Medical Entity and Relation Extraction from Narrative Clinical Records in Italian Language

Crescenzo Diomaiuta, Maria Mercorella(✉), Mario Ciampi, and Giuseppe De Pietro

National Research Council of Italy, Institute of High Performance Computing and Networking - ICAR, Via Pietro Castellino 111, 80131 Naples, Italy
{crescenzo.diomaiuta,maria.mercorella,mario.ciampi, giuseppe.depietro}@icar.cnr.it

Abstract. Applying Natural Language Processing techniques enables to unlock precious information contained in free text clinical reports. In this paper, we propose a system able to annotate medical entities in narrative records. Considering that existing NLP systems mainly concern entity recognition in English language, we propose an NLP pipeline to manage clinical free text in Italian. The overall architecture includes a spell checker, sentence detector, word tokenizer, part-of-speech tagger, dictionary lookup annotator, and parsing rules annotator. Essentially, it uses a rule-based approach to extract relevant concepts regarding patient's conditions, administered medications, or performed procedures, detecting their attributes, negated forms, and relations expressions. The indexing of the documents allows the user to retrieve relevant information, increasing his/her medical knowledge.

Keywords: Italian natural language processing · Medical entity recognition · Information Extraction · Unstructured medical records · UIMA

1 Introduction

Medical records contain a large amount of clinical data in the form of narrative text written by clinicians. This unstructured part is rich in useful mentions about patient's significant problems, performed medical procedures, and consumed drugs. The identification of key concepts from these text segments enables the automated processing of contained clinical data. Information Extraction (IE) concerns the extraction of predefined types of information from natural language text. In particular, Named Entity Recognition (NER) is a subfield of IE that aims at recognizing specific entities, such as diseases, drugs, or names, in free text documents [22]. IE, due to the ability to extract and convert data in a structured format, can support epidemiology studies, clinical decision, text mining, and automatic terminology management. Several issues make especially challenging the processing of clinical narrative text. Indeed, medical reports are usually full

© Springer International Publishing AG 2018
G. De Pietro et al. (eds.), *Intelligent Interactive Multimedia Systems and Services 2017*, Smart Innovation, Systems and Technologies 76, DOI 10.1007/978-3-319-59480-4_13

of misspellings, abbreviations, typos, and acronyms. Furthermore, one of the main tasks of NLP is to acquire contextual information for an accurate interpretation of the extracted entities. This requires the ability to detect associated attributes, negations, temporality, and relation among entities. There are two main approaches for designing NLP systems: a rule-based one, which requires the definition of dictionaries and a set of rules for matching patterns in the text; and a machine learning approach, which relies on learning algorithms to construct classifiers automatically using annotated examples. Rule-based approaches can be very effective, but require manual effort for writing the rules. Machine learning systems require less human effort, but require large annotated training corpora. Most contemporary NLP systems are hybrid, built from a combination of the two approaches [13]. In this work, we present a Medical Language Processing system able to extract relevant medical information from clinical records written in Italian. We have focused on a rule-based approach, mainly because of the lack of existing medical annotated corpora for the Italian language, able to train a learning algorithm. An NLP system in biomedicine includes two main components: a biomedical background knowledge and a framework to manage the NLP pipeline. The Unified Medical Language System (UMLS) constitutes a suitable biomedical knowledge resource in clinical NLP, bringing together several health standards and biomedical vocabularies of concepts. In general, an NLP pipeline includes the following basic components: sentence detector, word tokenizer, part-of-speech tagger, dictionary look-up annotator, and parsing rules to identify meaningful combinations of tokens [17]. The majority of existing NLP systems consider English as their domain language. To answer to the exigency of managing clinical free text in Italian, we propose an NLP system designed to analyze the content of narrative medical records, extracting from them patient medical problems, interested anatomical parts, performed procedures, dispensed devices, and prescribed medications. In addition to this, the system is capable of recognizing negated expressions, capturing concepts modifiers, and underlining useful relations among the entities. Finally, it allows a user to perform elaborated searches and enables to retrieve clinical information, improving the quality of patient care. The rest of the paper is organized as follows. Section 2 investigates the efforts to apply NLP to clinical text. Section 3 presents the system architecture. Section 4 evaluates the system. Finally, Sect. 5 includes indications for future works.

2 Related Work

Automatic detection of relevant entities in clinical documents, such as mentions about symptoms, examinations, diagnoses, and treatments, increases medical knowledge, providing the clinicians with a quick overview of the patient. There are quite a lot of studies concerning clinical entities recognition in English text, but very few studies on clinical text written in other languages [25]. Actually, several clinical NLP systems have been developed for converting unstructured text to structured data. The cTAKES is a modular system combining

rule-based and machine learning techniques, for recognizing medical entities. Its pipeline includes Sentence detector, Tokenizer, Normalizer, Part-of-speech tagger, Shallow parser, NER and negation annotators [24]. The cTAKES showed a high performance in medical concept extraction, making it suitable for successive applications [16,20]. Another widespread clinical NLP system is MedLEE, which aims at generating structured output from patient reports and assigning UMLS codes to relevant clinical information [15,18]. In literature, reviews about information extraction systems from clinical text are available, predominantly built on English. In [23], the authors based their review on several features, such as language, approach, clinical decision support task, and health outcomes, noticing that the majority of the approaches to support clinical decisions have been proposed for processing English free text. The approach presented by Byrd et al. [11] has shown the best results. They aim at identifying Hearth Failure signs and symptoms, through a rule-based NLP system, based on the Unstructured Information Management Architecture (UIMA) framework [5]. Furthermore, an increasing number of NLP research institutes are basing their software development on UIMA specifications [19], which offer a platform for NLP components integration. Moreover, one of the main requirements for developing clinical NLP systems is a suitable biomedical knowledge resource. For this purpose, UMLS offers several vocabularies to help users retrieving information from a wide variety of biomedical information sources [21]. One of the challenges of clinical NER in Italian is that medical terminologies are less extensive for Italian than for English [6]. In biomedicine, NLP offers a powerful instrument for text indexing and document coding. Recognition of relevant medical entities, understanding of their relationships, and, finally, a DB storage, allow powerful information retrieval [12]. In the Italian scenario, there is a small number of studies concerning the extraction of clinical entities from free text. In [10], Attardi et al., considering that Italian corpora annotated with mentions of medical entities are not easily available, have created their own corpus, using a rule-based approach built on regular expressions. They have used the Tanl NER, a statistical sequence labeler, to identify clinical entities, and SVM classifiers, to recognize negations and associations of measures to entities. In other works [8,9], they compensate for the lack of annotated medical Italian resources, creating a silver corpus through a machine translation of an existing one in English. In [7], they propose an unsupervised machine learning methodology for entity and relation extraction, grouping the relations of the same type with the spherical K-means clustering. Finally, Esuli et al. [14] present a solution for extracting a set of concepts of interest from radiological reports, based on a linear-chain CRF learning system, where clauses are the object of tagging. Considering the lack of Italian annotated corpora in biomedicine, we have chosen to base our work on the definition of rich dictionaries to look-up and the implementation of complex parsing rules to match textual patterns of interest. It requires manual efforts, but it can be very effective. Furthermore, annotations deriving from our approach can be employed to train a machine learning system, in the future.

3 System Description

The proposed system is founded on a rule-based approach, which uses manually constructed grammatical rules, built on an NLP pipeline, to extract relevant information from narrative text. The following paragraphs describe pre-processing and processing phases, illustrating technical implementation details.

3.1 Architecture

The implemented system is based on a modular architecture, composed of pre-processing and processing modules, as shown in Fig. 1. The system takes as input the clinical documents, in a PDF format, with information in the form of narrative text. It performs several *pre-processing* operations on each clinical record, obtaining as a result a cleaned and segmented document. In the *processing* operation, the client invokes a web server, by means of Restful APIs, in order to perform the annotation of keywords within the clinical document. Keywords are considered as meaningful words that are extracted from the textual content. The server side consists of a web application that exploits an NLP pipeline connected to a Knowledge Base (KB), which includes several dictionaries, obtained after appropriate *pre-processing* operations. The server sends a response to the client containing the annotation result, structured as an XML file containing the keywords found within the clinical document. In detail, the system invokes the NLP pipeline twice. Firstly, it invokes the pipeline giving as input the cleaned and segmented clinical document. Secondly, it passes to the pipeline a lemmatized document. Lemmatization is a methodical way of converting all the grammatical/inflected forms to the root of the word. Obtaining canonical forms, or lemma, of the words in the clinical record can increase the percentage of matching with words in dictionaries, which mainly contain terms in a base form. The entire process returns two lists of keywords, lemmatized and not, which have to be compared and integrated. The operation of comparison uses a string matching algorithm based on a cosine similarity metric, for measuring the difference between the two sequences. The system integrates the two keywords lists, building a new richer list, containing lemmatized and non-lemmatized entities. Finally, the cleaned/segmented document, the annotated document, and both lists of keywords are stored in a database. In the next subparagraph, we explain the technique used to perform the *pre-processing*. Text pre-processing is an important task and critical step in NLP. It reformats the original text into meaningful units, which contain important linguistic features before performing subsequent text processing strategies. We applied the text *pre-processing* on two different types of documents: health dictionaries and clinical documents.

3.2 Pre-processing

Firstly, this phase aims at creating a suitable Knowledge Base usable by the NLP pipeline. The KB has been derived from the UMLS metathesaurus, considering

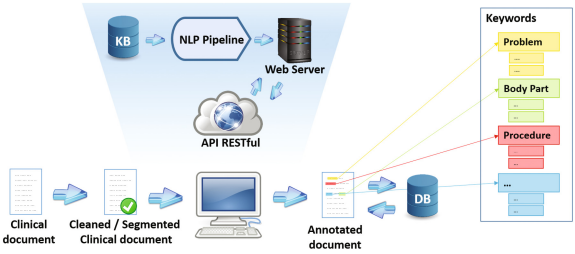

Fig. 1. System architecture.

two types of data files stored in an SQL database: *mrconso*, which contains medical concepts' names and their sources, and *mrsty*, which includes their semantic types. We have selected the following fields of interest, through several queries: CUI (Concept Unique Identifier), a UMLS code for identifying concepts, and STY (Semantic TYpes), a categorization of all concepts. The UMLS information sources, in their Italian translation, used for the creation of the custom dictionaries, are the following: MedDRA, the *Medical Dictionary for Regulatory Activities*; MeSH, the *Medical Subject Headings*; ICPC, the *International Classification of Primary Care*; MTHMST, the *Metathesaurus Minimal Standard Terminology Digestive Endoscopy*. Based on the semantic types and sources of interest, the results of the queries have been integrated in five medical dictionaries. In order to obtain atomic dictionaries and to increase the percentage of matched keywords, a normalization of the dictionaries have been actualized, including: replacement of the accented characters, removal of brackets and their contents, transformation of the terms in lowercase to be case insensitive, and a stop word removal. The resulting custom dictionaries refer to the following semantic types of UMLS: **Problem** includes *Anatomic Abnormality*, *Disease or Syndrome*, *Sign or Symptom*, etc.; **Body** includes *Body Part*, *Organ or Organ Component*, *Body System*, etc.; **Procedure** includes *Diagnostic Procedure*, *Health Care Activity*, *Therapeutic or Preventive Procedure*, etc.; **Device** includes *Medical Device* and *Research Device*; **Medication** includes *Pharmacologic Substance* enriched with drugs dictionary of the *Italian Medicines Agency* (AIFA). Secondly, the system performs a *pre-processing* on the clinical document given in input to the system. It is able to detect errors and suggest corrections respect to a dictionary, using a spell check based on the Symmetric Delete Spelling Correction algorithm (SymSpell) [1]. To allow noise minimization, the spell check operation is not automatic: it lets the user choose whether to replace or not the word gradually found.

3.3 Processing

The *processing* operation, which is the core of the system, refers to the pipeline used to extract relevant entities from clinical documents. It is shown in Fig. 2. The first stage is the ***language identification***, useful to discover the language of the text. Then it actualizes a ***tokenization*** of the text. In particular, the text

is broken up into words, phrases, symbols, or other meaningful elements called tokens. At the end of the tokenization process, a ***lexical analysis*** is implemented. This process marks up a word in a text with a representative part of speech tagging (POS-tag), based on built-in dictionaries and custom dictionaries. Through the built-in dictionaries look-up, the lexical analysis component assigns a grammatical label to the tokens, such as pronoun, verb, noun, adjective, etc. Moreover, via the custom dictionaries look-up, the lexical analysis component identifies *problem*, *procedure*, *body part*, *device*, and *medication* terms inside the text. Based on the lexical analysis tagging, a set of parsing rules has been defined to identify meaningful combinations of tokens in the document.

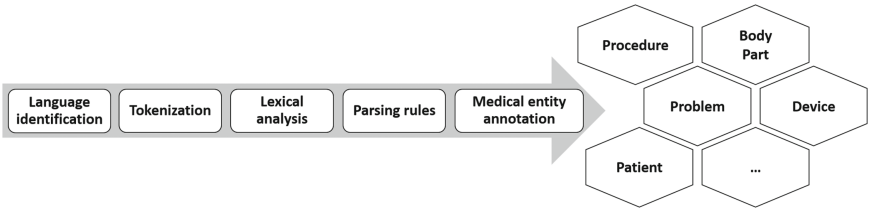

Fig. 2. NLP pipeline.

The ***parsing rules*** define a sequence of annotations that indicates something of interest inside the text. This stage examines annotations sequence created by preceding stage and selects token combinations that constitute meaningful expressions. Consider the following sentence as an example, "*Si evidenzia una lesione frammentata a livello del menisco mediale*", with its English translation, "*A fragmented lesion of the medial meniscus is appreciable*". It includes a mention about the problem "*lesione frammentata*", and a mention about the body part "*menisco mediale*". In detail, these concepts are fired by the following parsing rules: (i) *Problem = Problem dictionary + Adjective (0 or more) + group [Noun (0 or more) + Adjective (0 or more) + Verb (0 or more)]*; (ii) *Body = group [Body dictionary + Adjective (0 or more) + Adverb (0 or 1) + Verb (0 or 1)]*; (iii) *Body Part/Problem relation = group [Body + Problem]*. The first rule matches the problem *lesion* associated to its qualifier, the adjective *fragmented*. The second one matches the anatomical part. Finally, the third rule is able to connect the previous ones, finding the important relation between the problem and the involved body part. These are just some of the implemented parsing rules. Finally, there is the ***medical entity annotation*** block, which highlights medical entities in the text.

3.4 Technical Details

For the actualization of the NLP pipeline, the environment IBM Watson Explorer Content Analytics Studio© (ICA) ver. 11.0 has been used, in consideration of its capabilities of analyzing unstructured text [2]. It is based on the Apache UIMA

framework. Using the ICA, we have built an NLP pipeline to extract medical entities from free text. ICA allows configuring the following key annotators for text analysis: *Dictionary Lookup annotator*, which matches words from a dictionary with words in the text; Pattern Matcher annotator, which identifies patterns in the text, i.e. sequences of words, by using defined rules. Firstly, the pipeline language detection has been manually set to the Italian language, for removing any language ambiguity. Subsequently, Lexical Analysis parses the structure of the sentence and tags words, looking-up built-in dictionaries and custom dictionaries. During the creation of the custom dictionaries, the ICA studio allows the *generating inflections operation*, useful to add new surface forms and to match the word in different forms. Moreover, thanks to the possibility of adding features to the dictionaries, we have added a column containing the UMLS identifiers (CUI), in order to preserve the concept-code association. Furthermore, ICA Studio provides an interface with the ability to (i) create sophisticated parsing rules and (ii) specify matching criteria for the *Pattern Matcher annotator*. The matching criteria are based on tokens, dictionaries terms, and existing annotations. Subsequently, we have exported the pipeline as a PEAR file (Processing Engine ARchive), which is a fully compliant UIMA annotator, containing descriptor files, compiled classes, configuration files, jar files, and libraries. The PEAR file has been integrated into a Java EE web application, loaded on the Apache Tomcat application server. Furthermore, for ensuring the retrieval of annotated concepts, we have indexed and stored the processed documents, employing MongoDB [3], an open source non-relational database, oriented to the documents, and based on a JSON-style representation. This results in a very fast execution of queries on documents. Moreover, it offers the possibility to achieve a semantic search of information contained within a document, performing a stop word removal and a Snowball stemming [4]. Finally, the system user interface has been designed through the JavaFX framework, which uses Java programming language, to design, test, and deploy rich client applications.

4 System Evaluation

4.1 Use Case

A clinician can employ our pipeline to extract information, increasing his/her medical knowledge and acquiring patient data in a quick and easy way. In detail, our system provides a simple and intuitive user interface, admitting the following operations: (i) simultaneous annotation of one or more documents, (ii) clear displaying of relevant medical information, (iii) semantic search of annotated concepts. The user can select any number of reports to extract entities from them. Then, the system will inform him/her of the success of the operation. Furthermore, the user can display a single annotated report, to obtain a rapid overview of relevant data contained in the unstructured text. In particular, relevant concepts are underlined and a table summarizes recognized medical concepts, their associated UMLS codes, and relations among several entities, such as between a problem and the interested body part, or between a medication and its dosage.

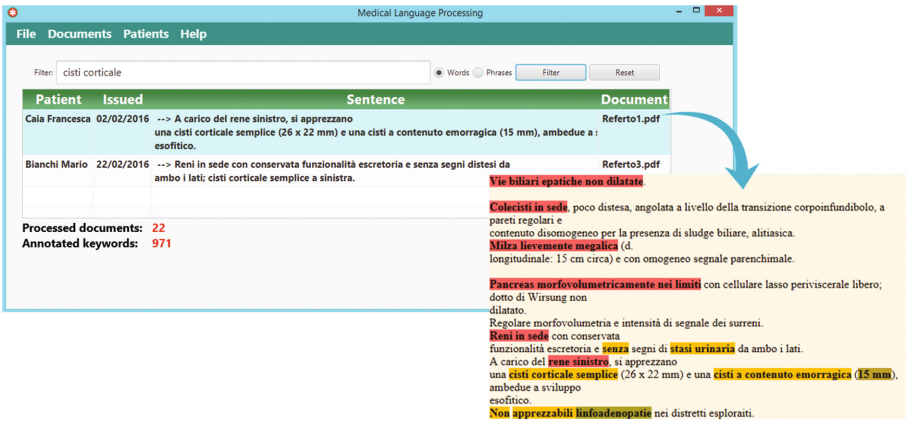

Fig. 3. Semantic search of medical concepts with the related annotated document.

Finally, Fig. 3 shows an operation that can turn to be very useful to a clinician, consisting in the semantic search of a concept among the documents previously annotated. This can support the process and the quality of patient care, due to the availability of a large amount of extracted data that can increase clinician medical knowledge.

Clearly, the knowledge base may be continuously enriched, to increase annotated entities. Furthermore, due to the definition of complex parsing rules, the annotator achieves the following results:

- Capturing negative expressions, thanks to the matching criterion requiring a medical entity and a term coming from a dictionary of common negations;
- Associating qualifiers to medical entities, due to a parsing rule that requires a medical entity combined with a certain type of adjectives or adverbs;
- Obtaining useful temporal expressions, through a parsing rule that matches date triggers of interest and a regular expression able to validate date format;
- Capturing quantifiers information, using a rule that links a medical concept to a regular expression able to match observation results;
- Underlining relations among entities, through a rule that combines entities from the *Problem* and *Body Part* dictionaries in a sentence.

In addition, we have characterized specific rules, not easily generalizable, to recognize the *Patient*, subject of the report, and the *Practitioner*, author of it. Therefore, we have defined dictionaries to find medical entities of interest and parsing rules to better understand their semantic meaning. General errors occur especially because of the complexity of the Italian medical language. To improve the annotations fired by the parsing rules, a procedure of iterative refinement, supervised by a domain expertise, may be adopted.

5 Conclusions and Future Works

This study has been performed on Italian clinical records, written as free text, with the aim of automatically extracting medical entities from them. In particular, we have proposed a system implementing an NLP pipeline, which recognizes keywords of interest in the narrative, extracts meaningful relations, and stores the annotated documents. The indexing of the records allows the retrieval of useful clinical information, increasing the medical knowledge and improving the patient quality of care. The recognition of the entities is mainly based on a dictionaries look-up, while, due to several parsing rules, it is possible to recognize negated mentions of the concepts, measurements associated with the keywords, and relations among them. We aim at improving the knowledge base that constitutes the dictionaries and the rules that have to match meaningful expressions. Moreover, our intent is to employ clinical documents annotated with our rule-based approach for training a machine learning system. Starting from the medical knowledge extracted from unstructured data, several applications can be developed. In the future, our intent is matching stored annotated keywords against a standard medical terminology. We plan to define a tool, based on the implemented system, able to map extracted entities regarding patient's medical conditions with the International Classification of Diseases (ICD). Furthermore, we aim at implementing an application that makes exhaustive and agile the research of information regarding a patient. In particular, we aspire at extracting information from the Patient Summary type of electronic record, a collection of the patient's most significant clinical data. Based on the implemented system, relevant medical entities could be extracted and presented to a clinician, offering a rapid patient overview.

References

1. FAROO spelling correction (2016). http://blog.faroo.com/category/spelling-correction/
2. IBM watson explorer (2016). https://www.ibm.com/us-en/marketplace/content-analytics
3. Mongo database (2016). https://www.mongodb.com/
4. Snowball resources (2016). http://snowball.tartarus.org/
5. UIMA home (2016). https://uima.apache.org/
6. UMLS documentation (2016). https://www.nlm.nih.gov/research/umls/
7. Alicante, A., Corazza, A., Isgrò, F., Silvestri, S.: Unsupervised entity and relation extraction from clinical records in italian. Comput. Biol. Med. **72**, 263–275 (2016)
8. Attardi, G., Cozza, V., Sartiano, D.: Adapting linguistic tools for the analysis of Italian medical records (2014)
9. Attardi, G., Cozza, V., Sartiano, D.: UniPi: Recognition of mentions of disorders in clinical text. In: Proceedings of the 8th International Workshop on Semantic Evaluation, pp. 754–760 (2014)
10. Attardi, G., Cozza, V., Sartiano, D.: Annotation and extraction of relations from Italian medical records. In: IIR (2015)

11. Byrd, R.J., Steinhubl, S.R., Sun, J., Ebadollahi, S., Stewart, W.F.: Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. Int. J. Med. Informatics **83**(12), 983–992 (2014)
12. De Bruijn, B., Martin, J.: Getting to the (c)ore of knowledge: mining biomedical literature. Int. J. Med. Informatics **67**(1), 7–18 (2002)
13. Doan, S., Conway, M., Phuong, T.M., Ohno-Machado, L.: Natural language processing in biomedicine: a unified system architecture overview. In: Clinical Bioinformatics, pp. 275–294 (2014)
14. Esuli, A., Marcheggiani, D., Sebastiani, F.: An enhanced CRFs-based system for information extraction from radiology reports. J. Biomed. Inform. **46**(3), 425–435 (2013)
15. Friedman, C., Shagina, L., Lussier, Y., Hripcsak, G.: Automated encoding of clinical documents based on natural language processing. J. Am. Med. Inform. Assoc. **11**(5), 392–402 (2004)
16. Garla, V., Re, V.L., Dorey-Stein, Z., Kidwai, F., Scotch, M., Womack, J., Justice, A., Brandt, C.: The yale cTAKES extensions for document classification: architecture and application. J. Am. Med. Inform. Assoc. **18**(5), 614–620 (2011)
17. Hardeniya, N.: NLTK Essentials. Packt Publishing Ltd. (2015)
18. Johnson, S.B., Bakken, S., Dine, D., Hyun, S., Mendonça, E., Morrison, F., Bright, T., Van Vleck, T., Wrenn, J., Stetson, P.: An electronic health record based on structured narrative. J. Am. Med. Inform. Assoc. **15**(1), 54–64 (2008)
19. Kunze, M., Rösner, D.: UIMA for NLP based researchers workplaces in medical domains. In: Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP, p. 20 (2008)
20. Lin, C.H., Lai, W.S., Lee, L.H., Tsao, H.M., Liou, D.M.: An entry generation pipeline for converting free-text medical document into clinical document architecture document with entry-level. In: 2014 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), pp. 505–508. IEEE (2014)
21. McCray, A.T., Aronson, A.R., Browne, A.C., Rindflesch, T.C., Razi, A., Srinivasan, S.: UMLS knowledge for biomedical language processing. Bull. Med. Libr. Assoc. **81**(2), 184 (1993)
22. Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., Hurdle, J.F., et al.: Extracting information from textual documents in the electronic health record: a review of recent research. Yearb. Med. Inform. **35**(128), 44 (2008)
23. Reyes-Ortiz, J.A., González-Beltrán, B.A., Gallardo-López, L.: Clinical decision support systems: a survey of NLP-based approaches from unstructured data. In: 2015 26th International Workshop on Database and Expert Systems Applications (DEXA), pp. 163–167. IEEE (2015)
24. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J. Am. Med. Inform. Assoc. **17**(5), 507–513 (2010)
25. Skeppstedt, M., Kvist, M., Nilsson, G.H., Dalianis, H.: Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: an annotation and machine learning study. J. Biomed. Inform. **49**, 148–158 (2014)