JLIS.it vol. 14, no. 2 (May 2023)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.36253/jlis.it-522

JLIS.it

Neurodegenerative clinical records analyzer: detection of recurrent patterns within clinical records towards the identification of typical signs of neurodegenerative disease history*

Erika Pasceri[(a)], Mérième Bouhandi[(b)], Claudia Lanza[(c)], Anna Perri[(d)], Valentina Laganà[(e)], Raffaele Maletta[(f)], Raffaele Di Lorenzo[(g)], Amalia C. Bruni[(h)]

a) Department of Culture, Education and Society, University of Calabria (CS), Italy, https://orcid.org/0000-0001-9917-2184
b) LS2N, UMR CNRS 6004, Nantes Université, Nantes, France, https://orcid.org/0000-0001-7899-8198
c) Department of Culture, Education and Society, University of Calabria (CS), Italy, https://orcid.org/0000-0002-3018-1987
d) Department of Culture, Education and Society, University of Calabria (CS), Italy, https://orcid.org/0000-0002-2852-0919
e) Association for Neurogenetic Research (ARN), Lamezia Terme, CZ, Italy, https://orcid.org/0000-0001-8379-400X
f) Regional Neurogenetic Centre, ASP CZ, Lamezia Terme, CZ, Italy, https://orcid.org/0000-0001-5848-9097
g) Regional Neurogenetic Centre, ASP CZ, Lamezia Terme, CZ, Italy
h) Regional Neurogenetic Centre, ASP CZ, Lamezia Terme, CZ, Italy, https://orcid.org/0000-0003-3471-3343

Contact: Erika Pasceri, erika.pasceri@unical.it; Mérième Bouhandi, merieme.bouhandi@ls2n.fr; Claudia Lanza, claudia.lanza@unical.it; Anna Perri, anna.perri@unical.it; Valentina Laganà, valelagana@gmail.com; Raffaele Maletta, maletta@arn.it; Raffaele Di Lorenzo, dilorenzo@arn.it; Amalia C. Bruni, amaliaceciliabruni@gmail.com
Received: 17 November 2022; **Accepted:** 09 January 2023; **First Published:** 15 May 2023

ABSTRACT

When treating structured health-system-related knowledge, the establishment of an over-dimension to guide the separation of entities becomes essential. This is consistent with the information retrieval processes aimed at defining a coherent and dynamic way – meaning by that the multilevel integration of medical textual inputs and computational interpretation – to replicate the flow of data inserted in the clinical records. This study presents a strategic technique to categorize the clinical entities related to patients affected by neurodegenerative diseases. After a pre-processing range of tasks over paper-based and handwritten medical records, and through subsequent machine learning and, more specifically, natural language processing operations over the digitized clinical records, the research activity provides a semantic support system to detect the main symptoms and locate them in the appropriate clusters. Finally, the supervision of the experts proved to be essential in the correspondence sequence configuration aimed at providing an automatic reading of the clinical records according to the clinical data that is needed to predict the detection of neurodegenerative disease symptoms.

KEYWORDS

Alzheimer; Categorization; Electronic health records (EHR); Machine learning; Semantic annotation.

* Authors all contributed in the realization of this paper, although Mérième Bouhandi specifically dealt with sub-section "Classifying Alzheimer-related indicators", "Results" and "Discussion" Sections, Claudia Lanza dealt with the "Introduction" Section, covered the "Materials and methods", "Text recognition" Sections and detailed in sub-section "Neurodegenerative categories matching with expressions" the categories schematization to run the association between medical expressions and pre-established categories; Erika Pasceri mainly focused on Section "Related works" dealing with the preliminary related works on the clinical health records digitization process and annotation as well as the KOSs frameworks, on sub-section "Sample definition and clinical records digitization" addressing the case of study framework digitization and on "Conclusion" Section , Anna Perri dealt with the sub-section "Objectives and context framework" and integrated the process of validating the correct correlation within the tags, Valentina Laganà, Raffaele Maletta, Raffaele Di Lorenzo and Amalia Cecilia Bruni provided the whole set of clinical data giving access to the archive and supported the evaluation process of the isolated linguistic structures with their domain field of knowledge.

Introduction

This paper presents a multidisciplinary research activity dealing with the realization of a semantic analyzer tool for the management of information contained in the digital clinical records of patients affected by Alzheimer's Diseases (AD), a progressive and disabling neurodegenerative disorder that, rarely, can be inherited in an autosomal dominant way (Bruni, Bernardi, and Maletta 2021; Alzheimer's Association 2016). AD is characterized by cognitive deficits, e.g., memory loss and behavioral and psychological symptoms of dementia (BPSD), including a wide range of non-cognitive symptoms involving perception, mood, behavior, personality, and basic functioning (Bruni, Bernardi, and Maletta 2021; Bruni, Bernardi, and Gabelli 2020).
The disposal of structured clinical data referring uniquely to specific under-treatment or post-treatment records results to be a leading task (Mills 2019). Indeed, the decision-making operations undertaken by doctors within specialized health sectors are generally based on the reference to specific parameters structured over the clinical documents (Shellum et al. 2016)in part by providing the capability for a broad range of clinical decision support, including contextual references (e.g., Infobuttons. Therefore, when it comes to understanding the logic behind a medical set of procedures, it becomes essential to evaluate them under the lens of the clinical information structure that can facilitate the appropriate data input as well as the subsequent inference processes over the acquired knowledge base. This study specifically describes the steps followed towards the construction of a semantic analyzer for the electronic health records (HER) referring to the AD. In particular, the paper is subdivided into several sections reflecting the different stages pursued to reach the development of a clinical supporting reader from a semantic perspective capable of retrieving the AD-related categories from the analysis of the linguistic expressions included in the anamnesis.

Related works

In the healthcare literature, considerable attention has been directed toward the clinical decision support (CDS) tools in the way they can provide medical operators with a pre-settled clinical workflow meant to orientate the health data insertion and the subsequent execution of specialized tasks (Kharbanda et al. 2018; Spineth, Rappelsberger, and Adlassnig 2018; Tolley et al. 2018; Beeler, Bates, and Hug 2014). To this end, it is necessary to apply information and knowledge management advanced techniques and methodologies which allow users to understand, share and use available information and transform data into knowledge. In this study, a focus will be given to semantic annotation, classification and evaluation of the clinical data with respect to a reference corpus made of the anamnesis referred to patients suffering from AD syndrome. The concept "semantic annotation" is intended as «the process of attaching to a text document or other unstructured content, metadata about concepts (e.g., people, places, organizations, products or topics) relevant to it». Specifically referring to the biomedical domain, it is worth mentioning three biomedical annotators: (i) Clinical Text Analysis and Knowledge Extraction System (cTAKES) (Savova et al. 2010) based on Unstructured Information Management Architecture (UIMA) and OpenNLP frameworks; (ii) MetaMap (Stewart, von Maltzahn, and Abidi 2012) which exploits the Unified Medical Language System (UMLS) Metathesaurus to process the mapping with the medical entities of the electronic health records and the concepts contained in the classification systems; (iii) MedCATTrainer annotator (Searle et al. 2019) which works in conjunction with Named Entity Recognition and Linking (NER+L) operators to extract medical information from texts. Despite the relevant outcomes found in exploiting the semantic annotator tools and the facility to apply their main functions to the source biomedical texts, the Natural Language Processing (NLP) tasks executed on unstructured texts via machine learning techniques intrinsically provide a more fine-grained systematization of the categories to be retrieved and used as tagging segmentation of clinical datasets. In this way, users can meet specific medical needs by collecting several important sequences of clinical records characterized by a medical recursive writing schema, offering early detection work on the patients' medical history. The detection of key textual units has also been addressed by (Hassanzadeh, Nguyen, and Koopman 2016), who exploit external officially shared semantic resources (MetaMap, NCBO annotator, Ontoserver, and QuickUMLS) to map the medical information in the EHR and obtain a more reliable set of data framework. (Klassen, Xia, and Yetisgen 2016) built NLP schemes to identify medical events in clinical notes in order to detect the diagnosis or coordination changes, as well as (Patel et al. 2018) who created a clinical entity recognition (CER) process using machine learning techniques classifying the desired outputs in categorized sequences to be retrieved. Tou et al. (2018) describe a study on the isolation of medical forecastable clusters referring to personal data, vital signs, or diagnosis results to detect the main forms of infections. The biomedical domain is also rich in Knowledge Organization Systems (KOSs), which differ in various aspects: their type (classification systems, thesauri, ontologies, etc.); their function and purpose (information retrieval, information sharing, indexing, etc.) (Mazzocchi 2018). Among these, Alzheimer's Disease Thesaurus is used for indexing and searching the ADEAR (Alzheimer's and related Dementias Education and Referral Center) database, which was created in 1990 by the US Congress as Alzheimer's Disease Education and Referral with the aim to «compile, archive, and disseminate information concerning Alzheimer's disease for health professionals, people with AD and their families, and the public». The OWL Ontology includes 156,869 classes belonging to different categories, such as organism, anatomy, biological process, neurological disease, neurological disorder, cellular anatomy, and so on.

Objectives and context framework

The neurodegenerative categorization system from which this study has taken its ground has been forged from the one created by the Italian Institute of neurodegenerative diseases (Laganà et al. 2022)from 2006 to 2018, were studied. Symptoms have been extracted from Neuropsychiatric Inventory (NPI located in the South of Italy, with the purpose of enhancing it by executing machine learning operations. The research tasks will be conducted through a semantic analysis of the expressions contained in a sample of clinical records related to the AD-patients. Indeed, the expected achievement of this activity is the development of the automatic classification of the typical expressions contained in the clinical records with the enhanced version of the categories' systematization. The pre-existing categories concerning cognitive and motor signs/symptoms, as well as the BPSD, have been developed and validated by the neurologists and psychologists working at the Neurogenetic Center of Calabria Region[1], Italy, where the archive containing the clinical records that have been used in this study (Laganà et al. 2022) from 2006 to 2018, were studied. Symptoms have been extracted from Neuropsychiatric Inventory (NPI is located. The archive consists of 12,860 paper-based handwritten medical records and each of them consists of a folder with an extremely variable number of sheets, as the pages are incremented after each follow-up visit carried out on patients (including diagnostic tests, other laboratory tests, structured or instrumental tests). Texts contained in medical records are handwritten, so for the integrated use of information clinical data need to be extracted in a structured formal way. Data collected are essentially made from narrative texts (Coronato et al. 2014) describing the patients' everyday life, cognitive disorders, and all signs and symptoms that in most cases lead to the disease outbreak.

The approach adopted could be considered interdisciplinary as it requires interaction with knowledge organization experts, natural language technicians, and medical experts. The final product, which will integrate the results and the resources developed during the project, is represented by a repository accessible both from the members of the project staff and by the final users, mainly represented by domain experts. In the next section, the methodological approach will be described.

Materials and methods

The work starts from a sample of clinical records stored in the CRN database about patients suffering from AD[2]. Specifically, the total number of records is 12,860[3]. In order to make the textual information shareable for the automatic medical entities detection (signs, symptoms), all paper-handwritten clinical records referring just to dead patients suffering from AD have been considered for the digitization and processing through a software for handwritten text recognition. This software has been semi-automatically trained to allow the recognition of several handwriting styles of the doctors who wrote clinical records[4].

Sample definition and clinical records digitization

The first phase concerned the medical records sample acquisition and, consequently, their extraction from the CRN archive for the digitization activity. The arrangement of the paper-based handwritten clinical records in the CRN's archive follows a shelving disposition (Casanova 1928), within which the documents are organized according to chronological order (Lodolini 2011). For the purpose of this study, a sample of medical records of dead patients has been selected, but only the anamnesis section has been taken into account.

[1] Regional Neurogenetic Centre (CRN), Lamezia Terme, Catanzaro (CZ), Italy.

[2] The study has been approved by the Ethics Committee of University of Calabria, Italy (Protocol Number 2022-UCAL-PRG-0008403 08/02/2022).

[3] The distribution is of 40,5% male and 59,5% female, the majority of them coming from the province of Catanzaro (Calabria Provincial Capital), and from other provinces of Calabria. Only 13.8% come from other Italian regions.

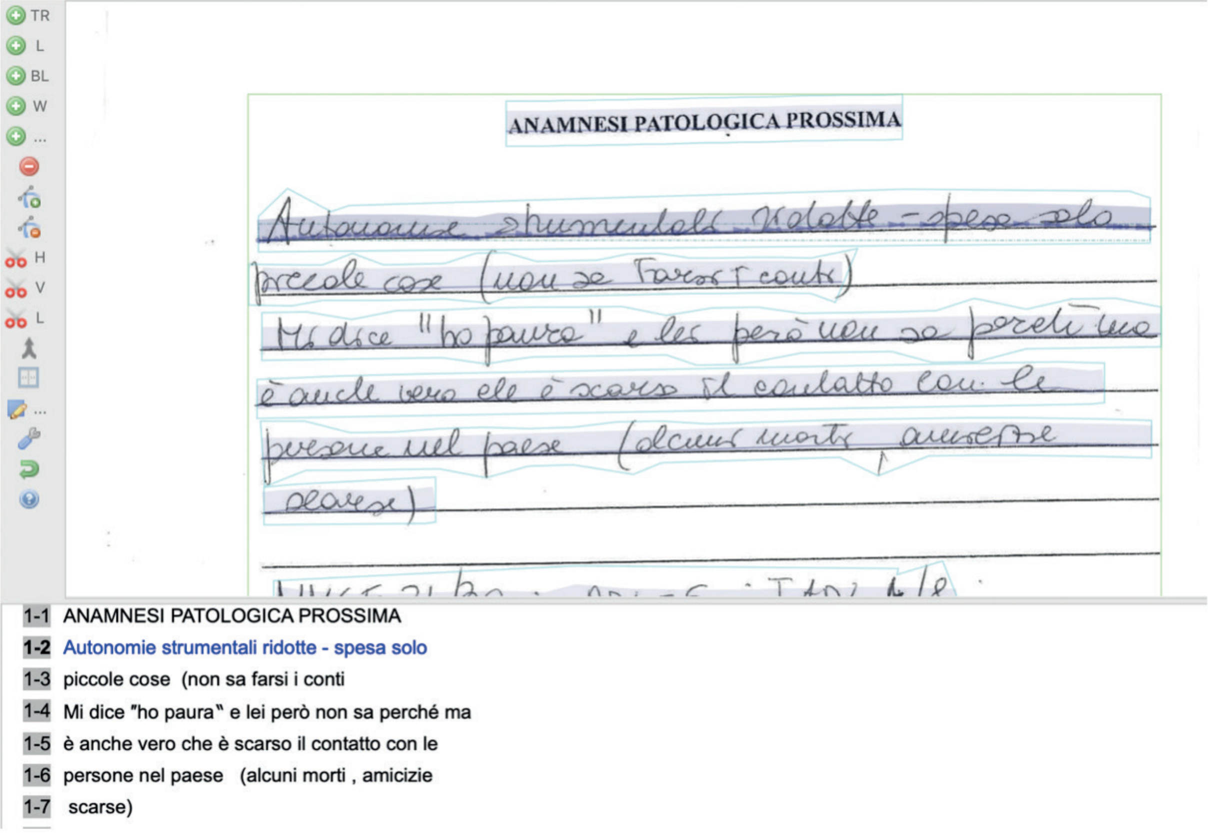
[4] To adequately train the tool, two different training sets have been set up corresponding to two different doctors' handwriting styles.

JLIS.it vol. 14, no. 2 (May 2023)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-522

Text recognition

In this second stage, part of the digitized records has been transcribed in order to obtain a reusable file format to be treated in the categories' identification task. To carry out this process Transkribus software has been employed. This text recognition tool specifically works on handwritten documents, and it offers a way to transcribe line by line the sections of these latter by providing a set for training the association of the characters' recognition every time a new document is imported written by the same authors (see Figure 1).



1-1 ANAMNESI PATOLOGICA PROSSIMA
1-2 Autonomie strumentali ridotte - spesa solo
1-3 piccole cose (non sa farsi i conti
1-4 Mi dice "ho paura" e lei però non sa perché ma
1-5 è anche vero che è scarso il contatto con le
1-6 persone nel paese (alcuni morti , amicizie
1-7 scarse)

Figure 1. Extract from the Transkribus working environment on a clinical digitized record.

As depicted in the previous image, each line of the digitized clinic record corresponds to a region of the document. In this way, Transkribus allows users to insert the matching transcription of the characters and, consequently, learns how to identify the future writing styles. For this very case of study, Transkribus has been deployed to perform the model training over 100 clinical records of dead patients[5] with confirmed AD syndrome[6], consisting of 243 pages subdivided as follows:

[5] The sample number has been set to 100 since this activity took its basis from an initial phase of clinical records' digitization, the number will be increased.

[6] The digitization will be enriched by other clinical records and the research activity will describe the progression of the symptoms from Mild Cognitive Impairment (MCI) to AD for the identification of the categories referring to patients suffering from these syndromes (Petersen and Negash 2008).

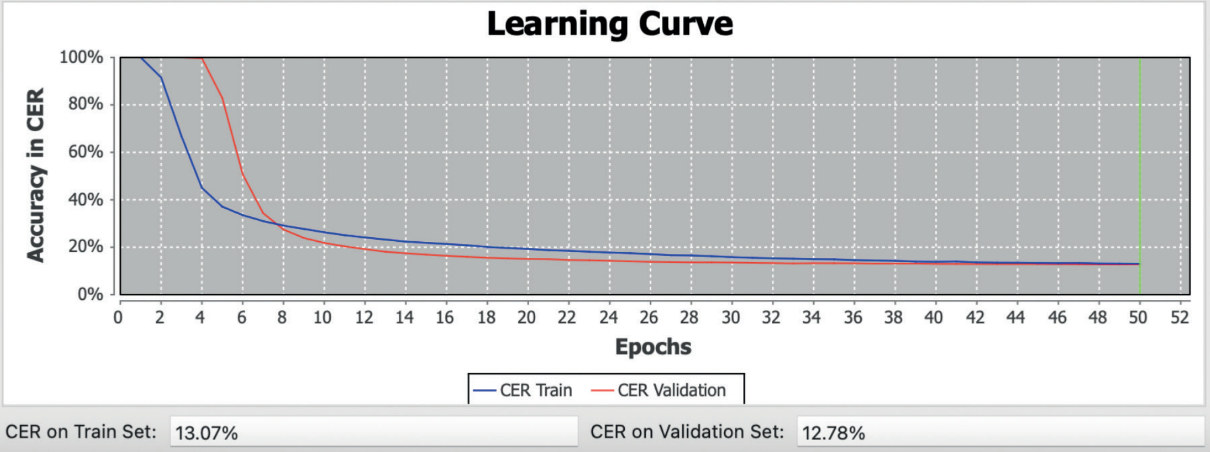
Table 1. Details of the training model for Transkribus.

Train set	Validation set	n. words	n. lines	n. epochs
233	*10*	*32730*	*5624*	*50*

On Transkribus this procedure is named Handwritten Text Recognition (HTR+). It implies the training of a set that successively tests itself over a test set. It runs over 50 document regions: changes in the numbers impact the length of the process: the more epochs users choose, the longer the activity will take. Figure 2 shows the accuracy percentages of the model trained to process the documents of neurodegenerative records automatically. The y-axis represents the "Accuracy in CER", where CER means the Character Error Rate detected during the transcription process by the model, this curve begins at a level of 100, and it decreases alongside the improvement of the model performances (indeed, the blue line is the progress of the training and red that of the evaluations over the test set). As indicated by the software main webpage:

> The value for the Test Set is the most significant as it shows how the HTR+ performs on pages that it has not been trained on. Results with a CER of 10% or below can be seen as very efficient for automated transcription. Results with a CER of 20-30% are sufficient to work with powerful Keyword Spotting technology.[7]

In this case, the CER on the train set corresponds to 13.07%, and this can be due to the fact that parts of the scanned clinical records were marked by several blank sections or some letters, such as the 'p' or the 'g' and 'q', the software was not able to correctly identify for the overlapping line of the letter with the others on the next rows. With this sample the CER on the validation set can be considered sufficiently at a good level considering that the two lines (the blue and the red) match at the end of the curve, meaning that the error is minimized as long as the training progresses.



CER on Train Set: 13.07%
CER on Validation Set: 12.78%

Figure 2. Levels of accuracy in the transcriptions.

[7] For more details: https://readcoop.eu/transkribus/howto/how-to-train-a-handwritten-text-recognition-model-in-transkribus/

Neurodegenerative categories matching with expressions

One goal of the study has been targeted to increase the clinical information about patients treated at CRN by extracting them through NLP techniques, since, to date, data about patients are manually imported into CRN database. The database has represented a solid starting point for implementing a network connection system between the clinical symptoms and signs sentence descriptions within the records and the corresponding categories. Along with the supervision of the CRN physicians and the analysis of the previous works on this subject, this study focused on a categories framework systematization onto two levels: (i) three top categories that have been, in turn, declined in (ii) sub-categories. The following list is meant to show the subdivision employed to reach an automatic identification of AD signs and symptoms descriptions.

	main categories		
	cognitive	**BPSD**[8]	**motor**
subcategories	memory	behavior	extrapyramidal signs
	orientation	affective disorders	
	speech disorders	psychosis	
	agnosia	emotionality	
	apraxia	sleep disorders	
	planning		
	handwriting		

Once defined this flat top-down signs and symptoms structure, the methodology pursued in this study has been based on the identification of the expressions used by doctors in their descriptions of clinical events within the clinical records to be linked to the categories and sub-categories, as shown in the following figures (Figure 3, Figure 4, Figure 5).

[8] For more details about Behavioral and Psychological Symptoms of Dementia see Cerejeira et al. (2012).

JLIS.it vol. 14, no. 2 (May 2023)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.36253/jlis.it-522

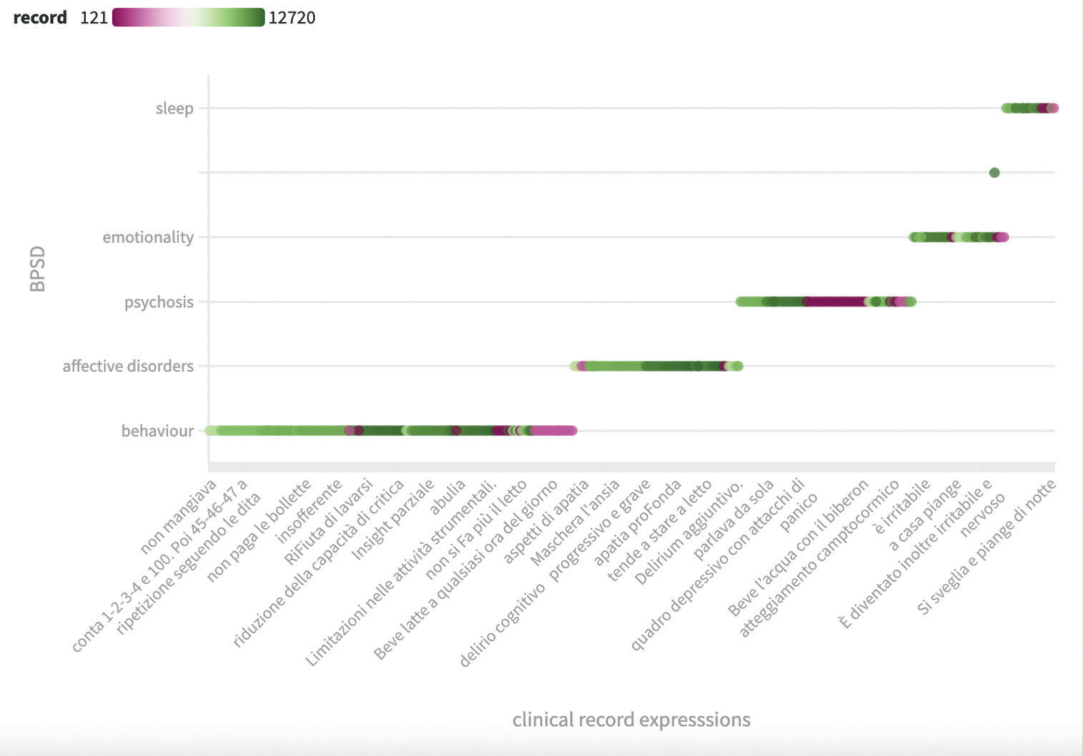


Figure 3. Distribution of subclasses in BPSD.

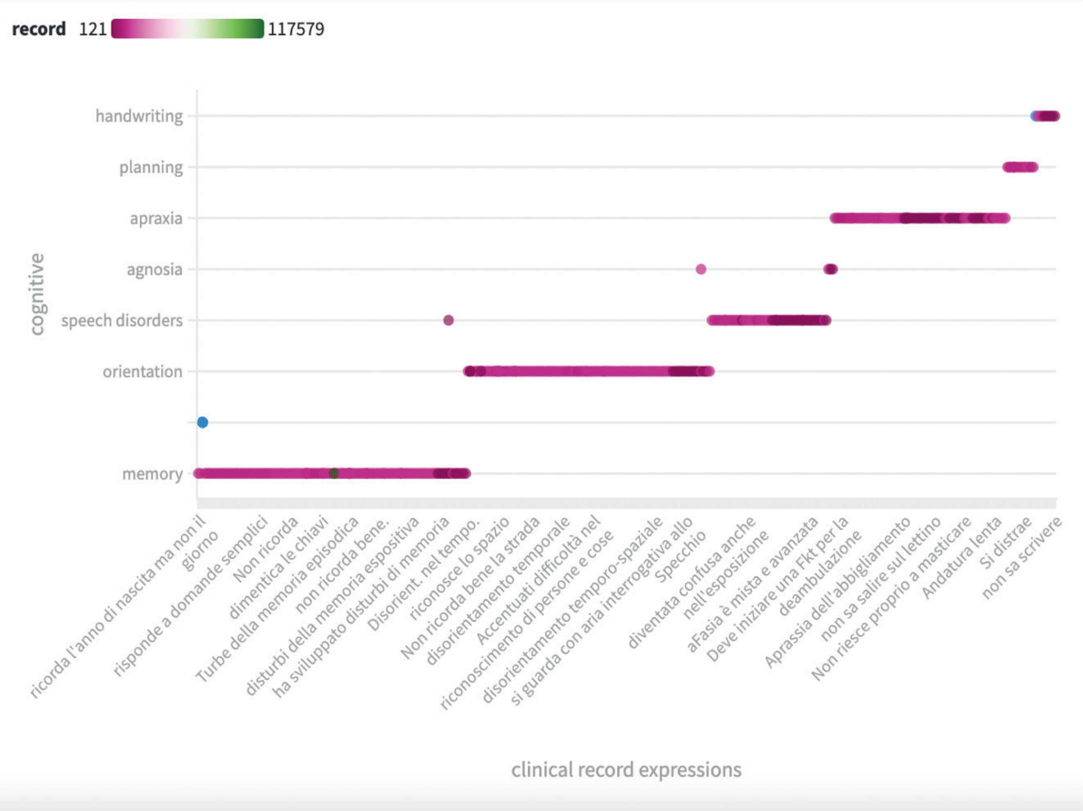


Figure 4. Distribution subclasses Cognitive.

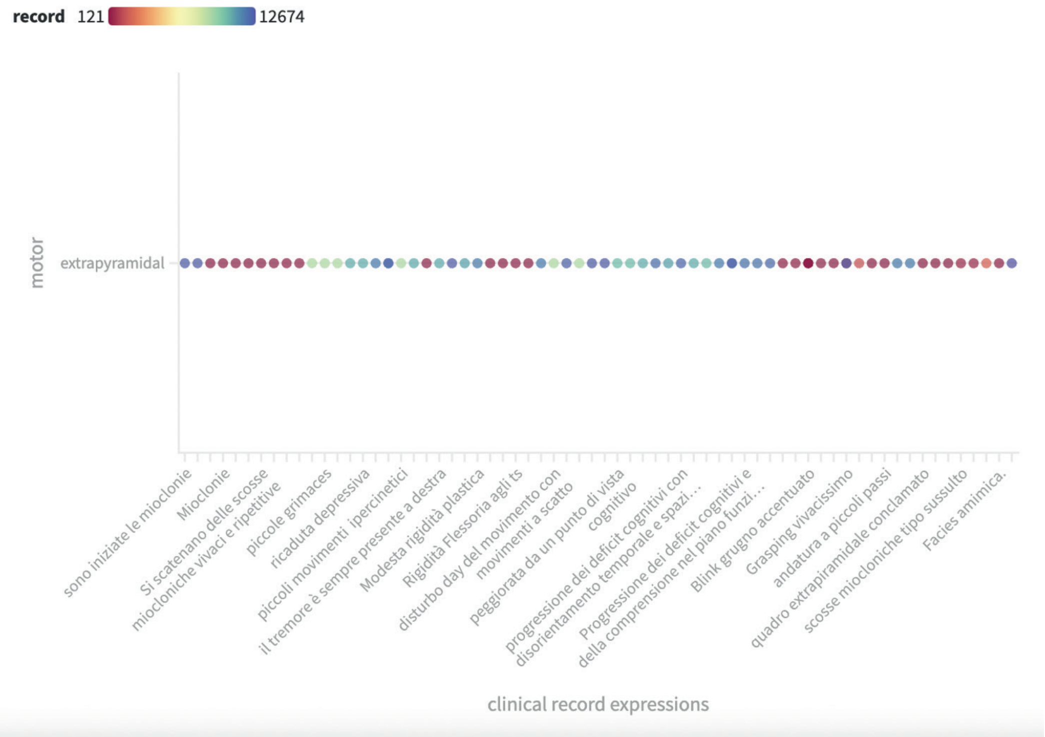


Figure 5. Distribution subclasses Motor.

The association of these expressions has been supported by a preliminary investigation of typical sentences used by the physicians in their descriptions within the clinical records' anamnesis compilation. This task implied the supervision of medical experts who supported the creation of a list of phrases per each category in order to develop a reliable training set for the future automatic identification of the matching expression plus categories. The total number of expressions retrieved in the train set sample is partitioned as the Table 2 shows:

Table 2. Clinical expressions partitioning.

		n. expressions		n. expressions		n. expressions
Top categories	**cognitive**	452	**BPSD**	379	**motor**	71
Sub-categories	memory	145	behavior	165	extrapyramidal signs	71
	orientation	128	affective disorders	75		
	speech disorders	58	psychosis	76		
	agnosia	4	emotionality	40		
	apraxia	88	sleep disorders	23		
	planning	16				
	handwriting	12				

Figure 6 depicts a scatter plot for the expressions related to each sub-categories. The following sections will detail the whole process developed to set a methodology aimed at automatically discovering the phrase segmentations related to the neurodegenerative signs and symptoms by implementing a machine learning schema.

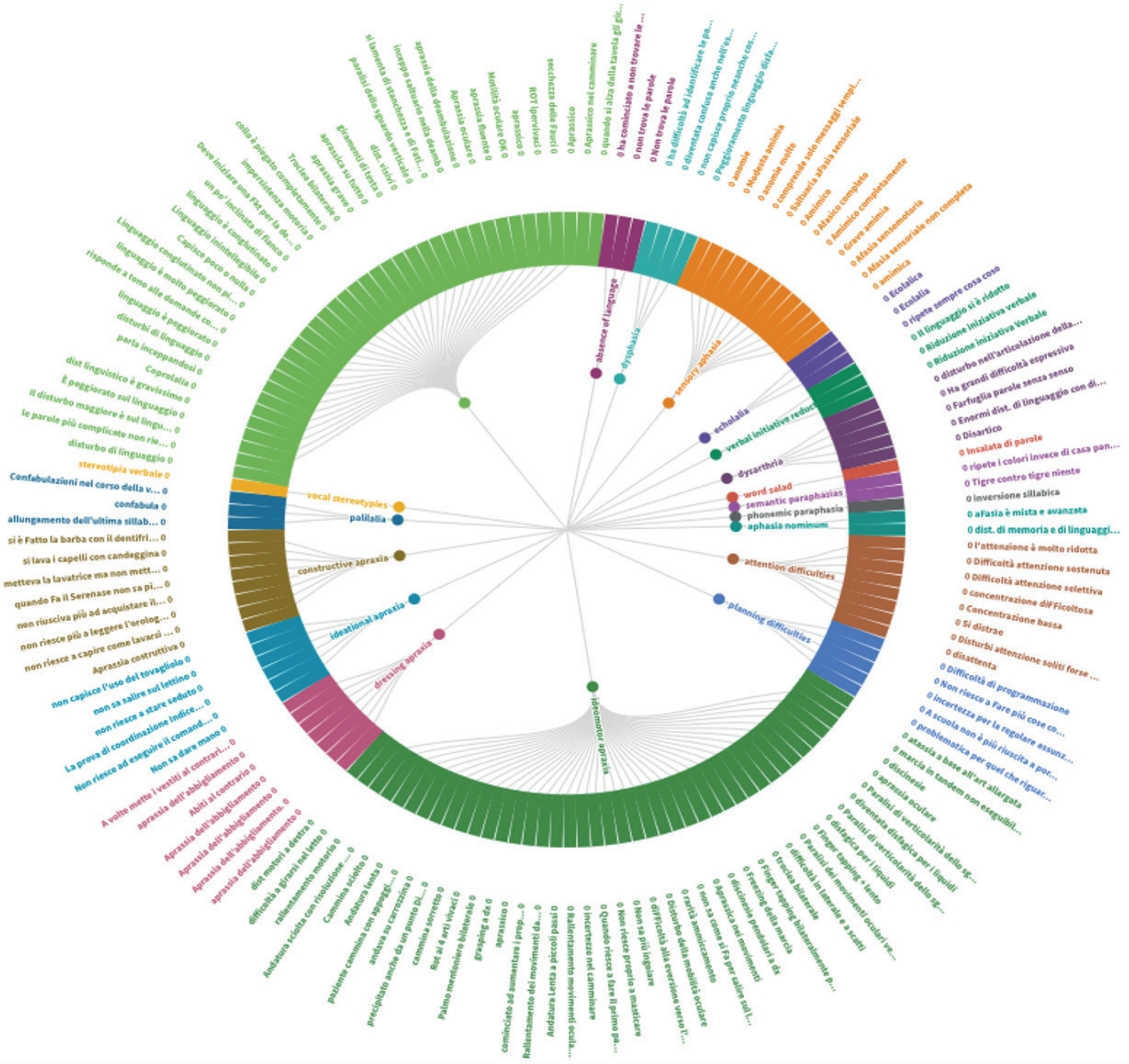


Figure 6. Scatterplot depicting the tags included in the source EHR corpus.

Classifying Alzheimer-related indicators

In electronic medical records, health indicators, medications, laboratory values, symptoms, and personal history are typically embedded in free text form as clinical, hospitalization, and intervention reports, progress notes, and discharge summaries. Many NLP tasks can be conducted on these corpora, we will focus on extracting cognitive, BPSD, and motor Alzheimer-related indicators. Different NLP methods can automate the identification and classification of linguistic entities that describe these essential concepts for a given domain, but they are quite challenging to be applied given the unstructured nature of linguistic data in medical records in the healthcare domain (Li et al. 2021). Less rigid methods, such as rule-based ones (Mykowiecka, Marciniak, and Kupść 2009), use token rules and regular expressions with some characteristics of the entities of interest to extract said entities. Finally, corpus-based methods use indicators from text corpora such as statistical information coupled with machine learning approaches for identifying and extracting these entities. Named-entity recognition tasks, knowledge extraction, and biomedical entities extraction, to cite a few, are all tasks that heavily rely on these processes (Lafferty, McCallum, and Pereira 2001; Settles 2004; Wu et al. 2015; Huang, Xu, and Yu 2015; Chalapathy, Borzeshi, and Piccardi 2016; Si et al. 2019).

Rule-based methods can be time-consuming to build and are prone to contextual conflicts, especially with more complex data, requiring a significant amount of human effort to build a complete set of tags, patterns, and domain-specific rules. For this, it results difficult to create a comprehensive and thorough list of rules due to the ever-evolving variability of the terms contained in the documentation under study. With these methods, however, the results are often satisfactory from an accuracy point of view, in terms of correlation between the exact expressions to be retrieved from the clinical records and the association to a pre-defined set of categories. Secondly, unknown and novel terms or rules are introduced unceasingly in active domains such as the healthcare, clinical or biomedical fields. In order to avoid the drawbacks of manual rules, machine learning approaches were proposed quite early on for NER and extraction tasks, with the usage, among other methods, of SVMs (Wu et al. 2015) and CRFs (Lafferty, McCallum, and Pereira 2001; Settles 2004; Si et al. 2019) for the classification and categorization step.

The neural approach to construct word representation (as well as sentences or document representations) can be seen as a crucial breakthrough in machine learning for NLP. Several methods exist for obtaining the word representations of all words in a predefined vocabulary of fixed size from textual corpora (Mikolov et al. 2013; Bojanowski et al. 2017; Devlin et al. 2019). Learning these representations is done in conjunction with training a neural network on a task, such as a document classification one. Thus, a matrix of weights from the network is called an embedding matrix. It can also be an unsupervised process, using statistical methods to represent the words in the corpus, as done in the earliest distributional methods.

Lower computation complexity is one of the main advantages of using the dense, low-dimensional vectors obtained from these methods compared to those obtained with classical distributional methods, eliminating the "curse of dimensionality" problem that early distributional methods based on high-dimensional co-occurrence matrices had. Furthermore, most neural methods output dense vector representations. The main advantage of these dense representations is their power of generalization. By choosing a small size for the word embeddings, the model is forced to choose the most relevant descriptors to populate the embedding matrix, discarding a good amount of the noise naturally existing in the corpus (Mikolov et al. 2013). These word representations are then used as input for actual task-oriented methods. In recent years, deep neural networks helped secure significant progress in NER and medical concept extraction by eliminating the necessity of feature engineering.

As shown in Figure 7, to process the X_t element, the model combines the representation of the input sequence up to the X_{t-1} element with the information of this new X_t element, thus creating a new state representing the input sequence up to the X_t element. For this reason, by maintaining a state vector that represents each element after it has been processed, it is impossible to parallelize the calculations, which is one of the major drawbacks of these recurrent models.

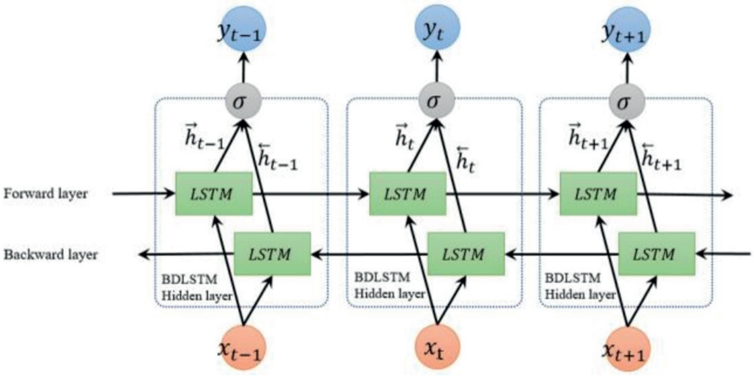


Figure 7. Layer description of a BiLSTM model.

Recurrent Neural Networks (RNNs) can keep track of sentence structure and various dependencies and allow information to be persistent over the network. However, vanilla RNNs often struggle to learn long-term temporal dependencies since their gradients can explode or completely vanish over multiple time steps. The vanilla RNN cell can then be replaced by a Long Short-Term Memory (LSTM) cell (Schuster and Paliwal 1997) or Bidirectional Long-Short Term Memory (BiLSTM) cell (Graves, Fernàndez, and Schmidhuber 2005) to solve this issue via a set of different gates. The addition of a CRF layer was often shown to surpass simple LSTM models for both NER and MCE (Chalapathy, Borzeshi, and Piccardi 2016; Panchendrarajan and Amaresan 2018). Conditional random fields are a class type of statistical modeling methods for prediction tasks where contextual information, i.e., the state of the neighboring tokens, affects the current prediction.

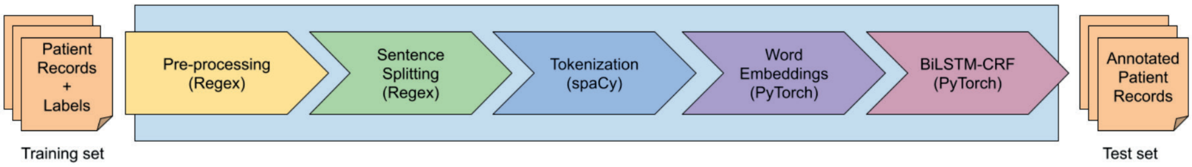


Figure 8. The NLP pipeline of the proposed work.

Our RNN uses two vertically stacked and fully-connected BiLSTM with a CRF layer on top, each LSTM cell uses 256 hidden units, and its dropout is set to 0.3. We only keep sequences that are 50 tokens longer or shorter and tagged expressions up to 8 tokens. We use the training-test sets with 30% withheld for the test sets. We train our model on 50 epochs, with Adam with Nesterov momentum (NAdam) as an optimization algorithm. Figure 9 displays the details of the BiLSTM-CRF module for sequence labeling:

JLIS.it vol. 14, no. 2 (May 2023)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.36253/jlis.it-522

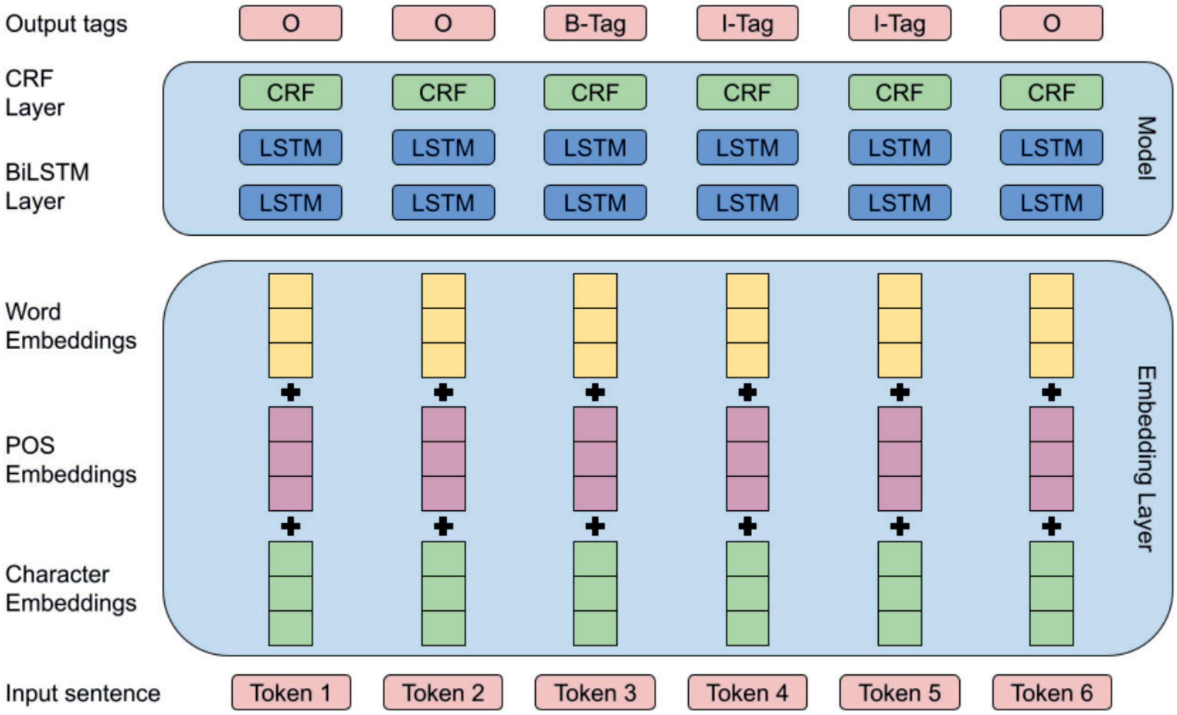


Figure 9. Diagram of the models architecture. The top part covers the language modeling and the tokens extraction and classification, the bottom parts show the different embeddings used to represent the tokens.

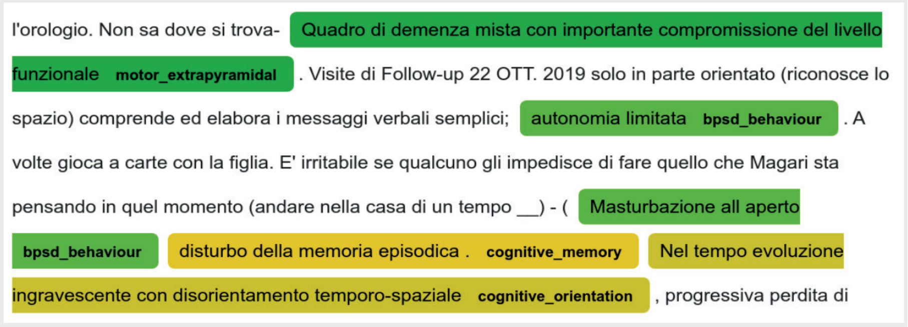
In this work, we tackle the problem with an end-to-end architecture. Given a dataset (split into training-test sets) and a set of entities with labels, the steps undertaken have been the following:

1. **Text preprocessing:** uses an extensive set of regular expressions to clean and process the text. This step is crucial for any NLP task, and it transforms text into a more digestible form so that the methods and algorithms can perform better. This step is even more crucial in tasks where records are used since the records are often unstructured, free-form, and not normalized.
2. **Sentence splitting:** splits the medical record into sentences by relying on a set of regular expression-based rules that define sentence breaks.
3. **Word tokenizing:** splits the sentences into meaningful segments, i.e, tokens, using spaCy.
4. **Token embeddings:** each token is represented using three different embedding types. Word embeddings are typically learned using words from the corpus vocabulary during the training phase. We conjointly learn character embeddings and POS embeddings: these two types of embeddings don't encode the same information that word embeddings contain. Character-level embeddings can be considered encoded lexical information, and POS embeddings encode syntactic context.
5. **Entity extraction:** the model learns the embeddings of the given tokens and directly uses them to predict the label for each token. We use the tags and sub-tags, the entities provided by the doctors, and the "I-O-B" labels for the tags. "I-O-B" Tagging is a standard tagging format for tagging tokens in tasks like name entity recognition. The "B-" prefix indicates

that the tag is the beginning of a chunk, and an "I-" prefix indicates that the tag is inside a chunk. An "O" tag indicates that a token belongs to no predefined entity and indicates that that token is not to be extracted.

Results

In this section we present the result of the evaluation of the proposed methods results using the test corpora. We report the precision, recall, and F1-score, the classical evaluation metrics for entity recognition and sequence labeling tasks. 70% of the dataset is used for training and 30% for testing.



l'orologio. Non sa dove si trova- Quadro di demenza mista con importante compromissione del livello funzionale motor_extrapyramidal . Visite di Follow-up 22 OTT. 2019 solo in parte orientato (riconosce lo spazio) comprende ed elabora i messaggi verbali semplici; autonomia limitata bpsd_behaviour . A volte gioca a carte con la figlia. E' irritabile se qualcuno gli impedisce di fare quello che Magari sta pensando in quel momento (andare nella casa di un tempo __) - (Masturbazione all aperto bpsd_behaviour disturbo della memoria episodica . cognitive_memory Nel tempo evoluzione ingravescente con disorientamento temporo-spaziale cognitive_orientation , progressiva perdita di

Figure 10. Excerpt from outputs of our algorithm for one of the clinical records. The tagging is learned via the neural network, using "displaCy".[9]

The data is randomly split before training the model and for each run, the seed is randomly initialized too. Each configuration is run three times for our experiments, and the reported results are the average of these runs. Both methods aim to extract Alzheimer's indicators in uploaded clinical records corresponding to the following categories of medical reports: cognitive, BPSD and motor. The following table reports the results for the main categories:

Table 3. Analysis (in % Precision, Recall and F1-score) of the model's outputs for the extraction task. This is an averaging of 3 runs using only the category labels.

Top categories	*Precision*	*Recall*	*F1-Score*
cognitive	0.17	0.37	0.24
BPSD	0.23	0.53	0.31
motor	0.19	0.72	0.30

[9] https://explosion.ai/demos/displacy

Discussion

The results differ for most classes, some of these are much more represented in the task corpus than others. The recall is much higher than precision, indicating that too many false positives can be found in the output lists. There can be five main types of problems for each predicted entity and a set of ground-truth entities E_T:

1. Type-1: e_p is not present in E_T, false positive.
2. Type-2: An E_T entity is not predicted, false negative.
3. Type-3: e_p and an E_T entity have the same span but different labels.
4. Type-4: e_p and an E_T entity have overlapping spans and different labels.
5. Type-5: e_p and an E_T entity have overlapping spans and the same labels.

Table 4 presents examples of entities extracted by our model:

Table 4. Examples of entities extracted by the model and matching errors.

Entity	**Expected Tags (from E_T)**	**Actual Tags (for e_p)**	**Type of Error**
modesta rigidità plastica	B-Motor, I-Motor, I-Motor	B-Motor, I-Motor, I-Motor	No error
riduzione iniziativa verbale	B-Cognitive, I-Cognitive, I-Cognitive	B-Cognitive, I-Cognitive, I-Cognitive	No error
voluta uscire più da sola	O, O, O, O, O	B-Motor, I-Motor, I-Motor, I-Motor, I-Motor	Type-1
episodi di disorientamento spaziale	B-Cognitive, I-Cognitive, I-Cognitive, I-Cognitive	O, O, O, O	Type-2
non sa dare mano	B-Cognitive, I-Cognitive, I-Cognitive, I-Cognitive	B-Motor, I-Motor, I-Motor, I-Motor	Type-3
si guarda circospetto dietro	B-BPSD, I-BPSD, I-BPSD, I-BPSD	O, B-Motor, I-Motor, O	Type-4
sviluppato un atteggiamento disforico	B-BPSD, I-BPSD, I-BPSD, I-BPSD	O, O, B-BPSD, I-BPSD	Type-5

Using morpho-syntactic, lexical, semantic, and distributional information allows us to use much richer information as input data for our model. In future work, an ablation study will be conducted with the different representations used to test the contribution of each and every vector representation in the final scores.

The results can directly suffer from the small amount of input data. Transfer learning (Ruder et al. 2019) can be used here to alleviate this issue, it is the set of methods that allow the transfer of knowledge acquired from solving a given problem to another problem or from a domain to another. Our models require significant resources to be tuned appropriately. However, by using pre-trained models as a starting point, transfer learning allows to train or fine-tune our model without needing much training data.

The perspective is to take advantage of unsupervised training methods (such as BERT language model, Devlin et al. 2019) in future work. It must be noted that our entity extraction work is mainly preliminary, and the methods presented here can be considered baselines for further work.

Conclusion

This study configured a methodology to retrieve categorized medical expressions to define the correct classification of AD's signs and symptoms. The purpose of this investigation addressed the identification of typical sentences in the digitized clinical records to be automatically mapped with a two-level system categorization. The research work developed a multidisciplinary approach: from paper-based handwritten clinical records to a digitized corpus from which to detect in an automatic way salient medical information to be mapped with normalized neurodegenerative-related categories and sub-categories. The corpus analyzed has been built from anamnesis texts totally written in natural language that for its nature is rich of irregular expressions. This has impacted the configuration of a twofold categorization model meant to contain the mapping between the medical recursive expressions related to AD signs and symptoms and the sub-categories selected with the supervision of the physicians working in this sector. In future work activities, along with the enrichment of the documents, the analyses will be targeted to the classification of the clinical records according to the declared syndromes doctors have assigned to each patient and to the correlation of these diseases with the corresponding automatic symptoms detection.

References

Alzheimer's Association. 2016. «2016 Alzheimer's disease facts and figures». *Alzheimer's & Dementia* 12 (4): 459–509.

Beeler, Patrick Emanuel, David Westfall Bates, and Balthasar Luzius Hug. 2014. «Clinical decision support systems». *Swiss Medical Weekly* 144 (w14073): 1–7. https://doi.org/doi.org/10.4414/smw.2014.14073.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. «Enriching Word Vectors with Subword Information». *Transactions of the Association for Computational Linguistics* 5: 135–46.

Bruni, Amalia Cecilia, Livia Bernardi, and Carlo Gabelli. 2020. «From beta amyloid to altered proteostasis in Alzheimer's disease». *Ageing research reviews* 64: 101126.

Bruni, Amalia Cecilia, Livia Bernardi, and Raffaele Maletta. 2021. «Evolution of genetic testing supports precision medicine for caring Alzheimer's disease patients». *Current Opinion in Pharmacology* 60: 275–80.

Casanova, Eugenio. 1928. *Archivistica*. Siena: Stab. arti grafiche Lazzeri.

Chalapathy, Raghavendra, Ehsan Zare Borzeshi, and Massimo Piccardi. 2016. «Bidirectional LSTM-CRF for Clinical Concept Extraction». https://doi.org/10.48550/arXiv.1610.05858.

Coronato, Antonio, Giuseppe Di Pietro, Amalia Cecilia Bruni, Erika Pasceri, Maria Teresa Chiaravalloti, and Giovanni Paragliola. 2014. «ALPHA: an eAsy inteLligent service Platform for Healthy Ageing». In *Ambient Assisted Living*, edited by Bruno Andò, Pietro Siciliano, Vincenzo Marletta, and Andrea Monteriù. Springer.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». http://arxiv.org/abs/1810.04805.

Graves, Alex, Santiago Fernàndez, and Jürgen Schmidhuber. 2005. «Bidirectional LSTM networks for improved phoneme classification and recognition». In *ICANN'05: Proceedings of the 15th international conference on Artificial neural networks: formal models and their applications – Volume Part II*, edited by Duch Włodzisław, Janusz Kacprzyk, Zadrozny Sławomi, and Oja Erkku. Berlin, Heidelberg: Springer-Verlag.

Hassanzadeh, Hamed, Anthony Nguyen, and Bevan Koopman. 2016. «Evaluation of Medical Concept Annotation Systems on Clinical Records». In *Proceedings of the Australasian Language Technology Association Workshop 2016*, 15–24. https://aclanthology.org/U16-1002.

Huang, Zhiheng, Wei Xu, and Kai Yu. 2015. «Bidirectional LSTM-CRF Models for Sequence Tagging». http://arxiv.org/abs/1508.01991.

Kharbanda, Elyse O., Steve E. Asche, Alan R. Sinaiko, Heidi L. Ekstrom, James D. Nordin, Nancy E. Sherwood, Patricia L. Fontaine, Steven P. Dehmer, Deepika Appana, and Patrick O'Connor. 2018. «Clinical Decision Support for Recognition and Management of Hypertension: A Randomized Trial». *Pediatrics* 141 (2): e20172954. https://doi.org/10.1542/peds.2017-2954.

Klassen, Prescott, Fei Xia, e Meliha Yetisgen. 2016. «Annotating and Detecting Medical Events in Clinical Notes». In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 3417–21. European Language Resources Association. https://aclanthology.org/L16-1545.pdf.

Lafferty, John, Andrew McCallum, and Fernando C. N. Pereira. 2001. «Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Dataand Labeling Sequence Data». In *Proceedings of the Eighteenth International Conference on Machine Learning*, 282–89. Morgan Kaufmann Publishers Inc.

Laganà, Valentina, Francesco Bruno, Natalia Altomari, Giulia Bruni, Nicoletta Smirne, Sabrina Curcio, Maria Mirabelli, Rosanna Colao, Gianfranco Puccio, Francesca Frangipane, Chiara Cupidi, Giusy Torchia, Gabriella Muraca, Antonio Malvaso, Desirèe Addesi, Alberto Montesanto, Raffaele Di Lorenzo, Amalia Cecilia Bruni, and Raffaele Maletta. 2022. «Neuropsychiatric or Behavioral and Psychological Symptoms of Dementia (BPSD): Focus on Prevalence and Natural History in Alzheimer's Disease and Frontotemporal Dementia». *Frontiers in Neurology* 13 (June): 832199. https://doi.org/10.3389/fneur.2022.832199.

Li, Irene, Jessica Pan, Jeremy Goldwasser, Neha Verma, Wai Pan Wong, Muhammed Yavuz Nuzumlalı, Benjamin Rosand, Yixin Li, Matthew Zhang, David Chang, R. Andrew Taylor, Harlan M. Krumholz, and Dragomir Radev. 2021. «Neural Natural Language Processing for Unstructured Data in Electronic Health Records: a Review». http://arxiv.org/abs/2107.02975.

Lodolini, Elio. 2011. *Archivistica. Principi e problemi*. Milano: Franco Angeli.

Mazzocchi, Fulvio. 2018. «Knowledge organization system (KOS)». *Knowledge Organization* 45 (1): 54–78.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. «Efficient Estimation of Word Representations in Vector Space». http://arxiv.org/abs/1301.3781.

Mills, Sherri. 2019. «Electronic Health Records and Use of Clinical Decision Support». *Critical Care Nursing Clinics of North America* 31 (2): 125–31. https://doi.org/10.1016/j.cnc.2019.02.006.

Mykowiecka, Agnieszka, Małgorzata Marciniak, and Anna Kupść. 2009. «Rule-Based Information Extraction from Patients' Clinical Data». *Journal of Biomedical Informatics* 42 (5): 923–36. https://doi.org/10.1016/j.jbi.2009.07.007.

Panchendrarajan, Rrubaa, and Aravindh Amaresan. 2018. «Bidirectional LSTM-CRF for Named Entity Recognition». In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation,* 531-540. Hong Kong: Association for Computational Linguistics.

Patel, Pinalkumar, Disha Davey, Vishal Panchal, and Parth Pathak. 2018. «Annotation of a Large Clinical Entity Corpus». In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2033–42. Brussels, Belgium: Association for Computational Linguistics.

Petersen, Ronald C., and Selamawit Negash. 2008. «Mild Cognitive Impairment: An Overview». *CNS Spectrums* 13 (1): 45–53. https://doi.org/10.1017/s1092852900016151.

Ruder, Sebastian, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. «Transfer

Learning in Natural Language Processing». In *Proceedings of the 2019 Conference of the North*, 15–18. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-5004.

Savova, Guergana K, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. «Mayo Clinical Text Analysis and Knowledge Extraction System (CTAKES): Architecture, Component Evaluation and Applications». *Journal of the American Medical Informatics Association* 17 (5): 507–13. https://doi.org/10.1136/jamia.2009.001560.

Schuster, M., and K.K. Paliwal. 1997. «Bidirectional recurrent neural networks». *IEEE Transactions on Signal Processing* 45 (11): 2673–81. https://doi.org/10.1109/78.650093.

Searle, Thomas, Zeljko Kraljevic, Rebecca Bendayan, Daniel Bean, and Richard Dobson. 2019. «MedCATTrainer: A Biomedical Free Text Annotation Interface with Active Learning and Research Use Case Specific Customisation». https://doi.org/10.48550/arXiv.1907.07322.

Settles, Burr. 2004. «Biomedical Named Entity Recognition using Conditional Random Fields and Rich Feature Sets». In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, 107–10. Geneva, Switzerland: Coling.

Shellum, Jane L., Robert R. Freimuth, Steve G. Peters, Rick A. Nishimura, Rajeev Chaudhry, Steve J. Demuth, Amy L. Knopp, Timothy A. Miksch, and Dawn S. Milliner. 2016. «Knowledge as a Service at the Point of Care». *AMIA … Annual Symposium Proceedings. AMIA Symposium* 2016: 1139–48.

Si, Yuqi, Jingqi Wang, Hua Xu, and Kirk Roberts. 2019. «Enhancing Clinical Concept Extraction with Contextual Embeddings». *Journal of the American Medical Informatics Association: JAMIA* 26 (11): 1297–1304. https://doi.org/10.1093/jamia/ocz096.

Spineth, Martin, Andrea Rappelsberger, and Klaus-Peter Adlassnig. 2018. «Implementing CDS Hooks Communication in an Arden-Syntax-Based Clinical Decision Support Platform». *Studies in Health Technology and Informatics* 255: 165–69.

Stewart, Samuel Alan, Maia Elizabeth von Maltzahn, and Syed Sibte Raza Abidi. 2012. «Comparing Metamap to MGrep as a Tool for Mapping Free Text to Formal Medical Lexicons». In *Knowledge Extraction and Consolidation from Social Media (KECSM 2012)*, 63–77.

Tolley, Clare L., Sarah P. Slight, Andrew K. Husband, Neil Watson, and David W. Bates. 2018. «Improving Medication-Related Clinical Decision Support». *American Journal of Health-System Pharmacy* 75 (4): 239–46. https://doi.org/10.2146/ajhp160830.

Wu, Yonghui, Jun Xu, Min Jiang, Yaoyun Zhang, and Hua Xu. 2015. «A Study of Neural Word Embeddings for Named Entity Recognition in Clinical Text». *AMIA … Annual Symposium Proceedings. AMIA Symposium* 2015: 1326–33.