NLP module can be denoted as term variation and filtering errors. While further improvement of the submodules dealing with these errors may be possible, it is more likely that improved dictionaries and disambiguation methods will help to reduce these types of errors. In this respect, further work on the generation of term variants would be useful. We also noticed that shallow parsing sometimes provides insufficient information to resolve errors in complex sentences. Such information may possibly derive from deep parsing, but exploring the usefulness of these techniques for our purposes will be left to future research.

We have shown the added value of the NLP module for the recognition and normalization of diseases with MetaMap and Peregrine. The module is general and can be applied in combination with any concept normalization system. Whether its use for the normalization of other concept types, such as genes or drugs, is equally advantageous still remains to be investigated.

Contributors NK developed the methodology, carried out the experiments, analyzed the data, and drafted the manuscript. BS and ZA helped to build the software infrastructure. EMM and JAK conceived and supervised the study, co-developed the methodology, and helped to write the paper. All authors read and approved the manuscript.

Funding This study was partially supported by the European Commission FP7 Program (FP7/2007-2013) under grant no. 231727 (the CALBC project).

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Krauthammer M, Nenadic G. Term identification in the biomedical literature. *J Biomed Inform* 2004;37:512–26.

2 Harris MA, Clark J, Ireland A, *et al*. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 2004;32:258–61.

3 Maglott D, Ostell J, Pruitt KD, *et al*. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 2007;35:26–31.

4 Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;32:267–70.

5 Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of the AMIA Symposium*. Philadelphia, PA, 2001:17–21.

6 Schuemie MJ, Jelier R, Kors JA. Peregrine: lightweight gene name normalization by dictionary lookup. *Proceedings of the BioCreAtIvE II Workshop*; Madrid, Spain, 2007:131–3.

7 Shah NH, Bhatia N, Jonquet C, *et al*. Comparison of concept recognizers for building the Open Biomedical Annotator. *BMC Bioinform* 2009;10:S14.

8 Mutalik PG, Deshpande A, Nadkarni PM. Use of general-purpose negation detection to augment concept indexing of medical documents. *J Am Med Inform Assoc* 2001;8:598–609.

9 Rebholz-Schuhmann D, Arregui M, Gaudan S, *et al*. Text processing through Web services: calling Whatizit. *Bioinformatics* 2008;24:296–8.

10 Hirschman L, Yeh A, Blaschke C, *et al*. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinform* 2005;6:S1.

11 Kim J-D, Ohta T, Pyysalo S, *et al*. Overview of BioNLP'09 shared task on event extraction. *Proceedings of the Workshop on BioNLP Shared Task*; Boulder, USA, 2009:1–9.

12 Voorhees EM, Tong RM. Overview of the TREC 2011 medical records track. *Proceedings of the twentieth Text REtrieval Conference (TREC)*; Gaithersburg, USA, 2011.

13 Uzuner O, South BR, Shen S, *et al*. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 2011;18:552–6.

14 Hirschman L, Colosimo M, Morgan A, *et al*. Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinform* 2005;6(Suppl 1):S11.

15 Morgan AA, Lu Z, Wang X, *et al*. Overview of BioCreative II gene normalization. *Genome Biol* 2008;9(Suppl 2):S3.

16 Lu Z, Kao HY, Wei CH, *et al*. The gene normalization task in BioCreative III. *BMC Bioinform* 2011;12:S2.

17 Hakenberg J, Plake C, Leaman R, *et al*. Inter-species normalization of gene mentions with GNAT. *Bioinformatics* 2008;24:126–32.

18 Wermter J, Tomanek K, Hahn U. High-performance gene name normalization with GeNo. *Bioinformatics* 2009;25:815–21.

19 Hanisch D, Fundel K, Mevissen H-T, *et al*. ProMiner: rule-based protein and gene entity recognition. *BMC Bioinform* 2005;6:S14.

20 Gerner M, Nenadic G, Bergman CMC-P. LINNAEUS: a species name identification system for biomedical literature. *BMC Bioinform* 2010;11:85.

21 Dai HJ, Chang YC, Tzong-Han Tsai R, *et al*. New challenges for biological text-mining in the next decade. *J Comput Sci Tech* 2010;25:169–79.

22 Struble CA, Povinelli RJ, Johnson MT, *et al*. Combined conditional random fields and n-gram language models for gene mention recognition. *Proceedings of the Second BioCreative Challenge Evaluation Workshop*; Madrid, Spain, 2007:81–3.

23 Buyko E, Tomanek K, Hahn U. Resolution of coordination ellipses in biological named entities using conditional random fields. *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING)*; Melbourne, Australia, 2007:163–71.

24 Baumgartner WA, Lu Z, Johnson HL, *et al*. Concept recognition for extracting protein interaction relations from biomedical text. *Genome Biol* 2008;9:S9.

25 Jimeno-Yepes A, Berlanga-Llavori R, Rebholz-Schuhmann D. Ontology refinement for improved information retrieval. *Inform Process Manag* 2010;46:426–35.

26 . Schwartz Hearst MAAS. A simple algorithm for identifying abbreviation definitions in biomedical text. *Proceedings of the 8th Pacific Symposium on Biocomputing*; Hawaii, USA, 2003:451–62.

27 Gaudan S, Kirsch H, Rebholz-Schuhmann D. Resolving abbreviations to their senses in Medline. *Bioinformatics* 2005;21:3658–64.

28 Okazaki N, Ananiadou S, Tsujii J. Building a high-quality sense inventory for improved abbreviation disambiguation. *Bioinformatics* 2010;26:1246–53.

29 Atzeni P, Polticelli F, Toti D. An automatic identification and resolution system for protein-related abbreviations in scientific papers. *Proceedings of the 9th European conference on Evolutionary computation, machine learning and data mining in bioinformatics*. Torino, Italy, 2011:27–9.

30 Schuemie MJ, Mons B, Weeber M, *et al*. Evaluation of techniques for increasing recall in a dictionary approach to gene and protein name identification. *J Biomed Inform* 2007;40:316–24.

31 Tsuruoka Y, McNaught J, Ananiadou S. Normalizing biomedical terms by minimizing ambiguity and variability. *BMC Bioinform* 2008;9:S2.

32 Ratkovic Z, Golik W, Warnier P. Event extraction of bacteria biotopes: a knowledge-intensive NLP-based approach. *BMC Bioinform* 2012;13:S8.

33 Wang X, Matthews M. Distinguishing the species of biomedical named entities for term identification. *BMC Bioinform* 2008;9:S6.

34 Leaman R, Miller C, Gonzalez G. Enabling recognition of diseases in biomedical text with machine learning: corpus and benchmark. *Proceedings of the 3rd International Symposium on Languages in Biology and Medicine (LBM)*; Jeju Island, South Korea, 2009:82–9.

35 Bada M, Hunter LE, Eckert M, *et al*. An overview of the CRAFT concept annotation guidelines. *Proceedings of the Fourth Linguistic Annotation Workshop*; Uppsala, Sweden, 2010:207–11.

36 Thompson P, Iqbal SA, McNaught J, *et al*. Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinform* 2009;10:349.

37 Doğan RI, Lu Z. An improved corpus of disease mentions in PubMed citations. *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing (BioNLP)*; Montreal, Canada, 2012.

38 Hahn U, Buyko E, Landefeld R, *et al*. An overview of JCoRe, the JULIE lab UIMA component repository. *Proceedings of the Language Resources and Evaluation Conference (LREC)*; Marrakech, Morocco, 2008:1–7.

39 Chowdhury M, Faisal M. Disease mention recognition with specific features. *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing (BioNLP)*; Uppsala, Sweden, 2010:83–90.

40 Hettne KM, Van Mulligen EM, Schuemie MJ, *et al*. Rewriting and suppressing UMLS terms for improved biomedical term identification. *J Biomed Semantics* 2010;1:1–5.

41 Buyko E, Wermter J, Poprat M, *et al*. Automatically Adapting an NLP Core Engine to the Biology Domain. *Proceedings of the Joint BioLINKBio-Ontologies Meeting*; 2006:2–5.

42 Vilares J, Alonso MA, Vilares M. Extraction of complex index terms in non-English IR: a shallow parsing based approach. *Inform Process Manag* 2008;44:1517–37.

43 Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Nat Lang Eng* 2004;10:327–48.