THE TOOLS AND RESOURCES FOR CLINICAL TEXT PROCESSING

Ulfeta Marovac*,
Aldina Avdić

Department of Technical Science,
State University of Novi Pazar,
Novi Pazar, Serbia

Abstract:
The expansion of electronic health has produced a large amount of information in medical information systems in a structured and unstructured format. The processing of unstructured data in the form of text is performed using natural language processing techniques. Natural language processing requires specific resources for processing text depending on the domain of the text as well as the language in which it is written. This paper aims to present the available tools, lexical resources, and corpora used in the analysis of clinical texts.

Keywords:
natural language processing, clinical texts, lexical resources.

INTRODUCTION

Most countries today use e-health and keep large amounts of data about patients and their illnesses that are usually used to manage individual cases as well as administration. Data stored in medical information systems can be structured and unstructured. Unstructured data contain free text such as anamnesis, radiological reports, patient care notes, and other similar clinical texts. This data carries information that is important not only for resolving an individual case of the disease but also for extracting general information. Health care is one of the top priorities of every state. We are witnessing that diseases have no borders and language barriers and that the fast and efficient availability of data can save people's lives. Most developed countries are seriously processing information from clinical texts to improve health, but most of the research is for the English-speaking world. Creating standards for storing medical data in a multilingual system could contribute to the faster development of medicine. If there was a single platform through which COVID-19 patients would be monitored, it would be possible to get information

Correspondence:
Ulfeta Marovac

e-mail:
umarovac@np.ac.rs

DOI: 10.15308/Sinteza-2021-134-140

about places of potential hotspots of infection [1], clinical characteristics and prognostic factors using natural language processing methods [2]. The ability to process reports of patients treated in countries where English is not an official language allows for global aggregation of data, which is extremely important especially for rare diseases [3]. This paper provides an overview of available resources for processing clinical texts in different natural languages. The paper is organized into six sections. The second section presents a description of the basic concepts on which the work is based. The third section shows the processing of the clinical text. An overview of the corpora used for different languages is given in the fourth section. The fifth section contains tools for medical classification and annotation of clinical texts. The last section contains the conclusion and directions for further research.

2. THE BASIC CONCEPTS

Introduce the basic concepts of clinical text processing that are the main points of consideration in this paper: clinical text, electronic medical reports, natural text processing, clinical text processing, resources for clinical text processing, and medical classification.

Clinical texts are written by doctors, medical staff, and other health care providers. They are used to document the patient's condition and the health services provided. They describe patients, their pathologies, their personal, social, and medical history. Clinical texts differ from scientific texts and are not prepared for publication. They do not use complete sentences, they use medically accepted expressions and abbreviations that do not belong to the natural language in which they were written.

Electronic health records (EHR) carry a lot of important information such as the patient's condition on admission to the hospital, the course of his recovery, then the state of health at discharge. This information is still most easily expressed in natural language, which makes information extraction more difficult [4]. Specific medical terminology is defined by different standards and classification systems. Classification and descriptions of diseases, treatments, and drugs control the vocabulary used in medical reports and administration reduce ambiguity and increase the degree of clarity.

Natural language processing (NLP) is a field of linguistics, computing, and artificial intelligence that explores ways in which computers can understand and use text or speech in natural language and apply them to some useful activities [5].

Clinical text mining is the extraction of information from clinical texts [6].

Resources for processing clinical text are all data sets that help in research, and they can be: sets of diagnoses, symptoms, drugs as well as clinical corpora.

Medical classification and terminology are classification systems and terms used in reports, administration, classification and description of diseases, treatments, and medications such as ICD coding diagnosis, SNOMED CT, MeSH, UMLS, ATC, and others [6].

3. THE CLINICAL TEXT PROCESSING

Clinical texts represent the basic form of communication between healthcare professionals. Using natural language processing methods, it is possible to extract information from these texts that are hidden in free text and which are not easily usable for computer analyzes. Most of the authors are engaged in the analysis of data from clinical texts written in English due to their public availability as well as the public availability of clinical text processing tools for English. Two approaches are used to process natural language in clinical texts: a rule-based approach; and machine learning algorithms. The first approach requires the existence of specialized clinical dictionaries that support complex clinical logic such as the MTERMS tool [7].

The second approach based on machine learning requires a set of manually annotated clinical data. An overview of the use of machine learning over clinical texts until 2020 is given in [8], where the results of 110 research available on PubMed from the period from 2015 to 2018, which concern the machine processing of clinical texts in English, are presented. Examining the properties of the data used, it was concluded that most of the research used hundreds or thousands of documents.

There is a small number with a very small data set of less than 50 and a very large of 10,000 documents (ten papers). Many of the data, although available, remained unused. The main reason for unused data is that they are not marked. If the data annotation is done manually then it is a hard job and prone to errors. Active learning algorithms enable the processing of documents even with a smaller number of manually annotated data, whereby new annotations use more algorithms and compare their results. Existing structured data are often used for annotation, so the textual part of the medical report can be labeled using the diagnosis code [9].

Semi-automated annotation can also be used. Often the data being processed comes from one institution, so the relevance of that data is questionable. Very often the results published on one data set did not give the same results on another set [10]. The clinical application of such data processing is wide from diagnostics, prognosis, protection, risk prediction, improvement of service provision, management, etc.

4. AVAILABLE CLINICAL CORPUS

It is very difficult to reach the clinical corpus due to the sensitivity of the data it contains. Each clinical corpus must have ethical approval for use. The data must pass the deidentification process to preserve patient privacy by taking into account names and identification numbers, telephone numbers, and addresses. There are several corpora available for both English and other languages.

For English, more data sets are annotated and consist of discharge lists, medical histories, nursing reports, radiological reports, sentences from the medical domain, and other medical reports. Some of them are available:

- Informatics for Integrating Biology & the Bedside (i2b2) [11]
- Computational Medicine Center (CMC) corpus [12]
- ShARe/CLEF eHealth [13]
- Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) [14]
- BioScope Corpus [15].

There are clinical corpora in non-English languages, but they are smaller and cover fewer different medical contents. The number of publications in the field of natural text processing in different languages in PubMed is shown in the graph (Figure 1).

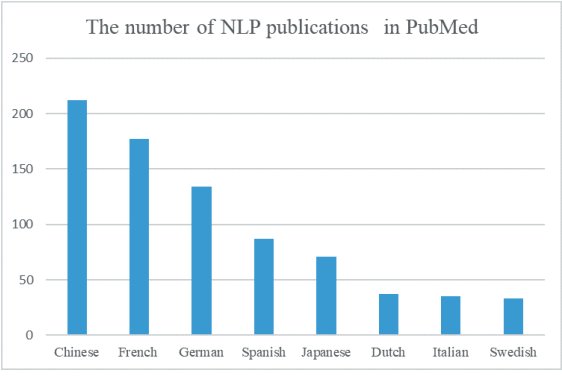


Figure 1 - PubMed's publication for query: "natural language processing and (French| German| Chinese| Spanish| Japanese| Dutch| Italian| Swedish) ".

Non-English corpora are commonly from the researcher's institutions and require special permits and contacts for their further use. These corpora contain labels for: diagnoses, symptoms, medications, and therapies. Some of the corpora of clinical texts in non-English languages used in the scientific papers are presented in Table 1.

Language	Description	No. of texts	Ref.
Spanish	discharge reports	142 154	[16]
Bulgarian	clinical texts	100 000 000	[17]
Bulgarian	outpatient records (diabetic)	500 000	[18]
Serbian	medical records (B05)	5000	[19]
Serbian	medical reports	4212	[20]
Serbian	medical documents	200	[21]
Swedish	Stockholm EPR Corpus	2 000 000	[22]
Danish	clinical narrative text	61000	[23]
Dutch	EMC Dutch clinical corpus	-	[24]
Finnish	intensive care nursing narratives	2800	[25]
French	clinical texts	170 000	[26]
Italian	clinical texts	23 695	[27]
Italian	clinical texts	100	[28]
German	clinical texts	18 000	[29]
German	leukemia laboratory results	12 743	[30]
German	nephrology records	6 817	[31]
German	discharge reports	118	[31]
Chinese	medical documents	1100	[32]

Table 1 - Non-English corpora

The lack of appropriate lexical resources can be solved by applying unsupervised methods [33].

5. TOOLS FOR MEDICAL CLASSIFICATION AND ANNOTATION OF CLINICAL TEXTS

Medical terminology and classification systems are used in healthcare to facilitate interoperability among institutions and to collaborate, medical professionals, scientists, and other stakeholders. There is a justified need to integrate the various medical terminology and classification.

International Statistical Classification of Diseases and Related Health Problems (ICD) has been used since the 18th century with constant revisions and additions. It is used in over 150 countries and is available in more than 40 languages and it is under the jurisdiction of World Health Organization [34]. Classification of diseases is a system of categories that are assigned to certain diseases according to defined criteria. The International Classification of Diseases is a standard tool used in epidemiology, health management, the analysis of population health, and monitoring health problems.

SNOMED CT [35] is a structured clinical vocabulary used in any electronic health record (EHR). It is the most comprehensive and accurate clinical health terminology product in the world.

It provides that data can be shared between health and social care institutions and service providers. SNOMED CT is available in American English, English, Argentine Spanish, Danish and Swedish. Translations into French, Dutch, Lithuanian, and several others. SNOMED CT is clinical hierarchical terminology that contains medical terms and their relationships as well as synonyms, including over 320,000 terms (Figure 2, Figure 3).

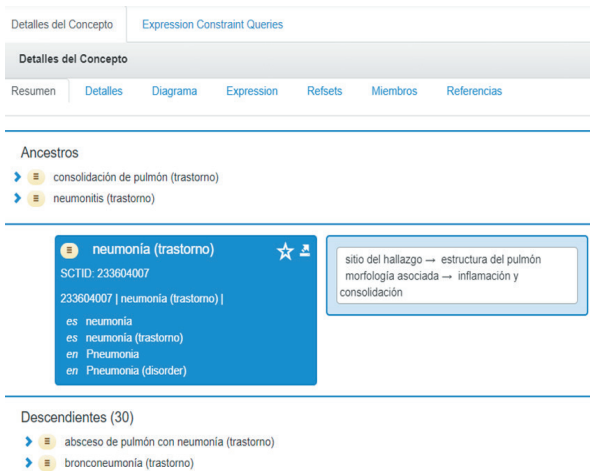


Figure 2 - SNOMED CT Spanish edition 2020

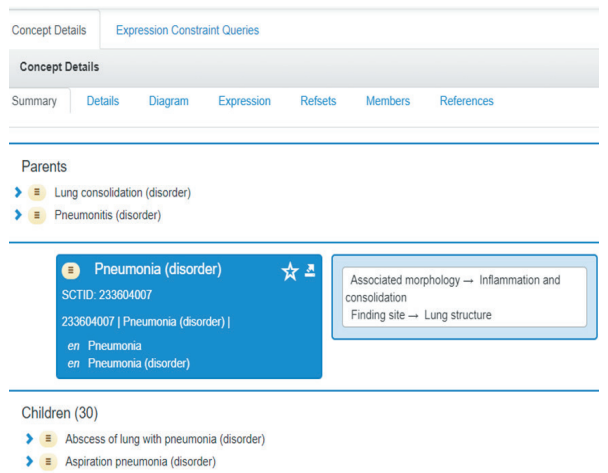


Figure 3 - SNOMED CT United States edition 2020

UMLS (Unified Medical Language Systems) [36] integrates and distributes key terminology, classification and coding standards, and associated resources to promote the creation of more effective and interoperable biomedical information systems and services, including electronic health records.. UMLS supports mapping between different terminologies. UMLS contains several million concepts derived from hundreds of bio (medical) dictionaries, such as ICD, SNOMED, OMIM, MeSH, GO, as well as medical abbreviations (Figure 4). It consists of three parts:

1. Metathesaurus - very large, multi-purpose, and multi-lingual vocabulary database that contains information about biomedical and health-related concepts
2. Semantic Networks - categorization and connections between all resources in metathesaurus
3. Specialized Lexicons - lexicons for biomedical and general English.

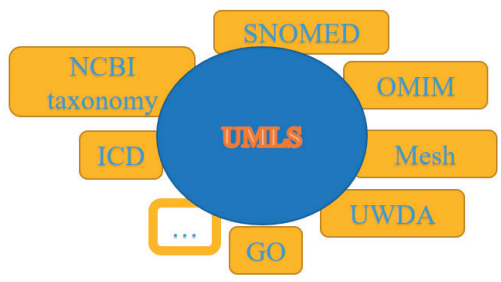


Figure 4 – UMLS concepts

Metathesaurus contains 215 different lexicons for 25 languages, of which the most are resources for English (144), followed by German, Spanish and French (Figure 5). There is interest in constantly updating these resources for different languages.

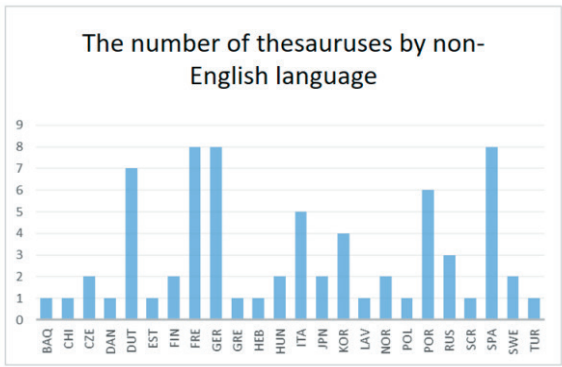


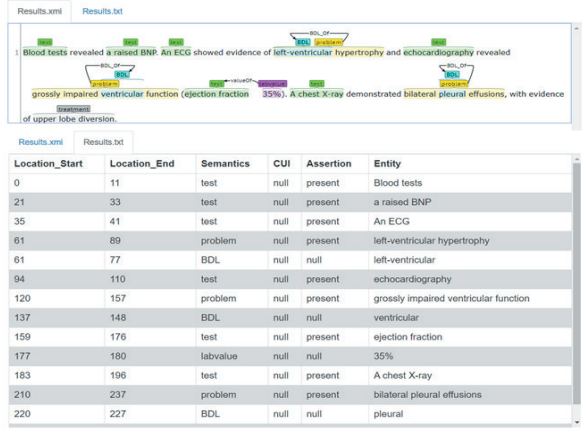
Figure 5 – The number of thesauruses by different languages

The most commonly used UMLS products are metathesauri, followed by MetaMap [37] which is used to map concepts from metathesaurus in the text. The creation of reference corpora is crucial in the process of developing appropriate methods for solving the problems of machine translation, deidentification, drug interaction, etc. [38] [39].

The three most popular information retrieval tools are MetaMap [40], cTAKES [41] and CLAMP [42]. The common feature of these tools is to perform mapping named entity based on UMLS. MetaMap is a tool for extracting biomedical information. cTAKES is a natural language processing system for extracting data from clinical free text from electronic medical records using machine-based rules. It contains all the basic functions of NLP processing for the English language, such as tokenizer, POS tagger, named entity recognizer, negation detection, machine learning functionality, etc. The latest NLP tool for clinical text CLAMP has greater flexibility in the development of custom schemes with the possibility of their application for information retrieval. CLAMP is a Java tool, it has built-in natural language processing modules for English text. By comparing these tools in [43], it was shown that CLAMP has the best performance in terms of F1 results, and higher accuracy, and slightly lower recall compared to cTAKES and MetaMap.

Figure 6 shows an example of the application of the CLAMP tool on the example of an EHR medical report in English:

"Blood tests revealed a raised BNP. An ECG showed evidence of left-ventricular hypertrophy and echocardiography revealed grossly impaired ventricular function (ejection fraction 35%). A chest X-ray demonstrated bilateral pleural effusions, with evidence of upper lobe diversion."



Location_Start	Location_End	Semantics	CUI	Assertion	Entity
0	11	test	null	present	Blood tests
21	33	test	null	present	a raised BNP
35	41	test	null	present	An ECG
61	89	problem	null	present	left-ventricular hypertrophy
61	77	BDL	null	null	left-ventricular
94	110	test	null	present	echocardiography
120	157	problem	null	present	grossly impaired ventricular function
137	148	BDL	null	null	ventricular
159	176	test	null	present	ejection fraction
177	180	labvalue	null	null	35%
183	196	test	null	present	A chest X-ray
210	237	problem	null	present	bilateral pleural effusions
220	227	BDL	null	null	pleural

Figure 6 - CLAMP tool application example

Figure 6 shows the .xml and .txt result of mapping different medical entities in the text such as: tests, symptoms, different laboratory analyzes, and more.

6. CONCLUSION

By analyzing the existing resources for processing clinical texts in different natural languages, it can be concluded that most resources and tools are made for the English language. Great efforts are being made to create tools for other natural languages as well. The specific tools for processing the appropriate natural language are needed to be able to process clinical texts, as well as lexicons of medical terminology in the appropriate language. Some of our future goals are to create appropriate resources for the Serbian language.

7. ACKNOWLEDGEMENTS

This paper is partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia under projects III44007 and ON 174026.

REFERENCES

[1] T. Varsavsky, M. S. Graham, L. S. Canas, S. Ganesh, J. C. Pujol, C. H. Sudre, ... and S. Ourselin, " Detecting COVID-19 infection hotspots in England using large-scale self-reported data from a mobile application: a prospective, observationa," *The Lancet Public Health*, vol. 6, no. 1, pp. e21-e29, 2021.

[2] J. L. Izquierdo, J. Ancochea, S. C.-1. R. Group and J. B. Soriano, "Clinical Characteristics and Prognostic Factors for Intensive Care Unit Admission of Patients With COVID-19: Retrospective Study Using Machine Learning and Natural Language Processing," *J Med Internet Res*, vol. 22, no. 10, p. e21801, 2020.

[3] C. Kothari, M. Wack, C. Hassen-Khodja, S. Finan, G. Savova, M. O'Boyle, ... and P. Avillach, "Phelan-McDermid syndrome data network: Integrating patient reported outcomes with clinical notes and curated genetic reports," *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, vol. 177, no. 7, pp. 613-624, 2018.

[4] C. Safran, C. Chute and J. R. Scherrer, "Natural Language and Medical Concept Representation," in *Preprints of the IMIA WG6 Conference,* Vevey, 1994.

[5] K. R. Chowdhary, "Natural language processing," *Fundamentals of Artificial Intelligence*, pp. 603-649, 2020.

[6] D. Hercules, Clinical text mining: Secondary use of electronic patient records, Springer Nature, 2018.

[7] L. Zhou, J. M. Plasek, L. M. Mahoney, F. N. Karipineni, X. Y. Chang, ... and R. A. Rocha, "Using Medical Text Extraction, Reasoning and Mapping System (MTERMS) to process medication information in outpatient clinical notes," in *AMIA Annual Symposium,* 2011.

[8] I. S. a. G. Nenadic, "Clinical text data in machine learning: Systematic review," *JMIR Medical Informatics*, vol. 8, no. 3, p. e17984, 2020.

[9] S. Horng, D. A. Sontag, Y. Halpern, Y. Jernite, N. I. Shapiro and L. A. Nathanson, "Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning," *PLoS One 2017*, vol. 12, no. 4, 2017.

[10] S. J. Fodeh, D. Finch and L. Bouayad, "Classifying clinical notes with pain assessment using machine learning," *Medical & Biological Engineering & Computing*, vol. 56, no. 7, pp. 1285-1292., 2018.

[11] "i2b2: Informatics for Integrating Biology & the Bedside," [Online]. Available: https://www.i2b2.org/.

[12] Ö. Uzuner, X. Zhang and T. Sibanda, "Machine learning and rule-based approaches to assertion classification," *Journal of the American Medical Informatics Association*, vol. 16, no. 1, pp. 109-115, 2009.

[13] H. Suominen, S. Salanterä, S. Velupillai, W. W. Chapman, G. Savova, N. Elhadad, ... and G. Zuccon, "Overview of the ShARe/CLEF eHealth evaluation lab 2013," in *International Conference of the Cross-Language Evaluation Forum for European Languages,* Berlin, 2013.

[14] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L. W. Lehman and G. Moody, "Multiparameter intelligent monitoring in intensive care II (MIMIC-II): A public-access intensive care unit database," *Critical Care Medicine*, vol. 39, no. 5, p. 952, 2011.

[15] V. Vincze, G. Szarvas, R. Farkas, G. Móra and J. Csirik, "The BioScope Corpus: Biomedical texts annotated for uncertainty, negation and their scopes," *BMC Bioinformatics*, vol. 9, no. 11, p. s9, 2008.

[16] M. Oronoz, K. Gojenola, A. Pérez, A. D. d. Ilarraza and A. Casillas, "On the creation of a clinical gold standard corpus in spanish: mining adverse drug reactions," *J. Biomed. Inform.* , vol. 56, p. 318–332, 2015.

[17] S. Boytcheva, G. Angelova, Z. Angelov and D. Tcharaktchiev, "Text mining and big data analytics for retrospective analysis of clinical texts from outpatient care," *Cybernetics and Information Technologies,* vol. 15, no. 4, pp. 58-77, 2015.

[18] S. Boytcheva, I. Nikolova, G. Angelova and Z. Angelov, "Identification of risk factors in clinical texts through association rules," in *Proceedings of RANLP Workshop on Biomedical Natural Language Processing*, 2017.

[19] A. R. Avdić, U. A. Marovac and D. S. Janković, "Normalization of Health Records in the Serbian Language with the Aim of Smart Health Services Realization," *Facta Universitatis, Series: Mathematics and Informatics*, pp. 825-841, 2020.

[20] A. Avdic, U. Marovac and D. Jankovic, "Automated labeling of terms in medical reports in Serbian," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 28, no. 6, pp. 3285-3303, 2020.

[21] J. Jacimovic, K. C. and D. Jelovac, "A Rule-Based System for Automatic De-identification of Medical Narrative Texts," *Informatica (Slovenia)*, p. 39, 2016.

[22] H.Dalianis, A. Henriksson, M. Kvist and S. Velupillai, "HEALTH BANK – A workbench for data science applications in healthcare," in *Proceedings of the CAiSE-2015 Industry Track Co-located with 27th Conference on Advanced Information Systems Engineering (CAiSE 2015)*, Stockholm, 2015.

[23] R. Eriksson, P. B. Jensen, S. Frankild, L. J. Jensen and S. Brunak, "Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text," *Journal of the American Medical Informatics Association*, vol. 20, no. 5, pp. 947-953, 2013.

[24] [Online]. Available: https://biosemantics.erasmusmc.nl/index.php/resources/emc-dutch-clinical-corpus.

[25] [Online]. Available: http://bionlp.utu.fi/clinicalcorpus.html.

[26] L. Campillos, L. Deléger, C. Grouin, T. Hamon, A. L. Ligozat and A. Névéol, "A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT)," *Language Resources and Evaluation*, vol. 52, no. 2, pp. 571-601, 2018.

[27] E. Chiaramello, F. Pinciroli, A. Bonalumi, A. Caroli and G. Tognola, "Use of 'off-the-shelf' information extraction algorithms in clinical informatics: a feasibility study of MetaMap annotation of Italian medical notes," *J. Biomed. Inform*, vol. 63, pp. 22-32, 2016.

[28] G. Attardi, V. Cozza and D. Sartiano, "Annotation and extraction of relations from Italian medical records," in *In Proceedings of the 6th Italian Information Retrieval Workshop*, Cagliari, 2015.

[29] S. Spat, B. Cadonna, I. Rakovac, C. Gütl, H. Leitner and S. G., "Enhanced information retrieval from narrative German-language clinical text documents using automated document classification," *Studies in Health Technology and Informatics*, vol. 136, p. 473, 2008.

[30] M. Zubke, "Classification based extraction of numeric values from clinical narratives," in *In Proceedings of RANLP Workshop on Biomedical Natural Language Processing*, 2017.

[31] R. Roller, F. X. H. Uszkoreit, L. Seiffe, M. Mikhailov and O. Staeck, "A fine-grained corpus annotation schema of German nephrology records," in *In Proceedings of the Clinical Natural Language Processing Workshop*, Osaka, 2016.

[32] B. He, B. Dong, Y. Guan, J. Yang, Z. Jiang, Q. Yu, ... and C. Qu, "Building a comprehensive syntactic and semantic corpus of Chinese clinical texts," *Journal of biomedical informatics*, vol. 69, pp. 203-217, 2017.

[33] A. Alicante, A. Corazza, F. Isgrò and S. Silvestri, "Unsupervised information extraction from Italian clinical records," in *Proceeding of Innovation in Medicine and Healthcare*, 2014.

[34] "International Statistical Classification of Diseases and Related Health Problems," [Online]. Available: https://www.icd10data.com.

[35] [Online]. Available: https://www.snomed.org/.

[36] O. Bodenreider, "The unified medical language system (UMLS): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. 1, pp. D267-D270, 2004.

[37] [Online]. Available: https://metamap.nlm.nih.gov/.

[38] C. Grouin, T. Lavergne and A. Névéol, "Optimizing annotation efforts to build reliable annotated corpora for training statistical models," in *In: 8th Linguistic Annotation Workshop – LAW VIII*, 2014.

[39] M. K. G. N. a. H. D. M. Skeppstedt, "Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study," *Journal of Biomedical Informatics*,, p. 148–158, 2014.

[40] A. R. Aronson and F. M. Lang, "An overview of MetaMap: historical perspective and recent advances," *J Am Med Inform Assoc.*, vol. 17, no. 3, p. 229–236, 2010.

[41] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler and C. G. Chute, "Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications," *J Am Med Inform Asso*, vol. 17, no. 5, pp. 507-513, 2010.

[42] E. Soysal, J. Wang, M. Jiang, Y. Wu, S. Pakhomov, H. Liu and H. Xu, "CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines," *J Am Med Inform Assoc,* vol. 25, no. 3, p. 331–336, 2018.

[43] J. Peng, M. Zhao, J. Havrilla, C. Liu, C. Weng, W. Guthrie, ... and Y. Zhou, "Natural language processing (NLP) tools in extracting biomedical concepts from research articles: a case study on autism spectrum disorder," *BMC Medical Informatics and Decision*, vol. 20, no. 11, 2020.