A supervised named entity recognition for information extraction from medical records

Darko Puflović*, Goran Velinov**, Tatjana Stanković*, Dragan Janković*, Leonid Stoimenov*
* Faculty of Electronic Engineering, University of Niš, 18000 Niš, Serbia
**Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje
*{darko.puflovic, tatjana.stankovic, dragan.jankovic, leonid.stoimenov}@elfak.ni.ac.rs
**goran.velinov@finki.ukim.mk

***Abstract*— Named entity recognition is a widely used task to extract various kinds of information from unstructured text. Medical records, produced by hospitals every day contain huge amount of data about diseases, medications used in treatment and information about treatment success rate. There are a large number of systems used in information retrieval from medical documentation, but they are mostly used on documents written in English language. This paper contains the explanation of our approach to solving the problem of extracting disease and drug names from medical records written in Serbian language. Our approach uses statistical language models and can detect up to 80% of named entities, which is a good result given the very limited resources for Serbian language, which makes the process of detection much more difficult.**

I. INTRODUCTION

Named entity recognition [1, 2, 3, 4] is part of the process called information extraction [5, 6, 7]. It is used to classify parts of text into predefined categories. Categories can vary, depending on task. Usually, text is divided into categories such as names of persons, organizations, locations, numbers, that can represent quantities of money, times, dates, percentages, etc. Entities relevant in this paper are names of diseases and medications and numbers that can represent dates, times and quantities as well as the abbreviations that medical staff often use.

Problem of named entity recognition is often solved using a grammar based or statistical methods [3, 4]. Commonly used statistical methods are supervised, semi-supervised and unsupervised methods.

The systems used for this task are primarily developed for English language and they use a variety of techniques to detect named entities, but most of them are useless for other languages. Statistical language models [15] can be of great help in dealing with languages that have sparse language resources, especially when used in combination with several other available techniques, like stemming [16]. Another advantage of using the statistical language models is the ability of their use on texts written in other languages with minor changes.

Today, hospitals and medical institutions produce huge amount of data about diseases and medications used in treatment and information about treatment success rate. Diagnoses written in text format are usually not structured and do not contain categorized information. This makes process of their understanding much more difficult. Computer systems which should recognize certain entities need to convert text into structured, and then to carry out the identification of specific parts that could be useful in the analysis.

Anamneses are composed of large amounts of useful information. It is possible to find the patient's history of illness, category the patient belongs like habits of smoking, drinking, etc. In addition to the historical background, text contains the tests that were carried out as well as the diagnosis that has been established. After the diagnosis, the patient is receiving a particular treatment that is contained in the document also. In addition, the record can have information about the amount in which the drugs are used, therapy duration and other useful details.

After that period, the patient undergoes examination when a doctor decides whether the treatment was successful or not. A system that allows obtaining such information from medical records can very easily highlight meaningful information in text or store them in structured way which can provide insights for better diagnosis or browsing history of the disease. Linking extracted named entities with the categories to which they belong can also expedite and facilitate the treatment process.

There are several approaches that can achieve these results. In the next section we will discuss the different methods used for solving these problems. In section III we will present our approach to solving this problem in documents written in Serbian language. The system is subject to changes and adding new functionality, which will be discussed in section 4 (Conclusion and future work).

II. RELATED WORK

There are several approaches to data extraction from text documents. Most of them are based on a method described in previous section, named entity recognition. This method can be accomplished in several ways, through supervised or semi-supervised learning algorithms, unsupervised learning algorithms [8, 9, 10].

Named entity recognition has the ability to recognize previously unknown entities using examples and rules. Examples are usually composed of positive and negative

ones. The algorithm uses these examples to create rules, later used to detect entities from new sentences.

Supervised learning [8] uses different techniques for named entity recognition, some of which are support vector machines [11], maximum entropy models [12], hidden Markov models [13], etc. Supervised learning approach uses huge set of training data, manually annotated, from which the system creates rules that are later used to identify entities in new sentences.

Unsupervised approach [9] uses unlabeled data to look for patterns in sentences. This is good approach to look for structure in data and to classify data into different categories.

Semi-supervised [10] learning approach is different because it uses smaller labeled training set and usually larger unlabeled one to create rules. This approach is useful in cases of insufficient data. Labeled data is expensive, but gives good results which makes this approach a good combination of supervised and unsupervised approaches.

A large number of named entity recognition systems for English language use unsupervised learning. This approach gives very good results, but it uses a large number of lexical resources such as WordNet and systems for part of speech tagging [24, 25, 26]. Also, semi-supervised learning [27] is widely used for bio-named entity recognition. Language resources in this approach are used to learn accurate word representations, but to a much smaller extent or even without using them in cases when this task is performed manually. Hidden Markov models [13], support vector machines [11] and conditional random fields [14] are often used as supervised learning [28] techniques. This method is not preferred for named entity recognition, because huge training dataset is needed. But, in some cases, like medical or biological texts, training data is already available.

Named entity recognition is used in the number of different tasks. Results depend on the methods used, as well as on the language over which those methods are applied. Typically, results are between 64% and 90%, but in some specific tasks can be near 100% [29, 30].

Some of the currently available tools for solving problem of named entity recognition in medical records are *Apache cTakes*[1] which is used to extract information from electronic medical records, written as free text, *CliNER*[2] (Clinical Named Entity Recognition system), an open source natural language processing system for named entity recognition in clinical text of electronic health records that uses conditional random fields (CRF) and support vector machines (SVM) and *DNorm*[3].

III. OUR APPROACH

The previous section provides a list of tools used in solving problem of named entity recognition in medical records. Listed tools are designed to work with documents written in English language. Medical records that we use are written in Serbian language. It is impossible to use any of these tools for documents written in any language other than English.

Our approach uses different technics for named entity recognition. Detection of disease and medication names is carried out using character and word n-gram models. Detection of dates, times and time intervals, on the other hand, uses parsed text to detect sequences of numbers and special abbreviations that are used to represent time intervals. Based on format, sequences obtained in this manner can distinguish between dates, times and time intervals. Other abbreviations are detected using dictionary that contains mostly used ones.

The process of named entity recognition that we used is carried out using statistical language models [15, 16]. It is necessary to divide the text first into words or characters and calculate the probability of their occurrence:

$$P(x_1^n) = \prod_{i=1}^{n} P(x_i \mid x_1^{i-1})$$

or

$$P(x_k \mid x_{k-3}, x_{k-2}, x_{k-1}) = \frac{Count(x_{k-3}, x_{k-2}, x_{k-1}, x_k)}{Count(x_{k-3}, x_{k-2}, x_{k-1})}$$

Character models [17] can be useful in recognizing words that do not appear in training data. Combinations of characters that appear in words are specific and can be indication of certain kinds of words.

For example, medications often contains "oxy" and "axy" group of characters. On the other hand, disease names frequently contain trigram "oza". Words with these groups of letters are not often used in medical records, so their presence is usually a sign that they represent names of medications or diseases.

Another way to reduce the number of "false positives" is stop words removal. [18] Stop words do not alter the meaning of the sentence, so that their removal does not affect the system accuracy. Depending on position, words can be lowercase of uppercase. Uppercase letters are not of big importance in this task, so text can be transformed to lowercase.

Another important transformation is lemmatization [19]. Words can be used in various forms which can make it difficult for recognition. Lemmatization transforms every word in text in its lemmatized form, the same form used in dictionaries, documents containing names of drugs and diseases mentioned before. This way, the grammatical rules used to create words in sentence are no longer a factor influencing the results.

Abbreviations in the text should be replaced by the words they represent. In case of codes present in list of diseases and medications, codes can be found in the documents. Other abbreviations are not easy to find. Incomplete list of commonly used English ones can be found online[4].

[1] http://ctakes.apache.org/index.html
[2] http://text-machine.cs.uml.edu/cliner/
[3] http://ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/DNorm
[4] http://studenti.mef.hr/abbrevations.doc

The numbers contained in the text do not necessarily represent categories that are of the interest to the system. Finding measures that are located next to the numbers is very important task. Fortunately, a list that contains measures is short and can be further divided in those that represent weight or time.

The process of creating of the models is applied over all data. The medication names and names of diseases do not include additional text that could interfere with the detection process, so the normalization [20] of the text could be skipped.

Once models are created, it is necessary to carry out a comparison over the parts of the text [21, 22, 23]. Elements of models that were created from medical records are sequentially compared with other models. If the similarity exceeds a certain minimum value, it is likely that part of the sentence is entity to be detected. Which category entity belongs, depends on the similarity with the different models created for those categories. The greatest similarity between model of some category and medical record model is an indication that named entity belongs to that category. However, the similarities between the different categories are rare, so resemblance with one of them usually means the affiliation to that category.

Numbers use a slightly altered approach. Primarily, system should detect numbers. Once this step is completed, detection of units of measure is performed. Units of measure which are located next to the numbers allow the identification of category that numbers belong. If being close to the number, units are an indication that number is an entity and they are remembered as one entity.

Medical document usually ends with a final diagnosis by a doctor established after the treatment. It may indicate that the patient is cured or it is necessary to continue treatment and further analysis. Detection of entity b.o. in the section that describes the condition of the patient after treatment means that he is healed successfully and does not need further actions. However, this is not always the case. On some occasions, the patient is referred for further treatment and tests or receives new therapy. This part of document can be similar to the previous parts, so it is possible to apply the same method for detection entities. This allows some of the entities present in that part of the document to be detected.

Entities that this system should recognize can be divided into the following categories:

– Names of diseases

List of disease names in Serbian language can be found in *International Statistic Classification of Diseases and Related Health Problems (ICD 10)*[5].

This list is divided into categories and contains code that represent every disease, category name which that disease belongs and names in Serbian and Latin.

Also, sometimes, doctors use these codes when writing diagnosis and this list makes it possible to decode them and replace with appropriate names. There are 14405 diseases listed, but using categories we can divide this list into smaller ones and to use them as training data.

– Names of medications

List of all medications used in Serbia can be found inside of *National Register of Medications (NRL)*[6]. Every entry in this register is represented by medication name, the company that produces it, the category to which the drug belongs, the code (*ATC code*), the date since when it is listed, the dosage and a detailed description of the cases where it is applicable, how it is used and what it consist of. This information can be very useful for linking with diseases that can be treated and to check whether the drug is suitable for treating disease.

– Abbreviations

Some of the abbreviations can be found in *NRL* described before, but doctors use other ones in medical records. A list of abbreviations can never be complete and it is hard to train system to recognize ones that are not listed, but they are important part of understanding meaning of medical record because doctors use them as substitute for many key parts of diagnosis.

– Numbers that represent dosage or dates and times

Numbers are often used in medical records as an indicator of the amount of medications prescribed by a doctor. In addition, the numbers may represent a time after which it is necessary to take medication or the dates when the patient needs to see doctor for examination. Amounts are usually accompanied by measures like milligrams (mg) or grams (g). On the other hand, dates and times are typically accompanied by measures of time like hours (h) or days. This can be helpful when determining category into which number belongs.

– Medical treatment success

Medical record usually ends with information about how successful treatment was. In the case of success, record typically ends with abbreviation b.o. (*English, N.A.D. nothing abnormal detected*). In case of unsuccessful treatment, doctor can list everything abnormal detected to that point and recommendations.

A. Overall Results

Records that we used in this paper consist of 42526 medical diagnoses from neurologic clinic. Those records consist of chief complains, description, history of disease and family diseases, psychosomatic and neurological symptoms all written as text in Serbian language. In the end, every record contains information about who prescribed therapy to the patient and was it successful.

Statistical linguistic models have proved to be a good choice because a large number of these diagnoses contain typographical errors. Despite attempts to rectify a number of them, some remained faulty. These, however, did not significantly affect the accuracy of the system, especially when order of some letters is replaced or a letter is misspelled.

[5] http://www.batut.org.rs/download/MKB102010Knjiga1.xls

[6] http://www.alims.gov.rs/ciril/files/2015/04/NRL-2015-alims.pdf

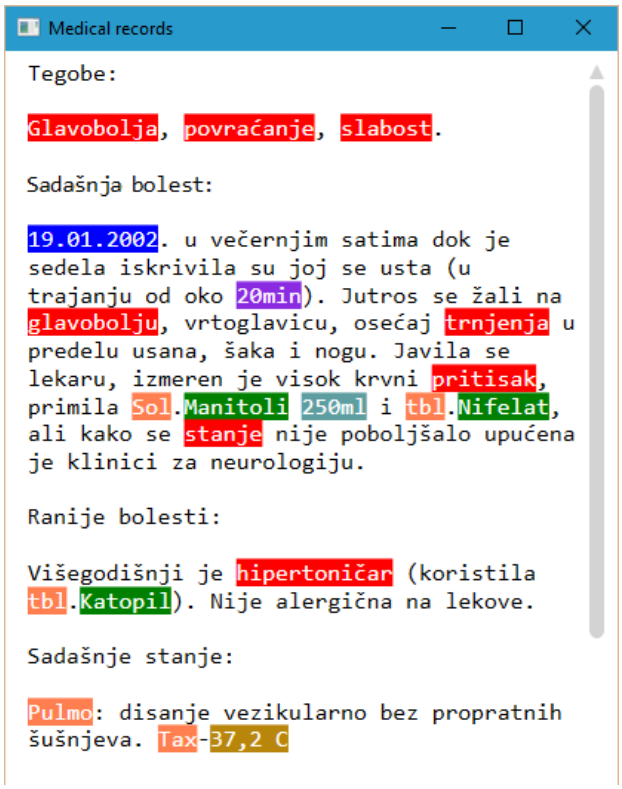

Figure 1. Detected entities shown in different colors

Figure 1 presents different entities shown in different colors. Words with red background color represent disease names and the green color is used for the names of drugs. The dates are marked in blue, time intervals in violet and numbers and abbreviations that represent quantities use cyan color. Words colored orange represent different kinds of abbreviations that doctors use in medical records. Also, dark yellow color is used to indicate other numbers that can provide more information, like temperature shown above.

Adjusting the length of the model gives different results. The best results were obtained using model length from 6 to 8 characters. Model used in this approach transform all disease and drug names in character models, but on the word level.

This approach gives good results in cases of small corpuses of disease and medications names, because of its ability to recognize similar words. This is not beneficial in cases of huge corpuses when there are a lot of false positives.

In order to remove false positives, models are produced differently. Producing models from large list of existing names word by word do not benefit from word relations in those names. Best way to solve this problem is to make character models from entire sentences of names. That raises another problem of variable length of those names. Solution we used is the use larger models and filling of empty spaces in cases of shorter ones with some special character that does not appear in text. This way, the process of the comparison comes down to a comparison of entire disease and medication names that removes possibility to detect similar words. This approaches have their advantages and disadvantages, therefore we use both.

Shorter records with a higher amount of specific data give the best results for models on entire names and sentences of length from 20 to 200 characters. Longer models eliminate the possibility of incorrectly detected named entities and results obtained using them are shown in Table I.

TABLE I.
CHARACTER MODELS MADE FROM SENTENCES AND NAMES

Text kind, length	Correct	Wrong	Not detected
Structured, 20	76%	2%	22%
Structured, 100	73%	3%	24%
Structured, 200	72%	1%	27%
Unstructured, 20	76%	1%	23%
Unstructured, 100	77%	0%	23%
Unstructured, 200	76%	0%	24%

Using 100 manually checked medical records

Results obtained using models of length 6 to 8 characters, created from words detect a large number of words that are not of our interest when used on small, structured text, but were good in case of long unstructured medical records. These results are shown in Table II.

TABLE II.
CHARACTER MODELS MADE FROM WORDS

Text kind, length	Correct	Wrong	Not detected
Structured, 6	35%	65%	0%
Structured, 7	37%	63%	0%
Structured, 8	38%	60%	2%
Unstructured, 6	78%	22%	0%
Unstructured, 7	77%	22%	1%
Unstructured, 8	81%	18%	1%

Using 100 manually checked medical records

It is very difficult to compare obtained results with those from other experiments. The task of finding a named entity depends on the category of entities that should be detected, but also on the language in which text is written. For this particular task the percentage of recognized entities is about 64%-90%. The complexity of the Serbian language makes this task even more difficult but the results obtained in our experiments are satisfactory, with the possibility for improvement.

Abbreviations that are impossible to recognize can be a problem during detection, but it is easy to differentiate them from other types of words used in text, so it is easy to detect and mark them. It is possible to request update of list of abbreviations or to simply use one in its original form.

All diagnoses are related to neurological diseases. It is possible to shorten list of medications and diseases to

speed up the system. This, however, can cause the problem in recognizing entities in areas of record that are containing history of illness or possible complications after received treatment. Yet the division into different categories of diseases and medications can help in classification of entities into groups that belong only to certain types of diseases or medications.

IV. Conclusion and future work

The information obtained by these methods, separated by categories make the overview of the diagnosis much easier. However that is not the only advantage of this system. The entities of all categories are directly related because they are obtained from the same medical record. This allows determination of the correlation between the relevant entities. Names of diseases are linked with medications used to treat them and many other features. As stated in previous sections, the list of drugs includes detailed descriptions of each of them, like substances they contain, but also list of possible replacements. This can help the doctor to select a suitable replacement for the drug if it is necessary, but also to suggest that a drug can create problems to the patient, in case of allergies or in case of medication intolerance.

No less important is the ability to determine in which cases the therapies proven effective and led to the healing of the patient, and which have caused additional complications and required further treatment.

A large amount of information and number of drugs and diseases makes it difficult for doctors to choose the most effective treatment that could be applied. Systems like this could provide a better insight and a lot of helpful information extracted from different sources.

System meets the requirements of detection on the medical records described in this paper, but there are possible improvements to make it better in more complex cases. Lack of the part of speech tagger is the biggest problem and a major handicap, but realization is not an easy task. Its use would facilitate the identification of potential parts of sentences that could be identified as a named entity.

Although the information is very useful, data obtained in this way can be used in many other purposes. The connection of the disease history with a current diagnosis and possible complications is one of the interesting approaches that might give good results in the creation of patterns, both in terms of diseases, and the time when they occurred.

The next step in improvement of this system should be the implementation of other detection methods for named entity recognition. Different techniques should be applied to large amounts of medical records in order to find a suitable method for various types of diagnoses. Medical documentation contains a wide range of diagnoses depending on the area in which they are written, so it is necessary to find a suitable combination of techniques that give the best results.

Huge amount of information requires finding better methods and the ways to minimize time required for their processing.

Acknowledgment

Research presented in this paper was funded by the Ministry of Science of the Republic of Serbia, within the project "Technology Enhanced Learning in Serbia", No. III 47003.

References

[1] Erik F. Tjong Kim Sang, Fien De Meulder, *Introduction to the CoNLL-2003 shared task: language-independent named entity recognition*, CONLL '03 Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003. - Volume 4, pp. 142-147, 2003.

[2] Nadeau, David; Sekine, Satoshi, *A survey of named entity recognition and classification*, Lingvisticae Investigationes, Volume 30, Number 1, pp. 3-26(24), 2007.

[3] GuoDong Zhou, Jian Su, *Named entity recognition using an HMM-based chunk tagger*, ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 473-480, 2002.

[4] Dan Klein, Joseph Smarr, Huy Nguyen, Christopher D. Manning, *Named entity recognition with character-level models*, CONLL '03 Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4, pp. 180-183, 2003.

[5] Stephen Soderland, *Learning Information Extraction Rules for Semi-Structured and Free Text*, Machine Learning, Volume 34, Issue 1, pp. 233-272, 1999.

[6] Ellen Riloff, Rosie Jones, *Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping*, AAAI-99 Proceedings, 1999.

[7] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, *Introduction to Information Retrieveal*, Cambridge University Press, 2008.

[8] Cezary Z. Janikow, *A knowledge-intensive genetic algorithm for supervised learning*, Machine Learning, Volume 13, Issue 2, pp. 189-228, 1993.

[9] Trevor Hastie , Robert Tibshirani, Jerome Friedman, *Unsupervised Learning*, The Elements of Statistical Learning, Part of the series Springer Series in Statistics, pp. 485-585, 2009.

[10] Olivier Chapelle, Bernhard Scholkopf, Alexander Zien, *Semi-Supervised Learning*, The MIT Press, Cambridge, Massachusetts, 2006.

[11] Jun'ichi Kazama, Takaki Makino, Yoshihiro Ohta, Jun'ichi Tsujii, *Tuning support vector machines for biomedical named entity recognition*, BioMed '02 Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain - Volume 3, pp. 1-8, 2002.

[12] Hai Leong Chieu, Hwee Tou Ng, *Named entity recognition: a maximum entropy approach using global information*, COLING '02 Proceedings of the 19th international conference on Computational linguistics - Volume 1, pp. 1-7, 2002.

[13] Sean R Eddy, *Hidden Markov models*, Current Opinion in Structural Biology, Volume 6, Issue 3, pp. 361–365, Elsevier, 1996.

[14] Burr Settles, *Biomedical named entity recognition using conditional random fields and rich feature sets*, JNLPBA '04 Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, pp. 104-107, 2004.

[15] ChengXiang Zhai, *Statistical Language Models for Information Retrieval, Synthesis Lectures on Human Language Technologies*, pp. 141, 2008.

[16] Darko Puflović, Leonid Stoimenov, *Plagiarism detection in homework assignments and term papers*, The Sixth International Conference on e-Learning, pp. 204-209., Belgrade, Serbia, 2015.

[17] Praneeth M Shishtla, Prasad Pingali, Vasudeva Varma, *A Character n-gram Based Approach for Improved Recall in Indian Language NER*, IJCNLP-08 Workshop on NER for South and South East Asian Languages, pp. 67-74, 2008.

[18] Akiko Aizawa, *Linguistic Techniques to Improve the Performance of Automatic Text Categorization*, Proceedings of the Sixth

Natural Language Processing Pacific Rim Symposium (NLPRS2001), pp. 307–314. 2001.

[19] Vlado Kešelj, Danko Šipka, *A suffix subsumption-based aproach to building stemmers and lemmatizer for highly inflectional languages with sparse resources*, INFOthecha, 2008.

[20] Andrei Mikheev, *Document centered approach to text normalization*, SIGIR '00 Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 136-143, 2000.

[21] Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, Supachanun Wanapu, *Using of Jaccard Coefficient for Keywords Similarity*, Proceedings of the International MultiConference of Engineers and Computer Scientists 2013., Vol I, IMECS 2013, Hong Kong, 2013.

[22] Shie-Jue Lee, *A Similarity Measure for Text Classification and Clustering*, IEEE Computer Society, Issue No.07 - July (2014 vol.26), pp. 1, 2014.

[23] Subhashini, R., Kumar, V.J.S., *Evaluating the Performance of Similarity Measures Used in Document Clustering and Information Retrieval*, First International Conference on Integrated Intelligent Computing (ICIIC 2010), pp. 27-31, 2010.

[24] Shaodian Zhang, Noémie Elhadad, *Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts*, Journal of Biomedical Informatiocs, Volume 46, Issue 6, December 2013, pp. 1088-1098, 2013.

[25] George A. Miller , *WordNet: a lexical database for English*, Communications of the ACM, Volume 38 Issue 11, Nov. 1995, pp. 39-41, 1995.

[26] Christopher D. Manning, *Part-of-speech tagging from 97% to 100%: is it time for some linguistics?*, CICLing'11 Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part I, pp. 171-189, 2011.

[27] Pavel P. Kuksa, Yanjun Qi, *Semi-supervised Bio-named Entity Recognition with Word-Codebook Learning*, Proceedings of the SIAM International Conference on Data Mining, SDM 2010, pp. 25-36, 2010.

[28] Bodnari, A., Deleger, L., Lavergne, T., Neveol, A., Zweigenbaum, P.: *A supervised named-entity extraction system for medical text*, Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop, September 2013.

[29] Xiao Fu, Sophia Ananiadou, *Improving the Extraction of Clinical Concepts from Clinical Records*, Proceedings of the Fourth Workshop on Building and Evaluation Resources for Biomedical Text Mining (BioTxtM 2014) at Language Resources and Evaluation (LREC) 2014.

[30] Giuseppe Attardi, Vittoria Cozza, Daniele Sartiano, *Annotation and extraction of relations from Italian medical records*, Proceedings of the 6th Italian Information Retrieval Workshop, Cagliari, Italy, May 25 – 26, 2015, CEUR Workshop Proceedings, vol. 1404, 2015.