Extracting Diagnostic Knowledge from MedLine Plus: a Comparison between MetaMap and cTAKES Approaches

Alejandro Rodríguez-González[*a], Roberto Costumero[*b], Marcos Martinez-Romero[c], Mark D.Wilkinson[a] and Ernestina Menasalvas-Ruiz[b]

[a]*Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid, Pozuelo de Alarcon, Spain;* [b]*Centro de Tecnología Biomédica, Universidad Politécnica de Madrid, Pozuelo de Alarcon, Spain;* [c]*Stanford Center for Biomedical Informatics Research, Stanford University*

Abstract: The development of diagnostic decision support systems (DDSS) requires having a reliable and consistent knowledge base about diseases and their symptoms, signs and diagnostic tests. Physicians are typically the source of this knowledge, but it is not always possible to obtain all the desired information from them. Other valuable sources are medical books and articles describing the diagnosis of diseases, but again, extracting this information is a hard and time-consuming task. In this paper we present the results of our research, in which we have used Web scraping, natural language processing techniques, a variety of publicly available sources of diagnostic knowledge and two widely known medical concept identifiers, MetaMap and cTAKES, to extract diagnostic criteria for infectious diseases from MedLine Plus articles. A performance comparison of MetaMap and cTAKES is also presented.

1. INTRODUCTION

Diagnosing a disease is a complex process that is focused on the identification and interpretation of clinical observations (findings) of a patient. These findings (symptoms, signs and diagnostic tests) allow the physician to determine (or discard) a list of possible diseases or evaluate the procedures that will help in the selection of the final diagnosis [1]. The physician tries to find patterns among the findings associated to the patient and findings related to known diseases. Findings are also essential for researchers to examine the relationships between diseases based on symptom-similarity, and to find genetic relationships between the molecular origins of diseases and their resulting phenotypes [2].

Clinical decision support systems (CDSS) have demonstrated to be helpful in the reduction of diagnosis errors, ensuring comprehensive treatment of patient illnesses and conditions and decreasing expenses over time. Nevertheless, the development of CDSS for diagnosis support requires having a high quality and comprehensive knowledge base with information about all possible clinical findings for each disease [3]. The Web contains a vast amount of resources with freely available diagnostic knowledge, but this knowledge is typically dispersed across a number of electronic sources, such as medical texts, databases and ontologies. Some collaborative sources (e.g. Freebase) contain rich and structured diagnostic knowledge, but their reliability can be questioned. Other sources (e.g. MedLine Plus) contain high quality and complete knowledge but it is extremely hard to reuse it because it is expressed in free text format. Extracting diseases and findings from unstructured medical text constitutes a basic enabling technology to unlock the knowledge within texts and support the development of advanced systems such as DDSS [4].

Several approaches for the extraction of medical knowledge from unstructured sources have been proposed during the last years. Some examples are Tsumoto [5], Tan et al. [6], Hahn et al. [7] and Amaral et al. [8]. However, most of previous efforts are focused on specific medical areas, such as radiology, or on complex, specialized tasks. They do not address the extraction of the basic clinical terms used to express the diagnostic criteria of a disease.

In this research, we have conceived, tested and evaluated a new way of extracting relevant medical diagnostic terms from a set of online MedLine Plus articles about infectious diseases as an extension of a previously published research [9]. We have developed a prototype capable of crawling the HTML code of the Web pages in order to extract all relevant diagnosis-related content (symptoms, signs and diagnostic tests). Then, we have applied a named-entity recognition approach to extract all relevant terms based on two of the most widely established biomedical entity recognizers: MetaMap [10] and cTAKES [11]. After that, the terms provided by MetaMap and cTAKES were validated using knowledge extracted from several reference terminological resources. This work also allowed us to compare the results provided by MetaMap and cTAKES to analyze their behavior when dealing with diagnosis-related texts. The output of our process is a list of diagnosis-related terms for each disease.

The reminder of the paper is organized as follows: Section 2 presents related work. Section 3 provides a detailed explanation about the method that has been conceived and the prototype that has been built. Section 4 presents the

results of the experiments performed to test the validity of the proposed approach. Finally, section 5 concludes the paper and sketches some future research directions.

*Address correspondence to this author at the Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid, Campus de Montegancedo, Pozuelo de Alarcon, 28223, Madrid, Spain. Tel: +34 914524900 (25550); E-mail: alejandro.rodriguezg@upm.es

2. RELATED WORK

There is a vast amount of diagnostic knowledge distributed across a variety of electronic resources, such as medical websites, databases and ontologies. Ontological sources such as the Unified Medical Language System (UMLS) [12] can provide valuable, but somewhat limited, diagnostic information. Some sources are focused on very specific domains, such as Online Mendelian Inheritance in Man (OMIM) [13] and the Human Phenotype Ontology (HPO) [14], and do not always contain the required knowledge. Textual databases such as Wikipedia and Freebase also contain valuable knowledge, but the reliability and completeness of their information is questionable. Another well-known resource is the DiseasesDatabase[1], however it is not possible to access its raw data.

MedlinePlus is an online free information service provided by the US National Library of Medicine, which is considered the world's largest medical library [15]. It provides reliable and up-to-date medical knowledge, including information about over 950 diseases and conditions. The data provided for each disease may vary, but it usually includes a description of the disease, causes, symptoms, exams and tests, and treatment. However, there are currently no methods or systems to extract all relevant information from Medline Plus in a structured and reusable format and make it available for further research and analysis.

Tools for Medical Term Extraction

Extracting relevant knowledge from narrative text and associating it to a logically structured domain vocabulary (i.e. a controlled terminology or ontology) is a challenging task that has been a topic of intensive study over the last decade. A variety of concept identification systems have been conceived to address this task. Two of the most comprehensive and broadly used tools are MetaMap [10] and cTAKES [11], which are freely available to the public and make it possible to discover medical entities in text and map them to ontology concepts.

MetaMap has been developed by the National Library of Medicine (NLM) and makes it possible to identify UMLS Metathesaurus concepts referred to in biomedical text. MetaMap is very flexible and easily customizable, and has been become a standard tool in mapping biomedical text to standard terminologies. MetaMap is, by design, very tightly coupled with the UMLS Metathesarus. Adapting it to use a custom dictionary of biomedical terms outside UMLS is non-trivial because MetaMap requires the external dictionaries to be in a specific format with certain databases always present, and this translation is not always possible [16]. This drawback does not negatively affect our approach because our concept identification process is UMLS-based.

The Apache Clinical Text Analysis and Knowledge Extraction System (cTAKES) [11] is an open-source NLP system that is focused on the extraction of information from clinical documents. It was built using the Apache UIMA Unstructured Information Management Architecture engineering framework [17] and the OpenNLP natural language processing toolkit [18]. cTAKES is able to identify named entities from various dictionaries, including a subset of UMLS (diseases and disorders, signs and symptoms, medications, anatomical sites and procedures). The main cTAKES components are sentence boundary detectors, tokenizers, normalizers, Part-of-Speech (PoS) taggers, shallow parsers and Named Entity Recognition (NER) annotators.

Although MetaMap and cTAKES are the most established "off the shelf" tools, there are other alternatives that have been developed to extract knowledge from medical text. MedLEE (Medical Language Extraction and Encoding System) [19], was originally designed to process radiology notes and later extended to other sub-domains. The NCBO Annotator [20] is an online service that identifies biomedical ontology concepts in unstructured texts. It works against the BioPortal ontology repository, which currently contains more than 400 biomedical ontologies. BeCAS (Biomedical Concept Annotation System) is a recently released service for biomedical concept recognition [21] that can be integrated on larger text-processing pipelines and that was tested on both abstracts and full-length scientific articles.

Diagnostic Knowledge Extraction using MetaMap and cTAKES

MetaMap and cTAKES have been previously applied in several notable studies to extract diagnostic knowledge from medical text. Okumura et al. [3] performed an analysis of the mapping between clinical vocabularies and findings in medical literature using OMIM as a knowledge source and MetaMap as the NLP tool. Another very interesting approach was undertaken by Okumura & Tatesi [22] where the authors analyze the application of MetaMap to the efficient extraction of symptomatic expressions, adding some heuristics to exploit patterns of tag sequences that frequently appear in typical symptomatic expressions.

Several authors have compared MetaMap and cTAKES in terms of their performance when extracting clinical terms from texts. Wu et al. [23] analyzed how well they recognize and interpret abbreviations in clinical texts (e.g. "abd" for "abdominal"). The results of this study showed that MetaMap handles abbreviations better than cTAKES (F-scores of 0.338 and 0.165 respectively) but they do not perform well overall because they are not been designed for this particular problem. MedLEE is a better choice for this abbreviation detection task (F-score: 0.601). They concluded that more advanced abbreviation recognition modules are necessary. Osborne et al. [24] applied MetaMap and YTEX [25], which is an extension of cTAKES to two concept recognition tasks. Their results suggested that YTEX would be a better system for "off the shelf" concept mapping. Also, YTEX scales better than MetaMap. However, MetaMap may

[1] http://www.diseasesdatabase.com/

be a better option for precisely identifying concept boundaries, that is, the start and end tokens that delimit a string associated to a complete concept (e.g. "white blood cell" instead of "blood cell" or "cell"). Collier et al. [26] applied MetaMap and cTAKES to the extraction of phenotypes and other related concepts that concern the diagnosis and treatment of diseases. They concluded that cTAKES performs well overall but that annotation performance varies widely across semantic types, and that MetaMap with the strict matching and word sense disambiguation features enabled can have superior precision. Another relevant work is the study performed by Denecke [27], who applied cTAKES and MetaMap to a real-world set of medical blog postings. The results of this study showed that both tools perform well when medical conditions or procedures are explicitly mentioned and described by nouns but that they often fail in mapping or produce wrong mappings for verbs, personal pronouns, adjectives and connecting words. When the text is dealing with diseases and clearly mentioned symptoms, which is the case of our work, both tools provide appropriate annotations. Ambiguity of terms led to errors in both systems.

Related work shows that current efforts are aligned with our approach. However, our research is not only focused on the extraction of diagnostic knowledge, but also on the validation of the results obtained by means of external information sources. Our analysis of previous work also allowed us to confirm that MetaMap and cTAKES, in spite of their limitations, are appropriate for our purpose of extracting diagnostic knowledge from narrative text.

3. MATERIALS AND METHODS

In this section we present our approach to extract and validate diagnostic knowledge from MedLine Plus. We describe the architecture and workflow of our approach and explain the different steps involved in the knowledge extraction and validation processes.

3.1 Architecture and general workflow

Figure 1 shows the high-level architecture and workflow of our approach. Our workflow is grounded on three fundamental steps:

1. **Medical Text Extraction and NLP Procedures**. MetaMap and cTAKES are applied to MedLine Plus articles to identify all relevant diagnostic-related terms.
2. **Validation Terms Extraction**. It comprises the extraction of diagnostic-knowledge from a set of structured information resources.
3. **Term validation**. The knowledge extracted from MedLine Plus (step 1) is validated against the knowledge extracted from other resources (step 2).

A more comprehensive explanation of the above steps is provided in the following sections.

Medical Text Extraction and NLP Procedures (MTENP Procedure)

Our analysis on the HTML code of the pages provided by MedLine Plus reveals that it is predictably structured, in contrast to other well-known health web pages with disease information such as the Centers for Disease and Prevention (CDC) website [28]. This reliability of source-data structure, together with the reliability of the MedLine Plus data itself, was the determining factor in its selection as the information source for our research.

This procedure starts from the Medline Plus URL for a particular disease. Web scraping using on the JSoup API[2] is applied to extract the text from the sections that we considered relevant for our purpose. These sections are "Symptoms" and "Exam and Tests".

Figure 1 Architecture of the proposed solution.

[FIGURE1]

After the Web scraping step, the NLP Procedure (MetaMap Filter) applies MetaMap or cTAKES to the text extracted from the webpage. The proposed solution lets the user decide whether to use MetaMap or cTAKES to annotate the text. Each subsystem restricts the semantic types in a different way. On the one hand, MetaMap uses "Diagnostic Procedure", "Disease or Syndrome", "Finding", "Laboratory Procedure", "Laboratory or Test Result" and "Sign or Symptom". On the other hand, cTAKES uses only three semantic types for the same categories, which are "Disease Disorder Mention", "Sign Symptom Mention", and "Procedure Mention". The correspondence between the semantic types used by MetaMap and cTAKES can be found in Table 1. This procedure returns a list of medical terms that are considered relevant based on the previous semantic types.

Table 1. Correspondence between MetaMap and cTAKES semantic types.

MetaMap	cTAKES
Diagnostic Procedure (diap)	Procedure Mention
Disease or Syndrome (dsyn)	Disease Disorder Mention
Finding (fndg)	Sign Symptom Mention
Laboratory Procedure (lbpr)	Procedure Mention
Laboratory or Test Result (lbtr)	Procedure Mention
Sign or Symptom (sosy)	Sign Symptom Mention

Validation Terms Extraction Procedure (VTE Procedure)

Some of the results returned by MetaMap and cTAKES may not be correct. For example, MetaMap classifies the word "red" as a finding. cTAKES also classifies terms such as "pale" as a finding. As a consequence, the goal of this

[2] http://jsoup.org/

procedure is to improve the overall accuracy of our approach by acquiring diagnostic knowledge from other information resources different from MedLine Plus. This knowledge will be used later to validate the knowledge returned by the MTENP Procedure.

The VTE Procedure extracts medical terms from several publicly available information sources (see Table 2), which can be classified into four categories:

- Trusted sources: Sources created and curated by widely known institutions or organizations, such as the World Health Organization. For instance, the International Classification of Diseases 10 – Clinical Modification (ICD10CM) is a classification provided by the Centers for Medicare and Medicaid Services (CMS) and the National Center for Health Statistics (NCHS), for medical coding and reporting in the United States. It classifies diagnoses and reason for visits in all American health care settings.
- Research sources: Sources created as part of a research. Includes: CCSO Signs and Symptoms Ontology, TM Signs and Symptoms Ontology (TM SSO) and Symptoms Ontology.
- Collaborative sources: Wikipedia and Freebase. The reliability of these sources is, in general, not as good as the sources mentioned above, but they contain a vast amount of information that can be used in combination with other sources for validation purposes.
- Other sources: Other medical webs. Includes: Medicinet[3] (Tests section).

VTE obtains the list of terms differently depending on the source (see Table 2).

Table 2. Summary of validation sources and extraction method used.

Extraction method and origin	Sources
Manual (webpage)	Medicinet Wikipedia
Automatic (Bioportal OpenLifeData)	ICD9CM, ICD10CM Symptoms Ontology TM SSO
Automatic (Jena and MQL)	Freebase (MQL) CSSO (Jena)

Terms from BioPortal[4] and OpenLifeData[5] sources were obtained through their SPARQL Endpoint using Jena API[6]. CSSO Ontology terms were automatically extracted from the ontology using Jena API.

Term Validation Process (TV Procedure)

VTE and MTENP Procedures can be executed separately in order to obtain the list of terms that will be used by the TV Procedure. The TV Procedure is in charge of analyzing the terms provided by the MTENP procedure to ensure they match the VTE-provided terms. If the TV procedure finds a match, the term will be returned as a valid diagnostic term. Validation is performed by TV as follows:

Given a term (t) from the list provided by MTENP (being independent if the execution is done using MetaMap or cTAKES), the process attempts to find a matching term (mt) from the list provided by VTE. The matching will be considered valid in any of the following situations (ordered by matching accuracy):

1. CUI Identification: For every concept classified under a UMLS source, UMLS provides a Concept Unique Identifier (CUI) (e.g. C0015967). If the CUI of "t" and "mt" are the same, there is a matching between them. This is the situation in which the match is considered most reliable.
2. Equals: If the string that represents "t" (or any of its associated synonyms) is the same as the string that represents "mt" (or any of its associated synonyms), then there is a match between them.
3. Similarity: A similarity score between "t" (and any of its synonyms) and "mt" (or any of its synonyms) is calculated, by means of the Levenshtein distance algorithm, with a threshold value of 0.85. We have used the implementation of the Levenshtein distance algorithm provided by the SimMetrics Java API[7]. The terms "t" and "mt" are pre-processed previously in order to maximize the possibility of finding a similarity. Pre-processing includes removal of stop words, symbols and trimming of the string among others.

If a matching is found, it is assumed that the term "t" is a valid diagnostic term and it is added to the final list of results.

The process carried out by VTE tries to give preference to those validation terms who came, first, from trusted sources; second, from research sources; third, from collaborative sources; and finally from other sources, in order of priority.

The source code of the prototype is publicly available at GitHub[8].

4. EVALUATION

The approach has been tested by executing the prototype over data associated to 30 different infectious diseases[9],

[3] http://www.medicinenet.com/procedures_and_tests/article.htm

[4] http://bioportal.bioontology.org/
[5] http://www.openlifedata.org/
[6] https://jena.apache.org/
[7] http://sourceforge.net/projects/simmetrics/
[8] https://github.com/alejandrorg/medlineplus2ddx

selected manually by our researchers. Infectious diseases typically have a large number of symptoms and diagnostic tests, providing a large variety of terms that should be extracted by our platform.

The evaluation was performed by doing a manual analysis of the results provided by our approach. For each disease, we compared: (1) the list of terms provided by our approach; with: (2) a list of terms manually extracted from the disease Web page.

True positive (TP), false positive (FP), true negative (TN) and false negative (FN) parameters were computed in order to calculate precision, recall, specificity and F1 score values. The mean values obtained using both the approach of MetaMap and cTAKES for these parameters based on the individual values for each disease are shown in Figure 2 for an easier comparison. Detailed results for each disease are available online[10],[11] including the list of terms manually extracted from the disease web page, the matchings with the list of terms provided by our approach and some extra information such as the primary type of source used for the matching.

The results show that our method performed well. A detailed analysis of the results, disease-by-disease, shows that some of the false negatives were a consequence of our validation method used. Thus in those cases the term as correctly classified either by MetaMap of cTAKES but our validation method had discarded it. This effect is especially noteworthy for the case of acronyms within diagnostic tests. The sources used to generate the list of validation terms were impoverished for terms related to diagnostic tests (and diagnostic tests results) resulting in a high number of false negatives.

Figure 2. Comparison between statistical results for MetaMap & cTAKES executions.

[FIG2]

Another problem that we have identified relates to "classical" false negatives (a term that has been incorrectly rejected). Several terms were not identified by the NLP process. Most of these terms are "sentences" or composite phrases, which complicates their identification by the NLP process. Finally, there were very few false positives, though it could be relevant that cTAKES found more than MetaMap.

As it can be seen in Figure 2, the mean results are quite similar between the executions using MetaMap or cTAKES. Precision is higher on MetaMap, while recall is higher in cTAKES. Specificity and F1 score are roughly the same, the former being higher in MetaMap, while the latter is higher in cTAKES.

The main differences are found in the analysis of individual diseases. cTAKES typically performs better on laboratory or test results or locating rare symptoms, but increases in most cases the number of false positives, incorrectly annotating several elements as findings. In this case, it could also be relevant that the number of true negatives is higher because the NLP process annotates more elements, but the validation usually classifies them correctly. The results for the different diseases including precision, recall, specificity and F1 scores can be found in Figures 3, 4, 5 and 6, respectively.

Figure 3. Comparison of precision results for MetaMap & cTAKES executions for each disease.

[FIG 3]

In Figure 3 we observe that the precision using cTAKES is higher than the one of MetaMap in only 5 of the diseases: Tonsilitis, Erypselas, Fifth disease, Roseola and Scarlet Fever. In most cases this is due to the increase in false negatives and the increase in true positives. For the rest of the diseases, the differences in precision are high enough to ensure the great difference in precision between both systems.

Figure 4. Comparison of recall results for MetaMap & cTAKES executions for each disease.

[FIG4]

In Figure 4 the recall is analyzed using both NLP systems and we observe the inverse effect than the one observed in Figure 3, as there are 6 diseases where MetaMap performs better than cTAKES: Cholera, Diphtheria, Impetigo, Mumps, Roseola and Pertussis. In this case the differences are less significant between both systems but the improvement using cTAKES is observable.

Figure 5. Comparison of specificity results for MetaMap & cTAKES executions for each disease.

[FIG 5]

In Figure 5 we compare the specificity between MetaMap and cTAKES for all the 30 diseases used in our experiments. In this particular case is worth mentioning cases such Poliomyelitis od Rubella, where both systems perform at their best, or the great differences encountered for example in the analysis of Roseola, where the difference in specificity is greater than 0.4 between both systems.

Figure 6. Comparison of f1 score results for MetaMap & cTAKES executions for each disease.

[FIG 6]

In Figure 6 the comparison between the F1-scores is shown. The lines in this chart are approximately overlapping, meaning that both systems are performing similarly. It is worth noting the two big differences encountered in Tetanus and Mumps with more than 0.2 difference between both

[9] https://github.com/alejandrorg/medlineplus2ddx/blob/master/diseasesList.txt

[10] https://github.com/alejandrorg/medlineplus2ddx/blob/master/MetaMap.xlsx

[11] https://github.com/alejandrorg/medlineplus2ddx/blob/master/cTAKES.xlsx

systems, performing cTAKES better on the former and MetaMap on the latter.

CONCLUSIONS AND FUTURE WORK

This paper presented a novel approach to extracting diagnostic knowledge for infectious diseases from MedLine Plus articles. Evaluation of the prototype developed reveals that the proposed method is accurate enough to be used to extract diagnostic-related terms from a variety of unstructured information sources, such as Web pages and clinical notes. However, the evaluation also reveals that improvements could enhance performance.

As future work we consider that an expansion of the VTE procedure through adding new data sources to increase the number of validation terms might improve the quality of the results through reduction of false negatives. Another future line would be to increase the number of NLP tools to contrast the results and even create further filters to complement and gather the best information from all of them, merging the results obtained. Finally, we plan to extend our work to the domain of treatment information to enrich the knowledge extracted.

LIST OF ABBREVIATIONS

If abbreviations are used in the text either they should be defined in the text where first used, or a list of abbreviations can be provided.

CONFLICT OF INTEREST

Authors declare no conflict of interest.

Alejandro Rodríguez González's and Mark Wilkinson's work is supported by Isaac Peral Programme of the UPM.

ACKNOWLEDGEMENTS

All the authors have contributed equally to the research and manuscript.

SUPPLEMENTARY MATERIAL

MetaMap Raw Results: Raw results obtained with MetaMap are available online (see footnote 10).

cTAKES Raw Results: Raw results obtained with cTAKES are available online (see footnote 11).

REFERENCES

[1] Rodríguez-González, A, Martinez-Romero, M, Egaña-Aranguren, M, Wilkinson, MD. Nanopublishing Clinical Diagnoses: Tracking Diagnostic Knowledge Base Content and Utilization. IEEE 27th International Symposium on Computer-Based Medical Systems (CBMS), 2004, May 27-29, New York, USA

[2] Zhou X, Menche J, Barabási AL, Sharma A. Human symptoms–disease network. Nat Commun, 2013;5:4212

[3] Okumura T, Aramaki E and Tateisi Y. Clinical Vocabulary and Clinical Finding Concepts in Medical Literature. Proceedings of the International Joint Conference on Natural Language Processing Workshop on Natural Language Processing for Medical and Healthcare Fields, 7–13, 2013

[4] Jonnalagadda S, Cohen T, Wu S, Gonzalez G. Enhancing clinical concept extraction with distributional semantics. J Biomed Inform, 2012;45(1):129-140.

[5] Tsumoto S. Automated extraction of medical expert system rules from clinical databases based on rough set theory. Inf Sci, 1998;12(1-4):67-84.

[6] Tan KC, Yu, Q, Heng CM, Lee TH. Evolutionary computing for knowledge discovery in medical diagnosis. Artif Intel in Medi, 27:129–154, 2003

[7] Hahn U, Romacker M, Schulz S. medSynDiKATe—a natural language system for the extraction of medical information from findings reports. Int J Med Inform, 2002;67(1–3):63–74

[8] Amaral MB, Roberts A, Rector AL. NLP techniques associated with the OpenGALEN ontology for semi-automatic textual extraction of medical knowledge: abstracting and mapping equivalent linguistic and logical constructs. Proceedings if the AMIA Annual Symposium, 2000

[9] Rodríguez-González A, Costumero R, Martínez-Romero M, Wilkinson MD, Menasalvas-Ruiz E. Diagnostic knowledge extraction from MedlinePlus: an application for infectious diseases. 9th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB 2015), 2015, June 3-5, Salamanca, Spain

[10] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings of the AMIA Annual Symposium, 2001

[11] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc, 2010; 17(5):507-513.

[12] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res, 2004;32(1):267-270

[13] Hamosh, A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res, 2005;33(1):514-517

[14] Köhler S, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res, 2014;42(D1):966-974

[15] Lindberg DAB. About MEDLINEplus. 1999. URL: http://www.nlm.nih.gov/medlineplus/aboutmedlineplus.html (Accessed 25 August 2015).

[16] Bhatia N, Shah NH, Rubin DL, Chiang AP, Musen MA. Comparing concept recognizers for ontology-based indexing: MGREP vs. MetaMap. AMIA Summit on Translational Bioinformatics, 2009.

[17] Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Nat Lang Eng, 2004; 10(3-4):327-348

[18] OpenNLP. Apache software foundation. URL http://opennlp.apache.org.

[19] Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc, 1994;1(2):161

[20] Jonquet C, Shah N, Youn C, Callendar C, Storey MA, Musen M. NCBO annotator: semantic annotation of biomedical data. In International Semantic Web Conference, 2009

[21] Nunes T, Campos D, Matos S, Oliveira JL. BeCAS: biomedical concept recognition services and visualization. Bioinf, 2013;29(15):1915-1916

[22] Okumura T, Tateisi Y. A Lightweight Approach for Extracting Disease-Symptom Relation with MetaMap toward Automated Generation of Disease Knowledge Base. Health Inf Sci, 2012, 164-172

[23] Wu Y, Denny JC, Rosenbloom ST, Miller RA, Giuse DA, Xu H. A comparative study of current Clinical Natural Language Processing systems on handling abbreviations in discharge summaries. Proceedings of the AMIA Annual Symposium, 2012

[24] Osborne JD, Gyawali B, Solorio T. Evaluation of YTEX and MetaMap for clinical concept recognition. arXiv preprint, 2014. arXiv:1402.1668.

[25] Garla V et al. The Yale cTAKES extensions for document classification: architecture and application. J Am Med Inform Assoc, 2011;18(5):614-620

[26] Collier N, Oellrich A, Groza T. Concept selection for phenotypes and diseases using learn to rank. J Biomed Semantics, 2015;6(1):24

[27] Denecke K. Extracting Medical Concepts from Medical Social Media with Clinical NLP tools: A Qualitative Study. Proceedings of the Fourth Workshop on Building and Evaluation Resources for Health and Biomedical Text Processing, 2014

[28] CDC. Center for Diseases and Prevention. URL http://cdc.gov (Accessed 25 August 2015).