Extraction of Medical Concepts from Italian Natural Language Descriptions

(Discussion Paper)

Patrizia Agnello[1], Silvia Maria Ansaldi[1], Fabio Azzalini[2,3], Giovanni Gangemi[2], Davide Piantella[2], Emanuele Rabosio[3] and Letizia Tanca[2]

[1]*INAIL - Dipartimento Innovazioni Tecnologiche*
[2]*Politecnico di Milano - Dipartimento di Elettronica, Informazione e Bioingegneria*
[3]*Human Technopole - Center for Analysis, Decisions and Society*

Abstract

In this paper we present a Natural Language Processing (NLP) pipeline to automatically extract medical concepts from a free text written in a language other than English. To do so, we use common NLP techniques and the metathesaurus of Unified Medical Language System (UMLS). Specifically, our goal is to automatically extract ontological concepts representing which part of the human body is injured and what is the nature of the injury, given an Italian textual description of a work accident. We start by partitioning the text into tokens and assigning to each token its part-of-speech, and then use an appropriate tool to extract relevant concepts to be searched within UMLS. We tested our system on a public large repository containing textual descriptions of work accidents produced by INAIL. Experimental results confirm that our system is able to correctly extract relevant medical concepts from texts written in Italian.

Keywords

Ontology, EHR, NLP, Work accident

1. Introduction

The term Electronic Health Records (EHRs) describes the concept of a comprehensive, cross-institutional, and longitudinal collection of healthcare data, trying to group the entire clinical life of a patient [1]. EHRs store information both in structured (e.g. diagnosis codes, laboratory results, etc) and unstructured (e.g. clinical notes, discharge summaries, etc.) formats. Unstructured data usually contain a more complete, and broader, view of the patient's conditions, as well as additional valuable information that would be difficult to represent in a structured manner (e.g. social history, special conditions, etc.). To leverage all the advantages of a systematic adoption of EHRs, many technical and non-technical requirements must be fulfilled [2], such as

SEBD 2021: The 29th Italian Symposium on Advanced Database Systems, September 5-9, 2021, Pizzo Calabro (VV), Italy

p.agnello@inail.it (P. Agnello); s.ansaldi@inail.it (S. M. Ansaldi); fabio.azzalini@polimi.it (F. Azzalini); giovanni.gangemi@mail.polimi.it (G. Gangemi); davide.piantella@polimi.it (D. Piantella); emanuele.rabosio@fht.org (E. Rabosio); letizia.tanca@polimi.it (L. Tanca)
0000-0003-0631-2120 (F. Azzalini); 0000-0003-1542-0326 (D. Piantella); 0000-0003-3722-7789 (E. Rabosio); 0000-0003-2607-3171 (L. Tanca)

privacy, data security, portability, performance, maintainability, reliability, interoperability, and usability.

In this work we focus on the natural-language texts included in EHRs, since unstructured data analytics is one of the most challenging task of EHRs automated analysis. Our main contributions are the development of a system capable of automatically extracting medical ontological concepts from an Italian natural language text, and the experimental evaluation of our system using a real-world dataset regarding work injuries.

The paper is organized as follows. Section 1 gives an overview of the problem, Section 2 reviews the state of the art regarding medical concept extraction, Section 3 describes our methodology, Section 4 presents the experiments and, finally, Section 5 concludes the paper.

2. State of the Art

Concept extraction from natural language texts related to clinical information consists of three phases [3]: (1.) Identification of concept mentions such as medications, drugs, anatomical parts, and diseases; (2.) Coreference resolution regarding relationships between different mentions referring to the same entity; and (3.) Extraction of relationships between concepts.

We now briefly describe one of the most complete medical ontology system (UMLS) and two state-of-the-art frameworks for clinical concept extraction: *cTAKES* and *MetaMap*.

2.1. Unified Medical Language System

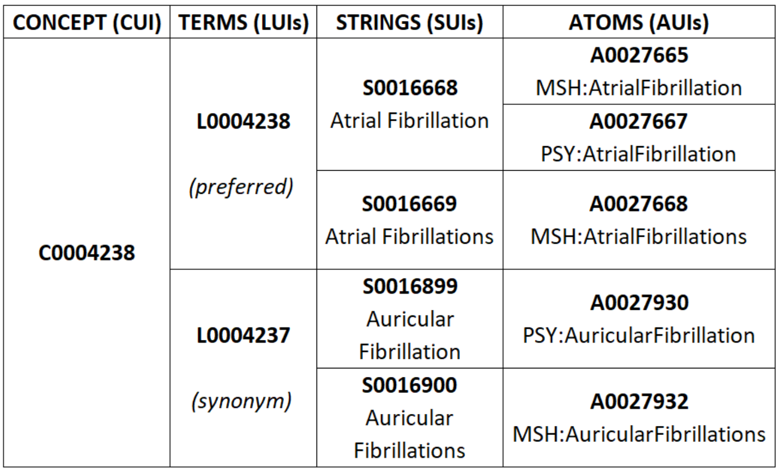
The *Unified Medical Language System (UMLS)* [4] is a compendium of many controlled vocabularies in the biomedical sciences, produced and distributed by the National Library of Medicine (NLM). It also provides a mapping structure among these vocabularies and thus allows to translate among the various terminology systems. It can be therefore considered a comprehensive thesaurus and ontology of biomedical concepts.

UMLS is composed by three modules: Metathesaurus, Semantic Network, and Specialist Lexicon. We now provide a brief explanation of each of these modules.

2.1.1. Metathesaurus

The Metathesaurus of UMLS includes over one million biomedical concepts and five million concept names, enclosing many vocabularies such as ICD-10, SNOMED CT, MeSH, and more. The Metathesaurus is structured to facilitate the identification of synonyms between concepts, also in different languages ensured by leveraging hierarchical concept identifiers, in turn linked to:

- Concepts (CUI): identifying the meaning to be expressed, it never changes over time, no matter the updates in the vocabularies.
- Strings (SUI): each string representing a concept is assigned with a permanent string identifier. Any character variation (e.g. case sensitivity, punctuation, etc.) will result in a different SUI, for each language. A SUI can in principle be linked to more than one CUI.
- Atoms (AUI): being building blocks of the Metathesaurus, atoms represent specific entries in the vocabularies included in UMLS. An AUI is linked to one and only one CUI.



CONCEPT (CUI)	TERMS (LUIs)	STRINGS (SUIs)	ATOMS (AUIs)
C0004238	**L0004238** *(preferred)*	**S0016668** Atrial Fibrillation	**A0027665** MSH:AtrialFibrillation
			A0027667 PSY:AtrialFibrillation
		S0016669 Atrial Fibrillations	**A0027668** MSH:AtrialFibrillations
	L0004237 *(synonym)*	**S0016899** Auricular Fibrillation	**A0027930** PSY:AuricularFibrillation
		S0016900 Auricular Fibrillations	**A0027932** MSH:AuricularFibrillations

Figure 1: Hierarchical concept identifiers of UMLS, as reported in [5]

- Lexical terms (LUI): a lexical term comprises different strings (i.e. SUI) that are lexical variants or minor variants. This layer is often used to reduce the computational complexity of exploration and for a more effective concept lookup. It is currently available for all the English strings, and only partially for other languages.

2.1.2. Semantic Network

The Semantic Network provides a consistent categorization of all concepts represented in the Metathesaurus along with a set of useful relationships between these concepts. The network contains 133 semantic types and 54 relationships. Each concept in the Metathesaurus is assigned one or more semantic types, which are linked to one another through semantic relationships. The major semantic types are *organisms, anatomical structures, biologic function, chemical, events, physical objects*, etc.

The possible relationships between semantic types range from simple ⟨*is-a*⟩ hierarchies to complex associations, such as ⟨*physically related to*⟩, ⟨*spatially related to*⟩, ⟨*co-occurs with*⟩. Relationships can be derived from associations already present in the vocabularies (*intra-source* relationships) or they can connect concepts from different vocabularies (*inter-source* relationships), including not only synonyms. Inter-source relationships enhance the integration of all the vocabularies present in UMLS and enable an easy exploration of the resulting ontology. A subset of the Semantic Network with ⟨*is-a*⟩ relationships is shown in Figure 2.

2.1.3. Specialist Lexicon

Specialist Lexicon is a module of UMLS that addresses the high degree of variability in natural language words, allowing the abstraction of lexical variants. It covers general English lexicon and many biomedical terms, including syntactic, morphological, and orthographic information. Since only English words are covered by this module, we adopted a different approach for Italian Natural Language Processing, which we will describe in Section 3.

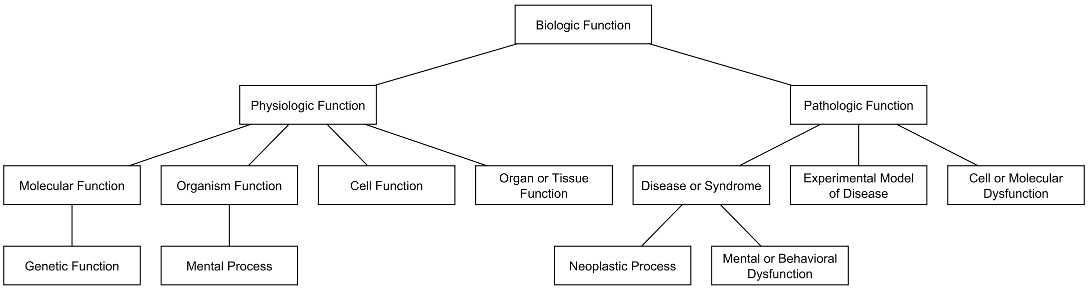


Figure 2: A portion of UMLS Semantic Network

2.2. cTAKES

Apache *clinical Text Analysis and Knowledge Extraction System* [6] (cTAKES, for short) is an open-source framework for knowledge extraction from clinical texts, exploiting NLP techniques including rule-based and machine learning approaches. It leverages a pipeline of six components. First of all the text is divided into sentences by the *sentence boundary detector*, a component which extends OpenNLP sentence detector [7]. Each sentence is then *tokenized* taking into consideration also context-specific occurrences (e.g. dates, time intervals, etc.). Each token is then *normalized* leveraging tools part of UMLS Specialist Lexicon (described in Section 2.1.3), in order to map each token in a normalized form with respect to many lexical properties (e.g. inflection, diacritics, symbols, stop words, etc.). Both the normalized and the original occurrences are maintained for further analysis. After a *part-of-speech tagging*, the *named entity recognition annotator* component performs a terminology-agnostic dictionary lookup on a subset of UMLS Metathesaurus (described in Section 2.1.1), searching all the noun-phrases identified and their respective unnormalized occurrences.

2.3. MetaMap

MetaMap was developed by the National Library of Medicine (NLM) with the goal of mapping biomedical text to the UMLS Metathesaurus [8]. It relies on a pipeline similar to cTAKES, apart from the leveraging of relationships and hierarchical identifiers, present in UMLS Metathesaurus, to better identify synonyms and lexical variants of the tokenized texts.

2.4. Comparison between cTAKES and MetaMap

A comparison between cTAKES and MetaMap has already been investigated in [9]: the results of the experiments proved that cTAKES slightly outperforms MetaMap, with the exception of texts in which abbreviations are present. It has been shown that abbreviations are quite common in natural language texts written by doctors and both tools have difficulties in correctly identify their correct meanings. With MetaMap, however, it is possible to partially solve this problem, specifying a list of strings that will be treated as special tokens. This possibility is

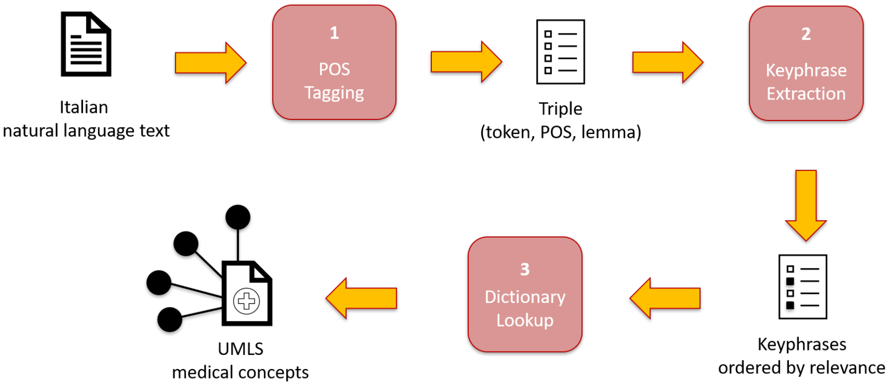


Figure 3: The framework of our system

not particularly investigated in the cited experimental comparison, and could be the subject of future studies.

The main disadvantage of both cTAKES and MetaMap, with respect to our study, is that they are strongly English-centric, since they both rely on UMLS Specialist Lexicon which, as we already described in Section 2.1.3, fully covers only the English language.

3. Methodology

We now present the methodology of our system. The final goal is to automatically extract ontological concepts representing which part of the human body is injured and what is the nature of the injury, given an Italian textual description of a work accident. Figure 3 displays the three phases of our workflow: Part-of-Speech (POS) Tagging, Keyphrase Extraction and Dictionary Lookup.

The first phase receives as input a textual description representing the dynamic of an accident and gives as output a preprocessed and enriched representation of the input text. Specifically, Tint takes as input a raw text in Italian and performs a series of natural language processing operations. Tint (The Italian NLP Tool) [10] is an open-source Java-based pipeline for Natural Language Processing (NLP) in Italian. It is very fast and accurate, and implements most of the common linguistic tools, such as part-of-speech tagging and dependency parsing. This first phase is necessary since the next stage needs the text divided into tokens, lemmas and parts of speech in order to continue the execution.

Example 1: Given the following accident description: "Erano in corso attività di produzione di acciaio. Mentre un agganciamento del nastro trasportatore vibrovaglio alla motopala, a causa di una manovra pericolosa rimaneva con le braccia in contrasto tra le due macchine decedendo per contusione al fegato", the pipeline produces the tagged text visible in Table 1.

The second phase receives as input the tagged text with lemmas and POS, and returns as output a new series of keyphrases ordered by importance and frequency. This step uses a tool called Keyphrase Digger (KD) [11] which analyzes the text file with tokens, lemmas and pos and returns as a result a new text file with a series of keyphrases ordered by importance and

Table 1
A possible output of phase 1

Token	POS	Lemma
Erano	Verb	essere
in	Preposition	in
corso	Noun	corso
attività	Noun	attività
di	Preposition	di
produzione	Noun	produzione
di	Preposition	di
acciaio	Noun	acciaio

frequency. Keyphrases are n-grams of different length, both single and multi-token expressions, which capture the main concepts of a given document [12]. Keyprhase extraction is essential to understand the topic covered in long text and has many applications, especially when integrated into pipelines, like this one, that perform more complex tasks.

Example 2: After the second phase the pipeline extracted the following concepts: “produzione di acciaio”, “nastro trasportatore vibro”, “contusione al fegato”, “vaglio alla motopala” and “manovra pericolosa”.

As can be seen from Example 2, not all the concepts extracted from the accident description contain information regarding the part of the body and the nature of the incident. The third and last phase, exploiting the Rest API provided by the UMLS, allows the system to query various databases and to discard all the concepts that do not relate to the medical field. This phase can be divided into two distinct sub-phases:

- Concept lookup: we create a query that queries UMLS to get the medical concept. We use as input every possible combination of the keyphrases obtained in the previous step.
- Semantic type lookup: after obtaining the medical concept, we check if it belongs to one of these four semantic types, which represent nature and location of an injury.

Example 3: After the third phase the only keyphrase not discarded is “contusione al fegato”.

4. Experiments

We tested our system on the accident descriptions contained in the InforMo dataset [13] made available by INAIL, a repository containing the results of a survey on mostly fatal accidents occurring during work time. This dataset contains 636 entries, each with detailed information on the incident. Concepts are extracted from the description written in natural language in the questionnaire (called dynamic) by those who compiled it. In the original questionnaire there is not always consistency between what is written in the dynamic section and the nature and location of the lesion’s attributes. As an example, we may find the concept “skull injury" in the dynamic of the accident, and “contusion" manually written as the nature of the injury. These two concepts might be considered as synonyms for someone who is filling out the questionnaire, but in an ontology they are two different concepts. To solve this problem, we decided to manually create a Golden Truth, analyzing all the dynamics to understand what could be extracted from them.

For each text, our system must extract two concepts that will form a pair constituting the nature and location of an injury. What is extracted is compared with the golden truth to assess the accuracy of the framework. What we want to achieve is an exact extraction of the nature-location pair directly from the textual description of the accident.

After analyzing each textual description, we have evidence to say that most of the times when a nature-location pair is present in the text, it is in the same period, so we analyze each period. If a couple is present in a period of the text we keep it, otherwise we delete the couple. This allowed to greatly reduce the extracted concepts while keeping the performance unchanged.

To evaluate the performance of our pipeline we used recall, precision and F1-score, whose definitions are reported reported here:

$$\text{Recall} := \frac{TP}{TP + FN} \quad \text{Precision} := \frac{TP}{TP + FP} \quad \text{F1-Score} := 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

TP represents a correctly guessed nature-location pair, the FP instead are all those incorrect but still extracted, FN are those that are mistakenly not recognized as a match.

In the UMLS query we can specify a parameter called “searchType" which can take two different values: “words" (by default) or “exact". With the first, a similarity search is carried out, resulting in the list of concepts most similar to the one given in input, ordered by decreasing similarity. With the second, on the other hand, a result is obtained only if the input word really exists in the database. We tested the system checking both of these parameters so as to understand which is the best one. In the Table 2 we list the results in the two cases. The results show that the “exact" case is better than the second. A deeper analysis highlights that this it is due to the fact that in the second case many more concepts are extracted, most of which are quite different from the original one.

Table 2
Experimental results

	exact	words
Recall	**0.90**	0.85
Precision	**0.51**	0.20
F1-Score	**0.65**	0.32

5. Conclusion and Future Work

In this paper we presented a methodology for extracting medical concepts from accident descriptions written in natural language, specifically tailored for the Italian language. The system, still being in a preliminary phase, suffers from some limitations: *(i)* there is a strong dependence on UMLS and its provided APIs, this often makes the system pretty slow in its computation *(ii)* the experimental campaign carried out is pretty limited, this may cause problems in the applicability of the framework to other input sources.

Acknowledgements

This work has been funded by INAIL within the BRiC 2018, ID09 framework, project RECKON.

References

[1] B. Shickel, P. J. Tighe, A. Bihorac, P. Rashidi, Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis, IEEE journal of biomedical and health informatics 22 (2017) 1589–1604.

[2] A. Hoerbst, E. Ammenwerth, Electronic health records, Methods Inf Med 49 (2010) 320–336.

[3] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, et al., Clinical information extraction applications: a literature review, Journal of biomedical informatics 77 (2018) 34–49.

[4] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Research 32 (2004) D267–D270. doi:10.1093/nar/gkh061.

[5] National Library of Medicine, UMLS reference manual, https://www.ncbi.nlm.nih.gov/books/NBK9676/, 2021. Online; accessed 24-April-2021.

[6] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, C. G. Chute, Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications, Journal of the American Medical Informatics Association 17 (2010) 507–513.

[7] Apache Software Foundation, openNLP website, https://opennlp.apache.org/, 2021. Online; accessed 28-April-2021.

[8] A. R. Aronson, F.-M. Lang, An overview of metamap: historical perspective and recent advances, Journal of the American Medical Informatics Association 17 (2010) 229–236.

[9] R. Reátegui, S. Ratté, Comparison of metamap and ctakes for entity extraction in clinical notes, BMC medical informatics and decision making 18 (2018) 13–19.

[10] A. Palmero Aprosio, G. Moretti, Tint 2.0: an all-inclusive suite for nlp in italian, Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 10 (2018) 12.

[11] G. Moretti, R. Sprugnoli, S. Tonelli, Digging in the dirt: Extracting keyphrases from texts with kd, CLiC it 198 (2015).

[12] P. D. Turney, Learning algorithms for keyphrase extraction, Information retrieval 2 (2000) 303–336.

[13] INAIL, Informo dataset, https://www.inail.it/sol-informo/, 2021. Online; accessed 28-April-2021.