Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Applied Soft Computing Journal 97 (2020) 106779

Contents lists available at ScienceDirect

Applied Soft Computing Journal

journal homepage: www.elsevier.com/locate/asoc

Crosslingual named entity recognition for clinical de-identification applied to a COVID-19 Italian data set

Rosario Catelli [a,b], Francesco Gargiulo [a,*], Valentina Casola [b], Giuseppe De Pietro [a], Hamido Fujita [c,d,e], Massimo Esposito [a]

[a] Institute for High Performance Computing and Networking (ICAR), National Research Council, Naples, Italy
[b] Department of Electrical Engineering and Information Technologies (DIETI), University of Naples Federico II, Naples, Italy
[c] Faculty of Information Technology, Ho Chi Minh City University of Technology (HUTECH), Ho Chi Minh City, Viet Nam
[d] Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Granada, Spain
[e] Faculty of Software and Information Science, Iwate Prefectural University, Iwate, Japan

ARTICLE INFO

Article history:
Received 7 August 2020
Received in revised form 28 September 2020
Accepted 3 October 2020
Available online 9 October 2020

Keywords:
COVID-19
Clinical de-identification
Named entity recognition
Deep learning
Annotated Italian data set

ABSTRACT

The COrona VIrus Disease 19 (COVID-19) pandemic required the work of all global experts to tackle it. Despite the abundance of new studies, privacy laws prevent their dissemination for medical investigations: through clinical de-identification, the Protected Health Information (PHI) contained therein can be anonymized so that medical records can be shared and published. The automation of clinical de-identification through deep learning techniques has proven to be less effective for languages other than English due to the scarcity of data sets. Hence a new Italian de-identification data set has been created from the COVID-19 clinical records made available by the Italian Society of Radiology (SIRM). Therefore, two multi-lingual deep learning systems have been developed for this low-resource language scenario: the objective is to investigate their ability to transfer knowledge between different languages while maintaining the necessary features to correctly perform the Named Entity Recognition task for de-identification. The systems were trained using four different strategies, using both the English Informatics for Integrating Biology & the Bedside (i2b2) 2014 and the new Italian SIRM COVID-19 data sets, then evaluated on the latter. These approaches have demonstrated the effectiveness of cross-lingual transfer learning to de-identify medical records written in a low resource language such as Italian, using one with high resources such as English.

1. Introduction

COrona VIrus Disease 19 (COVID-19) is a global threat opposed by experts, politicians and researchers from around the world [1]. In particular, everyone is rushing to keep pace with the influx of potentially relevant studies related to COVID-19 in order to gain timely knowledge to manage the current pandemic [2]. The availability of these new studies has led to an exponential increase in the amount of textual clinical data to be analyzed: unfortunately, this data cannot be used directly for medical investigations, due to the privacy restrictions provided by the relevant legislation, e.g. the Health Insurance Portability and Accountability Act[1] (HIPAA) in the United States or the General Data Protection Regulation[2] (GDPR) in the European Union.

A fundamental step to allow the sharing and publication of COVID-19 data is the de-identification, widely used in the medical area and termed as clinical de-identification, which aims to avoid the disclosure of a personal identity. But, in order to exploit health information for research purposes, it is necessary to aim at generalization through so-called surrogate terms rather than the deletion of privacy-sensitive information contained in medical records, safeguarding in this way also the readability of the documentation [3]. After proper de-identification, hence anonymization of the data, it is possible to release and share them publicly.

Initially the language domain of interest, i.e. English due to a greater worldwide availability of Electronic Health Records (EHRs), was taken for granted and de-identification challenges were organized by the Informatics for Integrating Biology & the

* Correspondence to: Institute for High Performance Computing and Networking (ICAR), National Research Council, Via Pietro Castellino 111 - 80131, Naples, Italy.
E-mail addresses: rosario.catelli@unina.it, rosario.catelli@icar.cnr.it (R. Catelli), francesco.gargiulo@icar.cnr.it (F. Gargiulo), valentina.casola@unina.it (V. Casola), giuseppe.depietro@icar.cnr.it (G.D. Pietro), h.fujita@hutech.edu.vn, hfujita-799@acm.org, issam@iwate-pu.ac.jp (H. Fujita), massimo.esposito@icar.cnr.it (M. Esposito).

https://doi.org/10.1016/j.asoc.2020.106779
1568-4946/

[1] https://www.hhs.gov/hipaa.
[2] https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu.

Bedside[3] (i2b2) group, founder in English. The problem of de-identification has benefited from the use of Natural Language Processing (NLP) techniques such as Named Entity Recognition (NER) in which entities have been assimilated, for example, to HIPAA identifiers, which are headed as the entities to be de-identified and then made anonymous through appropriate surrogates. Nowadays, the state of the art NER relies on deep learning techniques, thanks to the high amount of data available, to make the system able to recognize interesting entities.

Unfortunately, experiences in languages other than English remained confined to a few sporadic cases, such as ShARe/CLEF eHealth Evaluation Lab[4] and IberLEF 2019[5] with specific traces also in French and Spanish, as well as a few case studies in other languages. Hence, outside the Anglo-Saxon-speaking countries, the use of the best performing deep learning methods is severely limited both by the lack of resources suitable for their exploitation, i.e. large data sets, and by poor experimentation on such languages, which are consequently defined as low-resource languages.

This article tries to improve both these aspects of the literature, experimenting on a new data set based on COVID-19 medical records for a low resource language like Italian. The aim is to investigate the ability of cross-linguistic methods to transfer knowledge between different languages while retaining the features necessary to correctly perform NER, which is the basis of de-identification and anonymization systems. As far as is known, there are no multilingual approaches specifically designed for this task, nor any knowledge of the performance of existing systems with respect to the Italian language, which is the subject of study.

Two different system architectures have been tested that showed state of the art performance. To this end, the i2b2 2014 training data set in English was used and, in accordance with the i2b2 annotation guidelines [4], an Italian data set was created from the COVID-19 medical records provided by the Italian Society of Radiology[6] (SIRM). Different training approaches have been tested, both monolingual in English with zero-shot test on Italian, and cross-language with mixed language training and test on Italian. The results were promising and allowed to identify the best architectural solution for low-resource Italian language cases for the clinical de-identification task. The application of the method described here would allow a better de-identification of Italian COVID-19 medical records, speeding up their public dissemination: more accurate anonymization of privacy-sensitive information reduces the distrust of institutions to release data.

The remainder of this paper is structured as follows. In Section 2 the most important works related to the de-identification topic and a general background on cross-lingual approaches are drawn. In Section 3 both the data sets and the architectures used are described. In Section 4 the experimental setup and the evaluation metrics are explained, while in Section 5 the results are analyzed and discussed. Finally, in Section 6, conclusions are drawn and future works are advanced.

2. Background and related works

In the following the development of deep learning techniques for clinical de-identification is described in Section 2.1. Then in Section 2.2 the cross-lingual approaches are introduced.

2.1. Deep learning for de-identification

In terms of information, Protected Health Information (PHI) can be assimilated to an *named entity* [5]. The recognition of such entities occurs by implementing what is called NER, defined as *clinical* if applied on medical records in the form of unstructured text. The purpose is to be able to use the data contained in them, therefore it is necessary to identify the PHI and replace them with valid surrogates, a process called *anonymization*. For this reason it is important to also recognize the type to which the entity belongs, and it would be more correct to refer to Named Entity Recognition and Classification (NERC) [5].

Manual labeling of PHI, as stated by [6], does not allow either to reduce costs and errors related to human annotators or to outsource activity due to confidential data access. Among the automatic systems developed over time, those based on rules and machine learning are widely described by [7,8].

More recently, deep learning techniques have been developed. Indeed, algorithms and NER systems based on deep learning [9–12] have generally improved performance, also for clinical NER [6,13], exploiting two important elements: embeddings [14], that is a numerical representation of textual elements, and neural networks [9,11,12,15–17]. These findings have been applied to the clinical domain [18–21] then to de-identification [13,22] for which several routes have been tried: for instance, [23] have tried to better integrate the context, increasing the performance of the NER at the expense of engineering time, instead [24] have exploited a stacked learning ensemble, more effective but more time and resource intensive.

The latest architectures based on transformers, such as Bidirectional Encoder Representations from Transformers (BERT) [25] and subsequent variants related to the biomedical world [26,27], have paved the way for the use of techniques based on attention mechanisms [28]. Such techniques have been tried in different fields, such as chemical [29] or news [30]. Bidirectional Long Short-Term Memory + Conditional Random Field (Bi-LSTM+CRF) architectures continued to prove to be competitive in NER specific tasks, especially for low-resource languages, such as de-identification in Spanish [31] and, for instance, state-of-the-art results were also achieved by combining Bi-LSTM+CRF architecture with BERT embeddings [32].

In a completely transversal way to the task of clinical de-identification, important results have been obtained from [33,34]: the former have identified strategies to improve the NER in biomedical field by increasing the capacity of generalization of the CRF component, while the latter have provided an interesting reference point for gender assessment in the systems of named entities recognition, observing a lower recognition of female names as “Person” type entities.

2.2. Cross-lingual transfer learning approaches

This section examines how low-resource languages have been managed over time. In detail the techniques that preceded BERT are described in Section 2.2.1, while BERT and its multilingual version are illustrated in Section 2.2.2.

2.2.1. Non BERT-based multilingual techniques

In the field of transfer learning, a branch of particular interest applied to the NLP domain is that of cross-lingual transfer learning which, as stated by [35] is a type of transductive transfer learning where the source and target domain are different, i.e. training and prediction take place on corpora in different languages, and cross-linguistic transfer occurs through the use

[3] https://portal.dbmi.hms.harvard.edu/.
[4] https://clefehealth.imag.fr/.
[5] https://sites.google.com/view/iberlef-2019.
[6] https://www.sirm.org/.

of a single cross-linguistic representation space. Initially, task-specific models were popular, based on a coarse-grained representation such as Part of Speech (PoS) tags, and then exploited a delexicalized parser [36].

Recently, cross-lingual word embeddings have started to be used in combination with specific neural architectures, obtaining interesting results in various tasks, such as PoS tagging [37], NER [38] and dependency parsing [39]. In addition, several studies have been carried out analyzing the effects of different ways of constructing cross-lingual space: for example, [40] analyzed methods for learning cross-lingual embedding through both joint training and post-training mapping of monolingual embeddings, whereas [41] and [42] demonstrated that with the alignment of two monolingual word embedding spaces in unsupervised ways it is possible to get better results.

In detail, [41] introduced Multilingual Unsupervised and Supervised Embeddings[7] (MUSE), created by aligning the embedding spaces of monolingual word embeddings, without using parallel corpora, in an unsupervised way. Then a supervised version of MUSE, crosslingual fastText-based embeddings [43], was also released. These embeddings are generated by aligning the monolingual fastText embeddings in a common space using bilingual dictionaries as ground-truth. Only static embedding vectors have been released and, without the model, it is not possible to generate embeddings for Out-Of-Vocabulary (OOV) words.

Instead, [44] proposed Byte-pair Embeddings (BPEmb) to tackle the Out-Of-Vocabulary (OOV) problem. Based on Byte-Pair Encoding (BPE) [45], BPEmb create each word embedding by composing the necessary sub-word embeddings. In particular, [44] found that BPEmb offer nearly the same accuracy as word embeddings, but at a fraction of the model size, a valuable choice to train small models. BPEmb were released in 275 languages and were successfully used in several cross-lingual scenarios [46–48]. Moreover, [49] showed the importance of sub-word segmentation, due to the absence of any "one-size-fits-all" configuration, because performance is both task- and language-dependent. In addition, [50] noted that sub-word based models perform better than word-based models, such as MUSE and Word2Vec [14], in several low-resources languages scenarios.

Contextual language models, such as Embeddings from Language Models (ELMo) [51] and Flair [52], proved to be superior to static models such as Word2Vec [14] and Global Vectors for Word Representation (GloVe) [53] thanks to the ability to analyze the context, and this further improved performance in cross-lingual scenarios. Flair embeddings [54,55], which constitute a character-based contextual language model on which the Flair NLP framework [52] is based, prompted [56] to test a novel methodology for cross-lingual transfer learning for Japanese NER, based on a Bi-LSTM architecture and embeddings at both word and character level as input.

Furthermore, [57] proposed the Universal Language Model Fine-tuning (ULMFiT), an effective transfer learning method based on an appropriate fine-tuning strategy to improve language models performance, while [58] proposed generative pre-training techniques which led to the Generative Pre-trained Transformer (GPT) language model, which uses an encoder based on transformers [28]. Then [59] extended generative pre-training to cross-lingual models and obtained state of the art results, while [60] tested cross-lingual alignment with ELMo embeddings overcoming the state of the art for zero-shot dependency parsing.

Additionally, [61] experimented a polyglot system based on ELMo, showing relevant results. Indeed, to create a multilingual system, there are two possible alternatives: (i) train a specific model for each language and (ii) train only one model for all languages. In particular, [61] have shown how the choice (ii) provides better results especially in the case of low resources languages thanks to the enrichment of the model with the data of languages that, although different, can be linked together on different aspects of the language (e.g. semantics, morphology, syntax, and so on). Starting both from this principle and encouraging results obtained on Slavic languages by [62], it was decided to consider pre-trained multilingual models on large corpora and fine-tune them on the target language, Italian, which is a low resources language. This way the extremely computationally expensive training procedure can be totally avoided, initializing the model with the multilingual one.

While the world of research has made an effort to organize knowledge in order to better use it against the COVID-19 [63–71], on the other hand a series of research with pandemic focus has followed.

For instance, [72] released, during the COVID-19 global pandemic, a multilingual data set containing more than 5 thousand statements in English, Spanish, French and Spanglish (Spanish + English). This data set was used to study some cross lingual transfer learning techniques related to the Intent Detection task, observing performance improvement in most models with cross lingual training compared to models with mono lingual training. Based on this assumption, both zero shot and cross lingual training approaches were tested.

Finally, [73] have used a LSTM model for COVID-19 prediction, detailing evaluation criteria of the models under analysis and providing a prospective estimate of the total number of cases with the LSTM.

2.2.2. BERT-based multilingual techniques

BERT is a deep contextual language model, based on transformers [28]. Unlike ELMo and GPT, BERT is trained by Cloze Task [74], commonly known as masked language modeling, which is different from classic right-to-left or left-to-right language modeling, allowing it to encode information from both directions in each level freely. Furthermore, BERT also optimizes a target for the classification of the next sentence, so that the paired sentences during training are half consecutive pairs and half random pairs. Lastly, BERT uses a sub-word vocabulary based on the WordPieceModel segmenter [75], a data-based approach to break down a word into sub-words that is more effective than operating at the word level. As demonstrated by [25], BERT is able to achieve high performance in several sentence classification tasks thanks to the fine tuning of the transformer encoder followed by a softmax classification layer fine-tuned for 2-3-4 epochs with a learning rate in the order of e-5: in the case of NER, a sequence of shared softmax classifications produces sequence tagging patterns.

The multilingual BERT (mBERT[8]), differs exclusively for the different training data set consisting of Wikipedia data in 104 languages provided as they are, without the typical links of cross-lingual methods, but appropriately scaled. Leveraging WordPiece, mBERT thus generated is a model in which common sub-words are shared between languages even far apart in the form of a standalone lexicon. Many have recently started investigating the performance of mBERT. Among them, [76] have carried out a series of experiments showing how the transfer also happens in languages in different scripts, although it works better with typologically similar languages. Instead, [77] consider a broader spectrum of NLP tasks, comparing mBERT with different methods of zero-shot cross-lingual transfer and experimenting with different strategies to improve generalization capabilities. Moreover, [78] studied the contribution of the different components

[7] https://github.com/facebookresearch/MUSE.

[8] https://github.com/google-research/bert/blob/master/multilingual.md.

of mBERT to its cross-lingual skills, stressing that the depth of the network is more relevant than the lexical overlap between languages.

Furthermore, [79] found that although mBERT performs well in scenarios with medium and high language resources, non-contextual embedding working at the sub-word level, such as BPEmb, outperforms mBERT in low-resources scenarios. Finally, [80] explored cross-lingual transfer for Danish using several architectures for supervised NER, including Flair, fastText, BPE and both monolingual (Danish) and multilingual BERT, on a modestly-sized training set, testing different training and fine-tuning approaches.

Additionally, [81] used both BERT and Bi-LSTM+CRF architectures to create a drug extraction model to study the ability to respond quickly to emerging diseases such as COVID-19.

Finally, [82] proposed a new *Artificial Intelligence and NLP based Islamic FinTech Model* (based on several NLP techniques, from rules to deep learning) to analyze the impact of the COVID-19 pandemic on the poor and small and medium enterprises, predicting possible future scenarios by leveraging the use of specific taxes of Islamic countries to deal with them.

3. Material and methods

In this section the architectures used and their topological structures are progressively introduced, then the data sets employed are described and finally the different training and fine tuning strategies adopted are explained. For the sake of clearness, in Fig. 1 is given an overview of the research aspects covered by this paper. In detail, Italian Medical Records constitute the primary input information, while English medical records constitute the broader additional information indicated in the Figure with a red arrow and an extended graphical representation. The output is given by PHI predictions that represent the information to anonymize in order to make the Italian input documents compliant with privacy regulations. The central block represents all the different combinations of (i) network topologies, (ii) pre-trained embeddings and (iii) different training strategies as detailed in the yellow balloons and covered hereinafter.

3.1. Network topologies

The system architectures introduced in the following are currently considered the state of the art for NER tasks in NLP: the results obtained in the literature do not allow to identify a significantly superior architecture in the case of clinical de-identification but, depending on the specific scenario (conditioned by language, size of data sets, training strategies and so on), one architecture tends to prevail over the other.

A different discussion deserves the time complexity. Given a sentence of length N, systems based on transformers like BERT process it all together, so the time complexity is O(1) while for a Bi-LSTM+CRF it is O(N) [83,84]: this is mainly due to the fact that transformers were designed to run on parallel hardware architectures (such as GPU, TPU and so on) resulting faster [28,84–86] whereas the second is intrinsically serial.

3.1.1. Bi-LSTM + CRF based architecture

The first architecture used is a Bi-LSTM+CRF, whose network topology is shown in Fig. 2. It is possible to distinguish three main layers, input, middle and output, which are described in detail in the following paragraphs.

Embedding layer. Different types of embeddings have been selected and mixed, based on BPEmb subword embeddings [44] and Flair contextual string embeddings [54], for which a detailed analysis is provided below.

MultiBPEmb It[9] is the multilingual version of BPEmb.[10] [44] The basis of these embeddings is one large multilingual segmentation model. Consequently, corresponding embeddings with a sub-word vocabulary, i.e. pre-trained sub-word embeddings, are shared among all 275 supported languages. The training corpus is based on Wikipedia: thanks to the underlying algorithms, which are language-agnostic but not language-independent, the article texts of all Wikipedia editions can be concatenated. This way, a sub-word segmentation model and sub-word embeddings are learnt. In detail, SentencePiece,[11] [87] the open source version of Google WordPiece, is used to learn the BPE sub-words segmentation model, while GloVe [53] is used to train sub-word embeddings. In particular, the dimensionality of the sub-word embeddings is set at 300, while the vocabulary size can be 100,000, 320,000, 1,000,000. Generally, embedding a word though BPE means that the word is subdivided into sub-words, whose embeddings vectors are subsequently combined. In sequence tagging problems with word-based gold annotations, these sub-word embeddings vectors are usually condensed into one, and this procedure can be done in several ways (e.g. arbitrarily choosing one then losing some information, using a composition function such as addition, leveraging a RNN, and so on). In this case, in order to condense the sub-word embeddings into one, the first and last sub-words embedding vector have been concatenated, leaving GloVe as the embedding algorithm. The vocabulary size has been chosen equal to 1,000,000, so that words can be more easily represented through sub-words.

Flair: multi and multi-fast embeddings Recently [54] proposed their contextual string embeddings, called Flair, along with their Flair NLP framework [52]. The novelty of these embeddings is the ability to capture latent syntactic–semantic information, unseen by standard word embeddings, leveraging two important principles: firstly, they model words as sequences of characters because they are trained without any explicit notion of words and, secondly, the surrounding text contextualizes them so that the same word will have different embeddings derived from its contextual use. As in this, such embeddings are usually employed taking advantage of both forward and backward version. In regard to multilingual versions, it is possible to distinguish between the *multi* version, pre-trained on more than 300 languages using the JW300 corpus as proposed by [88], and the *multi-fast* version pre-trained on English, German, French, Italian, Dutch and Polish, mixing several corpora (Web, Wikipedia, Subtitles and News). The embeddings dimensionality is set at 1024 and 2048 for *multi-fast* and *multi* respectively, one for forward and one for backward embeddings. The interesting property of these character-level embeddings is related to their vocabulary size: it is not as computationally heavy as word-level embeddings that have millions of distinct words to consider, but it only counts a bunch of hundreds of distinct characters, so it is really easy to train. Finally, character-level models deal well with OOV and rare words and morphologically rich languages like Italian.

Stacked embeddings As many studies in the field of clinical NER have shown [21,89,90], combining different types of embeddings together or using the most advanced ones, with different techniques [54,91], can be a useful method to take advantage of their different characteristics and achieve better performance. In detail, the concatenation technique was used, getting the stacked embedding x_t of each word as:

$$\mathbf{x}(t) = \mathbf{R} * \begin{bmatrix} \mathbf{x}(t)^{MultiBPEmb} \\ \mathbf{x}(t)^{Flair} \end{bmatrix} \quad (1)$$

[9] https://nlp.h-its.org/bpemb/multi/.
[10] https://nlp.h-its.org/bpemb/.
[11] https://github.com/google/sentencepiece.

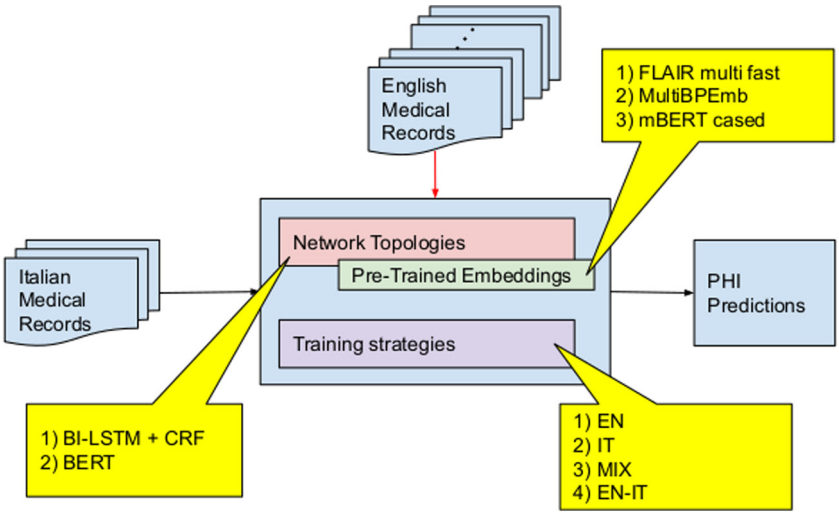

Fig. 1. Research analysis overview. The figure highlights the main research topics investigated in this paper. The red arrow indicates the inputs with extra-information for the system architecture.

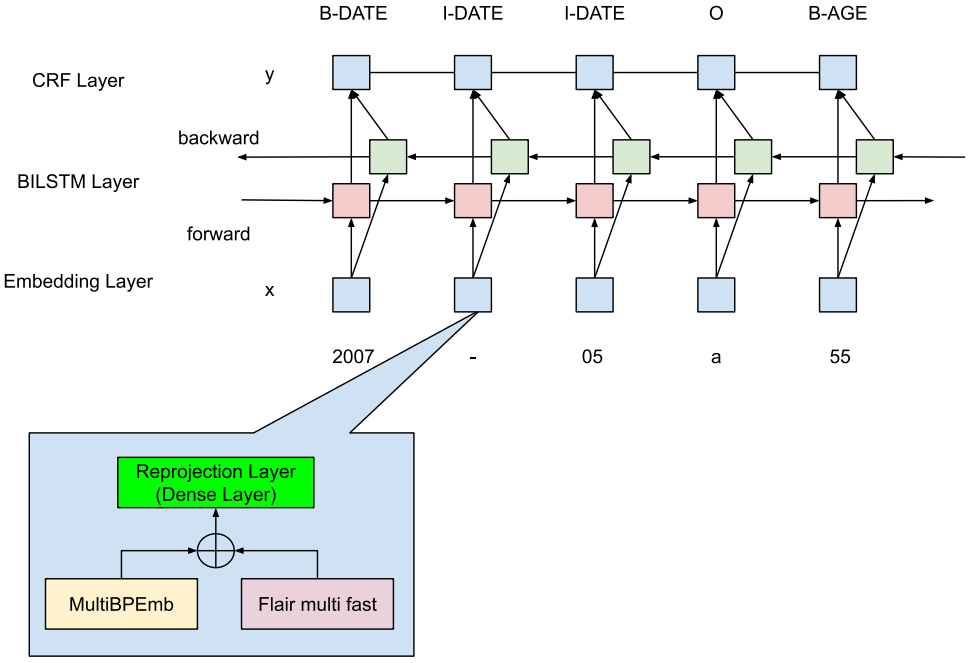

Fig. 2. A Bi-LSTM + CRF network topology. In addition, a detail regarding the embedding layer is shown.

where $\mathbf{x}(t)^{MultiBPEmb}$ and $\mathbf{x}(t)^{Flair}$ are respectively the MultiBPEmb word embedding and a type of Flair contextual string embedding and $\mathbf{R}$ is a weight matrix to remap the original stacked embedding, hereinafter *Original-Embedding*, into a new trainable embedding, called *Reprojected-Embedding*.

In detail, [54] have demonstrated how the combination of Flair embeddings with GloVe embeddings [53] is the one capable of achieving the best performance for NER. But the use of a stacked embedding in multilingual environments is able to achieve better performance when the pre-training languages have similar characteristics, otherwise the risk is to increase the confusion introduced in the network then degrade its performance. For this reason, in a bilingual scenario like the one under consideration, the optimal choice would have been to use English–Italian bilingual embeddings but, unfortunately, both Flair embeddings and GloVe embeddings are not available in such combinations. Therefore it was decided to use *Flair embeddings multi fast* (with far fewer languages than *Flair embeddings multi*) together with *MultiBPEmb*, which continue to use the GloVe algorithm but adding the ability to work at sub-word level, as already explained.

Bi-LSTM layer. It takes in input a sequence of embedding ($\mathbf{x}(1)$, $\mathbf{x}(2), \ldots, \mathbf{x}(n)$) composed by the d-dimensional vector representation of the corresponding words. It is composed by the so-called forward LSTM which produces a representation $\overrightarrow{\mathbf{h}(t)}$ of the left context of the sentence at every word t. Moreover it is also composed by another LSTM that reads the same sequence in reverse, the so-called backward LSTM obtaining a representation $\overleftarrow{\mathbf{h}(t)}$ of the right context of the sentence. The overall output is obtained by concatenating both left and right context representations: $\mathbf{h}(t) = [\overrightarrow{\mathbf{h}(t)}; \overleftarrow{\mathbf{h}(t)}]$. Therefore, the representation of a word obtained using this model is an effective representation of a word in context.

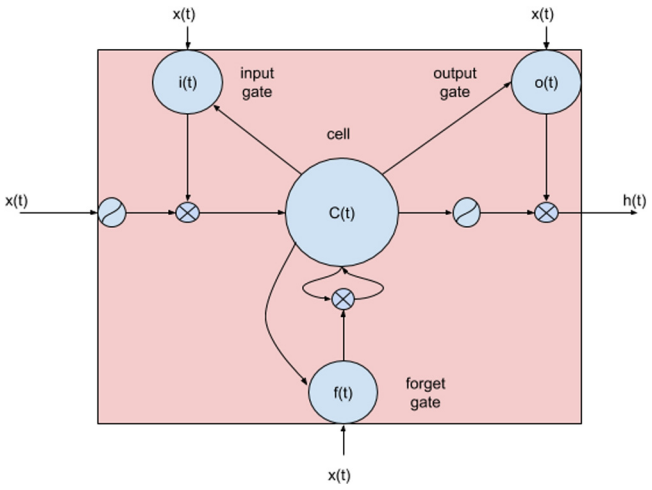

Fig. 3. Representation of a Long Short-Term Memory Cell as described in [12].

In the Bi-LSTM architecture, the hidden layer is a Long-Short Memory Cell as depicted in Fig. 3.

The implementation is managed through the following equations:

$$i(t) = \sigma(\mathbf{W_{xi}}x(t) + \mathbf{W_{hi}}h(t-1) + \mathbf{W_{ci}}c(t-1) + b_i) \quad (2)$$

$$f(t) = \sigma(\mathbf{W_{xf}}x(t) + \mathbf{W_{hf}}h(t-1) + \mathbf{W_{cf}}c(t-1) + b_f) \quad (3)$$

$$c(t) = f(t)c(t-1) + i(t)\tanh(\mathbf{W_{xc}}x(t) + \mathbf{W_{hc}}h(t-1) + b_c) \quad (4)$$

$$o(t) = \sigma(\mathbf{W_{xo}}x(t) + \mathbf{W_{ho}}h(t-1) + \mathbf{W_{co}}c(t-1) + b_o) \quad (5)$$

$$h(t) = o(t)\tanh(c(t)) \quad (6)$$

where σ is the logistic sigmoid function, and $i(\cdot), f(\cdot), o(\cdot)$ and $c(\cdot)$ are the input gate, forget gate, output gate and cell vectors. The $\mathbf{W}_{??}$ matrices represents the weight matrices to be calculated during the training process. For example, the notation $\mathbf{W_{xo}}$ represents the weight matrix of the input–output gate.

CRF Layer. It was demonstrated that the usage of CRF [92] network at the top of the Bi-LSTM prediction can improve the overall performances of sequence tagger classifier. A CRF is a variation of Markov Random Field where all the clique potentials $\phi(c)$, $1 \leq c \leq C$ are conditioned on input features. In the case of Bi-LSTM + CRF network topology, the features are the hidden layer at the top of Bi-LSTM, given that these features could be considered as a score matrix P for a given sequence, the CRF layer learns only the transition probability of the output labels.

Formally, considering a general definition of CRF, let $h = \{h(1), \ldots, x(n)\}$ and $y = \{y(1), \ldots, y(n)\}$ represent observed input tokens and corresponding output labels respectively. The distribution of a CRF linear-chain $p(y|h)$ is given by:

$$p(h|y) = \frac{1}{Z(h)} \prod_c \phi(y, h) \quad (7)$$

where

$$Z(h) = \sum_y \prod_c \phi_c(y_c, h) \quad (8)$$

3.1.2. BERT

The Bidirectional Encoder Representations from Transformers [25], is a general purpose language model trained on a large text corpus (like Wikipedia), which can be used for various downstream NLP tasks, such as NER, Relation Extraction, and Question Answering, without heavy task-specific engineering.

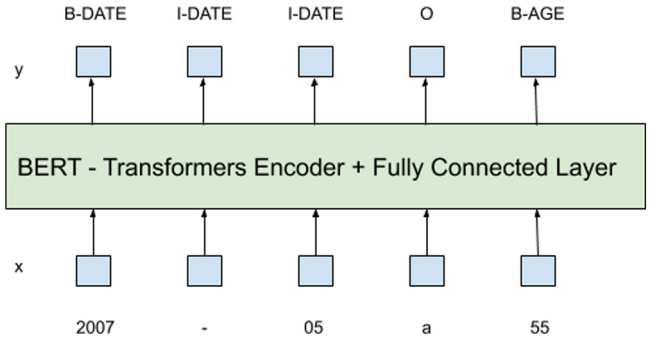

Fig. 4. A BERT network topology for Entity Recognition Task.

In detail, BERT architecture is based on 12 encoder layers, called Transformers Blocks, 12 attention heads (or Self-Attention, see [28]), and feed forward networks with a hidden size of 768. A simple network topology is shown in Fig. 4.

BERT accepts embedding and encoder input/output vectors that have a dimension of 512, called Maximum Sequence Length. Some special tokens are employed: the first is *[SEP]*, used for segments separation. The second one corresponds to the first input token supplied, the *[CLS]* token (*CLS* stands for *Classification*), which produces an output vector, of *hidden size* dimension, that can be used as the input for an arbitrarily chosen classifier.

In particular, for NER tasks, BERT is fine-tuned following a general tagging task approach without a CRF layer as output layer. As input to the token-level classifier, working over the NER label set, the representation of the first sub-token is used.

Formally, the final hidden representation h_i of each token i is passed into softmax function. The probability P is calculated as follows:

$$P(t|h_i) = softmax(W_o H_i + b_o) \quad (9)$$

where $t \in T$, W_o and b_o are weight parameters. Furthermore, during the training, categorical cross-entropy as loss function is used.

Transformer. At the base of BERT is the Transformer [28]. Say $\mathbf{x}$ and $\mathbf{y}$ a sequence of subwords from a couple of sentences. The token [CLS] is placed before $\mathbf{x}$ and after both $\mathbf{x}$ and $\mathbf{y}$ the token [SEP]. Called E the embedding function and called LN the normalization layer [28], it is possible to get the embedding in this way:

$$\hat{h}_i^0 = E(x_i) + E(i) + E(1_{\mathbf{x}}) \quad (10)$$

$$\hat{h}_{j+|x|}^0 = E(y_j) + E(j + |x|) + E(1_{\mathbf{y}}) \quad (11)$$

$$\hat{h}_\cdot^0 = Dropout(LN(\hat{h}_\cdot^0)) \quad (12)$$

Hence the embeddings follow M transformer blocks. Defined the element-wise Gaussian Error Linear Units (GELU) activation function [93] and called MHSA the Multi-Heads Self-Attention function and FF the Feed Forward layer, in each of these blocks it applies:

$$\hat{h}_\cdot^{i+1} = Skip(FF, Skip(MHSA, h_\cdot^i)) \quad (13)$$

$$Skip(f, h) = LN(h + Dropout(f(h))) \quad (14)$$

$$FF(h) = GELU(h\mathbf{W}_1^\top + \mathbf{b}_1)\mathbf{W}_2^\top + \mathbf{b}_2 \quad (15)$$

where $h^i \in \mathbb{R}^{(|\mathbf{x}|+|\mathbf{y}|)\times d_h}$, $\mathbf{W}_1 \in \mathbb{R}^{4d_h \times d_h}$, $\mathbf{b}_1 \in \mathbb{R}^{4d_h}$, $\mathbf{W}_2 \in \mathbb{R}^{4d_h \times d_h}$, $\mathbf{b}_2 \in \mathbb{R}^{4d_h}$ and one new position $\hat{h}_i$ is calculated as follows:

$$[\ldots, \hat{h}_i, \ldots] = MHSA([h_1, \ldots, h_{|x|+|y|}]) = \mathbf{W}_o Concat(h_i^1, \ldots, h_i^N) + \mathbf{b}_o \tag{16}$$

While in each attention, also called attention head, it applies:

$$h_i^j = \sum_{k=1}^{|\mathbf{x}|+|\mathbf{y}|} Dropout(\alpha_k^{(i,j)}) \mathbf{W}_V^j h_k \tag{17}$$

$$a_k^{(i,j)} = \frac{exp \frac{(\mathbf{W}_Q^j h_i)^\top \mathbf{W}_K^j h_k}{\sqrt{d_h/N}}}{\sum_{k'=1}^{|\mathbf{x}|+|\mathbf{y}|} exp \frac{(\mathbf{W}_Q^j h_i)^\top \mathbf{W}_K^j h_{k'}}{\sqrt{d_h/N}}} \tag{18}$$

where N is the number of attention heads, $h_i^j \in \mathbb{R}^{(d_h/N)}$, $\mathbf{W}_o \in \mathbb{R}^{d_h \times d_h}$, $\mathbf{b}_o \in \mathbb{R}^{d_h}$ and $\mathbf{W}_Q^j$, $\mathbf{W}_K^j$, $\mathbf{W}_V^j \in \mathbb{R}^{d_h/N \times d_h}$.

To date BERT is released in two sizes $BERT_{base}$ and $BERT_{large}$. $BERT_{base}$ is made of 12 layers (transformer blocks), 768 hidden size, 12 attention heads, and 110 million parameters, whereas $BERT_{large}$ is composed of 24 layers, 1024 hidden size, 16 attention heads and, 340 million parameters. The multilingual version of BERT is released only in the base size and is case sensitive. However, additionally, experiments by the scientific community have widely demonstrated that Cased versions of BERT and variants are superior to Uncased versions in the NER task, where the relevant entities often have capitalized initials [94]. For this reason, the cased version of mBERT has used in this work.

3.2. Training strategies: data sets

In order to analyze the crosslingual capabilities of the multilingual NER systems under examination, two different data sets were used: the English i2b2 2014 de-identification corpus and the Italian SIRM COVID-19 de-identification corpus, the latter created ad hoc for the investigation of the performance of low language resource systems with a specific case in Italian.

Four methods were tested as shown in Fig. 5:

1. EN. It provides a training set exclusively in language with high resources.
2. IT. It provides a training set exclusively in language with low resources.
3. MIX. It provides a mixed training set, both in high resource language and low resource language.
4. EN–IT. It provides two separate training sets for two distinct training phases: the first with the high resource language data set, the second with the low resource language data set.

The first data set, the i2b2 2014 de-identification corpus, was released by [8], members of the i2b2 National Center for Biomedical Computing, on the occasion of the NLP Shared Tasks Challenges. In full compliance with the HIPAA criteria and with some additional specifications, guidelines were issued for the annotation [4]. In detail, these guidelines expand the categories of 18 identifiers provided by HIPAA and group them into 7 main categories containing different subcategories. After manual tagging, data were surrogated before release in xml format including both text and annotations within appropriate tags. In particular, 1304 medical records were provided, from 2 to 5 for each of the 296 patients to whom they belonged. Of these medical records, respectively 521, 269 and 514 were assigned to the training, validation and testing data sets. In the specific case of this study, only the training data set was used.

The second data set, hereafter called SIRM COVID-19 de-identification corpus, was created specifically for this study. In particular, starting from the 115 medical records in pdf format provided by SIRM[12] without any annotation. Hence, in addition to pre-processing the data, they were annotated according to criteria similar to the first data set, so as to maintain uniformity between the recognition categories. Where the appropriate subcategory was not available, it was decided to opt for the closest one semantically: for example, the Italian regions were annotated as LOCATION: OTHER, since they belonged neither to the subcategory LOCATION: COUNTRY nor to the subcategory LOCATION: STATE or the street names identifying the hospitals were aggregated with LOCATION: HOSPITAL entities. In order to proceed with the planned experimental tests, 65 of 115 medical records and 50 of 115 medical records were used for training and testing purposes respectively.

Table 1
PHI distributions in the i2b2/UTHealth 2014 training data set and in the SIRM COVID-19 de-identification corpus.

C:Subcategory	TR_{i2b2}	TR_{SIRM}	TS_{SIRM}
AGE	810	63	55
C:EMAIL	3	0	0
C:FAX	5	0	0
C:PHONE	229	3	7
C:URL	2	66	76
DATE	5254	64	90
I:BIO ID	1	0	0
I:DEVICE	7	0	0
I:HEALTH PLAN	1	0	0
I:ID NUMBER	171	137	129
I:MEDICALRECORD	398	0	0
L:CITY	259	38	63
L:COUNTRY	53	1	5
L:HOSPITAL	928	134	131
L:ORGANIZATION	85	4	8
L:OTHER	4	3	6
L:STATE	221	0	0
L:STREET	144	0	0
L:ZIP CODE	139	0	0
N:DOCTOR	1932	302	425
N:PATIENT	879	3	0
N:USERNAME	219	0	0
PROFESSION	149	38	27
Category	**TR_{i2b2}**	**TR_{SIRM}**	**TS_{SIRM}**
AGE	810	63	55
CONTACT	239	69	83
DATE	5254	64	90
ID	578	137	129
LOCATION	1833	180	213
NAME	3030	305	425
PROFESSION	149	38	27
Total #	11,893	856	1022

Table 1 presents an exhaustive list of entity distributions in the de-identification corpora used. In particular, with reference to the first column *C:Subcategory*, *C:* stands for the category to which the entities belong if divided into subcategories, and in detail it can be C, I, L, N which stand for CONTACT, ID, LOCATION and NAME respectively. Instead the TR_{i2b2}, TR_{SIRM} and TS_{SIRM} columns indicate the i2b2 training data set and the SIRM COVID-19 training and testing data sets respectively. Finally, the i2b2 guidelines provided the subcategories C:IPADDRESS, I:ACCOUNT, I:LICENSE, I:SSN and I:VEHICLE, but the same i2b2 data set has none, so it was preferred not to include them in Table 1.

Finally, to represent the distribution of the entities within the data sets used for training and testing, a clustered column chart has been constructed, shown in Fig. 6.

The i2b2 training data set was converted from the xml format to brat standoff format through the use of the NeuroNER

12 https://www.sirm.org/category/senza-categoria/covid-19/.

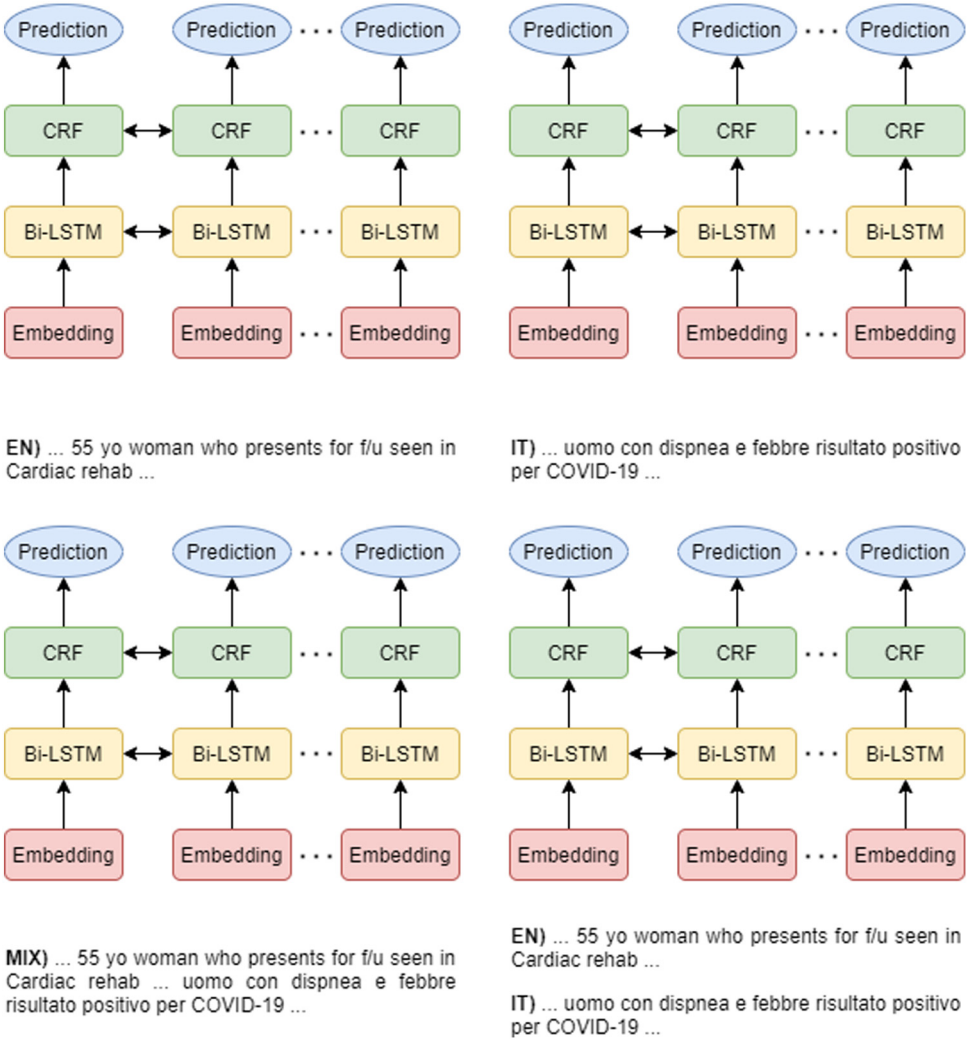

Fig. 5. Transfer learning strategies. From top left to bottom right it is possible to view the EN, IT, MIX and EN–IT strategies, where the input is modified accordingly.

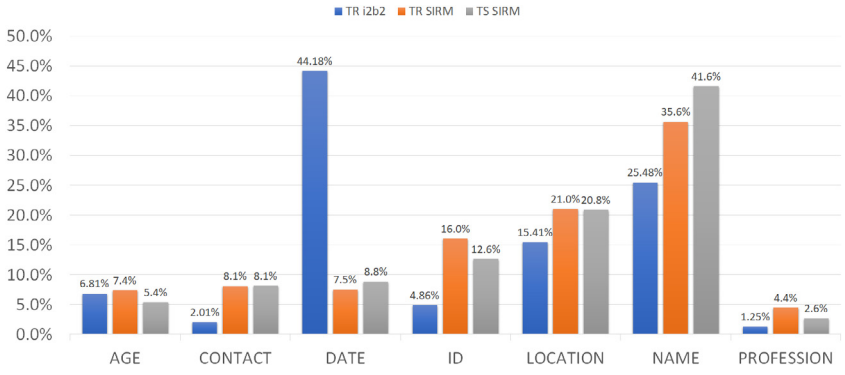

Fig. 6. Clustered column chart. Distribution of the entities in the data sets.

tool [22]. While the annotations for the SIRM COVID-19 data set were created directly in brat standoff format, after transforming the source pdf files into text using the python *pandas* library. If the annotated text was contiguous with unannotated text, they were separated to improve tokenization[13]: the subsequent misalignment was adjusted recalculating initial and final offset of the entity within the text. The data sets were then converted from brat standoff format to CONLL format, depending on the input needs of the framework used and according to the IOB tagging format [95] used where O, B-tag and I-tag represent an untagged token, the begin of a tagged token and all the rest of a tagged entity respectively.

[13] Spacy was used as tokenizer, see https://spacy.io/ for details.

Table 2
Bi-LSTM+CRF hyper-parameters.

Hyperparameter	Value
Annealing factor	0.5
Batch size	16
Dropout	0.5 (variational) 0.05 (word)
Epochs	up to 500
Gradient clipping	5
Hidden size	256
Learning rate	0.1 - 0.0001
Patience	3
RNN Layers	1

Table 3
$BERT_{base}$ and mBERT hyper-parameters.

Hyperparameter	Value
Attention heads	12
Batch size	32
Epochs	5
Hidden size	768
Languages	104
Hidden layers	12
Maximum Sequence Length	512
Parameters	110 M

4. Experiments and results

In this section, the experimental setup is shown in Section 4.1, the evaluation metrics are described in Section 4.2 and, finally, the results obtained are presented in Section 4.3.

4.1. Implementation details and experimental setup

To implement the systems described in this article, two different frameworks have been used that offer different possibilities for NLP tasks like classification, NER, Part-of-Speech tagging, sense disambiguation and so on. The first one is Flair,[14] [52] written in Python: this framework has been used to implement the neural network based system Bi-LSTM+CRF, leaving default values not of interest and setting as shown in Table 2 the values relevant to the experimentation. The second is Hugging Face Transformers,[15] also written in Python: this framework has been used to implement the system based on Transformers, so BERT. Similarly to what was done previously, only the values relevant to the experimentation have been modified and reported in Table 3. The hyper-parameters modified and reported in Tables 2 and 3 are described below.

Regarding Flair, the stochastic gradient descent (SGD) was used to update neural network parameters. Every 3 epochs without improvement the learning rate is reduced according to *Patience* hyper-parameter, by multiplying the annealing factor, so it goes from 0.1 to 0.0001, the latter being a system condition of early stopping. For this reason, the 500 limit of training epochs is never reached but the number of epochs used is different for each trained model. Other hyper-parameters are: gradient clipping 5.0, Bi-LSTM hidden size 256, variational dropout 0.5, word dropout 0.05 and batch size 16.

Regarding HuggingFace Transformers, $BERT_{base}$ and mBERT implementations have both 110*M* of parameters. Batch size and Maximum Sequence Length were set to 32 and 512 respectively, while the model was fine-tuned for 5 epochs. Attention heads, hidden size and hidden layers were 12, 768 and 12 respectively.

An IBM POWER9 cluster with NVIDIA V100 GPUs was used to run the experiments. In detail, the tested models were based on:

1. Bi-LSTM+CRF with stacked embedding consisting of *MultiBPEmb* and *Flair embedding multi-fast* (both forward and backward);
2. mBERT Cased.

The models were trained using the strategies introduced in Section 3.2. In particular, the EN and IT strategies perform a training on i2b2 2014 and SIRM COVID-19 data sets respectively, while the MIX strategy provides a single concatenated i2b2 2014/SIRM COVID-19 data set for training, finally the EN–IT strategy performs a first training on i2b2 2014 data set and a second training on SIRM COVID-19 data set. All models were tested on SIRM COVID-19 testing data set (50 of 115 clinical records).

Finally all models were trained and tested repeating the procedure five times for each configuration and reporting the arithmetic mean of the results, rounded to the fourth decimal place.

4.2. Evaluation metrics

From precision P and recall R it is possible to define their harmonic mean, called measure F_1 to evaluate the performance of the models and compare them. Said TP the number of true positives, FP the number of false positives and FN the number of false negatives, it is possible to define the metrics:

$$F_1 = \frac{2 * P * R}{P + R} \tag{19}$$

$$P = \frac{TP}{TP + FP} = \frac{\#\ of\ correct\ entities}{\#\ of\ predicted\ entities} \tag{20}$$

$$R = \frac{TP}{TP + FN} = \frac{\#\ of\ correct\ entities}{\#\ of\ expected\ entities} \tag{21}$$

The most common calculation methods for the F_1 value, which change its value in the case of multi-class problems, are as follows:

- Macro-Averaging. The precision and recall values are calculated for each class. Later precision and recall are calculated as the arithmetic average of the precision and recall values. Therefore F_1 is calculated by (19).[16] It should be reported its standard deviation also.
- Weighted Macro-Averaging. The precision and recall values are calculated for each class. Later precision and recall are calculated as the weighted average (concerning the number of expected entities for each class) of the precision and recall values. Therefore F_1 is calculated by (19)[16].
- Micro-Averaging. The number of correct, expected and expected entities of each class is summed up. Accuracy and recall are calculated with these total sums. In this case, for binary classification problems, $F_1 = Accuracy$.

The most common method is Micro-Averaging which has been used. The results were produced using five criteria: *binary*, *i2b2 category* and *i2b2 sub-category* on one side and *entity* and *token* on the other. In the case of the *binary* criterion it is sufficient to discriminate between entities and non-entities (or tokens and non-token), then for the *i2b2 category* and the *i2b2 subcategory* it is necessary to recognize the categories and subcategories to which the entities or tokens respectively belong. In addition, the *entity* criterion controls if a predicted entity matches precisely the correspondent in the so-called gold standard (i.e. ground truth), while the *token* criterion controls only if there is a token correspondence, which is considered correct even if it only partially fits the entity. So, in *entity-subcategory* cases the lowest scores are obtained, while in *token-binary* cases the highest scores are obtained.

[14] https://alanakbik.github.io/flair.html.
[15] https://github.com/huggingface/transformers.
[16] The final F_1 value should be calculated by (19) and not by arithmetic or weighted average of the F_1 values of the classes.

4.3. Results

The Micro-Averaged F_1 results are shown in Table 4. In particular, column *Model* indicates the trained model, while column *Strategy* indicates the training strategy adopted. On the other columns, as explained in Section 4.2, there are six evaluation criteria: S, C and B represent respectively *i2b2 subcategory*, *i2b2 category*, *binary*, while the subscripts E and T stand for *entity* and *token*.

In particular, two pre-trained Italian language models were used as baselines: the *Bi-LSTM+CRF: BPEmb (IT) + Flair (IT)* and the $BERT_{base}$ *(IT) Cased* models. In these cases the only possible training strategy involves the exclusive use of the Italian training set. Observing the results it is possible to understand how it is feasible to obtain better performance by using strategies based on transfer learning approaches: in this way it is easy to increase the training set by using data available in languages with high resources such as English.

Furthermore, the results further confirm what [25] have already expressed in the literature: although the results at token level suggest the use of BERT-based architectures for the NER task, this assumption is actually misleading. It is important to remember that the NER, hence de-identification as the basis of the anonymization process, should be evaluated at the level of multiclass entities, i.e. category and subcategory. In all other scenarios, in fact, entities could be replaced by the wrong surrogates, which would leave ample room for re-identification [3]. As a result, it is possible to consider the model Bi-LSTM+CRF with *MultiBPEmb + Flair multi fast* stacked embedding trained with strategy *EN–IT* as the most suitable for the clinical de-identification in a low-resources scenario such as that of the Italian language. Hereinafter, the *Bi-LSTM+CRF: BPEmb (IT) + Flair (IT)* will be referred to as monolingual system, while the *Bi-LSTM+CRF: MultiBPEmb + Flair multi fast* model trained with MIX or EN–IT strategies as crosslingual systems.

4.3.1. Embeddings ablation analysis

For the sake of completeness, it was analyzed the importance of each embedding type within the EN–IT crosslingual system. Results are reported in Table 5.

As can be easily seen from the results, neither Flair alone nor MultiBPEmb alone can achieve results comparable to their combination: exploiting a contextual model that works at character level proves to be a less performing choice compared to the use of a subword model in the case of a low-resources language. But the considerable detachment that is obtained by combining the two different embedding suggests that, in a clinical de-identification task such the one under analysis, the use of a subword model that can also exploit contextuality is particularly effective.

4.3.2. Embeddings space analysis

The Fig. 7 reports an embedding scatter plot obtained applying a 3D Principal Component Analysis (PCA) on the original embedding space representing the Original-Embedding (MultiBPEmb + Flair multi fast) on the first row and the Reprojected-Embedding after the double training strategy (EN–IT) on the second row. On the other hand, the first column presents the tokens related to English sentences whereas the second column the ones related to Italian sentences.

The two data sets used for the plots are:

- the Italian SIRM COVID-19 test set composed by 1185 sentences;
- the first 1185 sentences of the English i2b2 2014 training set.

This choice was made in order to have about the same data points for both English and Italian scenarios.

For the sake of clarity, only the two most represented categories were considered, *C:Name* and *C:Location*, respectively the red and blue points for the English set and the orange and black points for the Italian set. All other categories are labeled with *C:Other*, using the colors yellow (EN) and green (IT).

The scatter plots highlight two major insights:

1. *Column view*: the training process has generated a redistribution of clusters on the reprojected embedding space depending on the NER task adapting their own position on the analyzed categories;
2. *Row view*: the better alignment of embeddings clusters considering the relative position of the clusters related to each category: in fact, in the reprojected embedding space, the clusters of the same category remain in the same area for both languages.

5. Discussion

The results obtained allow to identify what may be the best approaches to manage clinical de-identification using NER systems for Italian language. First of all it is possible to notice that the use of the crosslingual system trained in English and tested in Italian does not obtain exciting results, on the contrary it obtains worse results than a monolingual system with training and testing in Italian.

In detail, this study allows to highlight that, even if used in a scenario of limited resources such as that of the Italian language, crosslingual systems properly used can obtain better results than monolingual systems provided some caution during training, so as to take full advantage of the beneficial effects due to the transfer learning. In fact, crosslingual systems that are trained with a mixed English–Italian data set or with a double training first in English then in Italian, can obtain better results than monolingual systems. Moreover, this study shows that it is slightly preferable to adopt a strategy with double training, rather than single training with a mixed data set: this finding leads to think that in the first case the "noise" introduced within the network is more limited, favoring a better settlement of the weights of the neural network.

Along with these aspects, it is important to add another crucial consideration: the world of research today is strongly interconnected, which is why it is increasingly common to come across bibliometric references, often in English or transliterated Chinese, within medical records written in any other language. Hence there is a phenomenon sometimes similar to code-switching, although references do not constitute expressions in different languages within the same expression, but rather sentences disconnected from the rest of the discourse and reported as notes at the bottom of the page. For these reasons, crosslingual systems can, in addition, succeed in obtaining superior performance, at least in terms of recognition of entities in languages other than the target language. An example can be the entity *State Administration of Traditional Chinese Medicine*, written in English within a predominantly Italian text and correctly recognized only by the crosslingual systems as LOCATION: ORGANIZATION.

5.1. Strengths and weaknesses: monolingual vs crosslingual systems

In order to clarify the advantages and disadvantages of the best crosslingual system, i.e. trained with EN–IT strategy, compared to the monolingual system, a comparative analysis of the entities surveyed is proposed below. For simplicity, the monolingual system will be indicated by the abbreviation IT while the crosslingual system with EN–IT training strategy will be indicated by the abbreviation EN–IT.

Table 4
Micro-Averaged F_1 results.

Model	Strategy	S_E	C_E	B_E	S_T	C_T	B_T
Bi-LSTM+CRF: BPEmb (IT) + Flair (IT)	IT	0.8110	0.8278	0.8317	0.8856	0.9115	0.9190
Bi-LSTM+CRF: MultiBPEmb + Flair multi fast	EN	0.2662	0.2948	0.3134	0.4103	0.4914	0.5797
	IT	0.7910	0.8118	0.8159	0.8826	0.9060	0.9183
	MIX	0.8371	**0.8602**	0.8618	0.8970	0.9304	0.9417
	EN–IT	**0.8391**	0.8595	0.8619	**0.9033**	0.9321	**0.9449**
$BERT_{base}$ (IT) Cased	IT	0.7553	0.7880	0.8561	0.7969	0.8979	0.9260
mBERT Cased	EN	0.4585	0.5029	0.6878	0.5498	0.6097	0.6878
	IT	0.7768	0.8207	**0.9449**	0.8923	**0.9353**	**0.9449**
	MIX	0.7696	0.8105	0.9379	0.8833	0.9245	0.9379
	EN–IT	0.7228	0.7576	0.8969	0.8241	0.8678	0.8969

Table 5
Micro-Averaged F_1 results for embeddings ablation analysis of the EN–IT crosslingual system.

Embedding	S_E	C_E	B_E	S_T	C_T	B_T
MultiBPEmb	0.7614	0.7743	0.7914	0.8201	0.8569	0.8835
Flair multi fast	0.7621	0.7851	0.7972	0.8529	0.8801	0.8963
MultiBPEmb + Flair multi fast	**0.8391**	**0.8595**	**0.8619**	**0.9033**	**0.9321**	**0.9448**

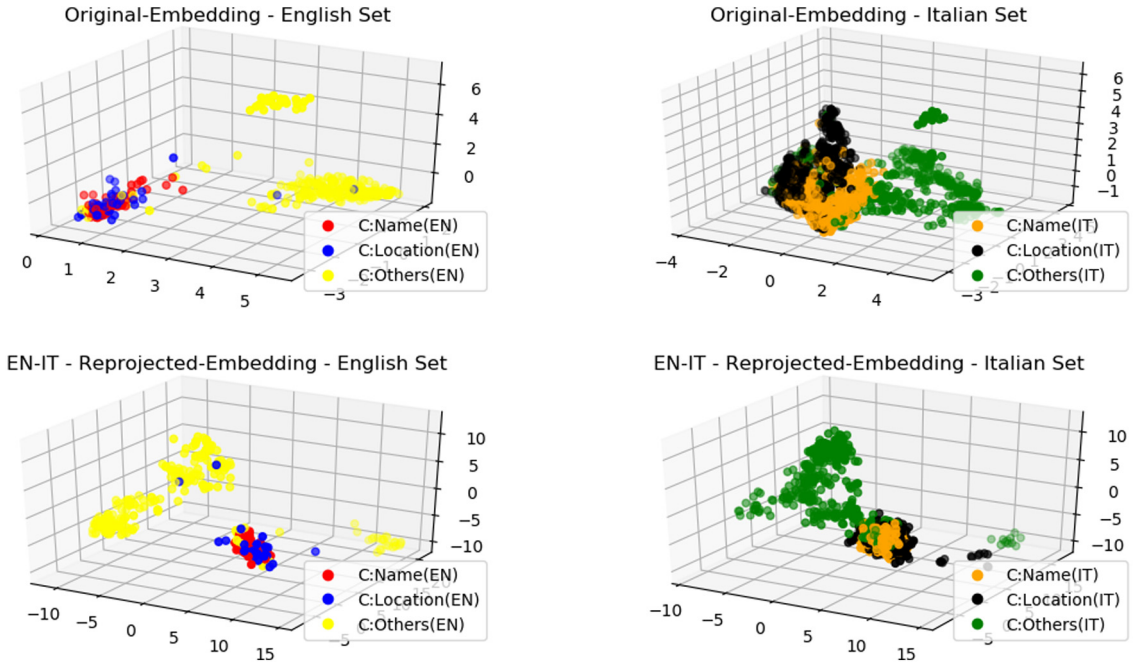

Fig. 7. Scatter plots of three dimensional principal component analysis of the embedding points.

5.1.1. Entities analysis

Table 6 shows the entities that are correctly recognized only by one system, IT or EN–IT. The output of the tokenization process is emphasized by the alternation of black and red colors.

First of all, the only type of case study in which the IT system has an advantage over the EN–IT system: it refers to all those situations in which there are multi-token entities in the target language that are complex and specific. An example is given by the entity of type LOCATION: HOSPITAL *Unità di terapia intensiva* "Intensive Care Unit": the IT system correctly recognizes the entity, instead the EN–IT system succeeds in a random way because of the complexity and peculiarity of the entity, which neither presents the same number of tokens as the English correspondent (4 vs 3) nor has the same roots for all the words in the other training language (*terapia* vs "care"). On the other hand, the EN–IT crosslingual system has a number of advantages in different scenarios, which can be grouped in three cases.

The first case concerns all those entities in languages other than the target language, but present because they may be quotes. A frequent example is the one given by foreign names, English or Chinese, such as the entity *Liu, Bin* of type NAME: DOCTOR. Even if the Beginning or Inside of the entity is not correct, maybe because of an unusual pattern compared to the Italian, i.e. *Surname, Name*, it is possible to get better results at token level. Here the crosslingual is clearly superior to the monolingual approach. The second case, on the other hand, concerns those entities which generally belong to the LOCATION category. What can be identified is a higher accuracy, especially finer-grained and therefore at subcategory level, of the EN–IT system than the IT system. In detail, some entities of type LOCATION: CITY or LOCATION: OTHER as *Milano*, *Marcianise*, *Vibo Valentia*, *Wuhan* and *Veneto* are generally correctly identified by the EN–IT system, instead with wrong subcategories or unseen by the IT system. The motivation is probably to be found in the ability of the crosslingual system to rely to a greater extent on contextual patterns derived also from the English language that suggest the presence of an entity of type LOCATION: CITY. Instead, the IT system tends to identify them as LOCATION: HOSPITAL: this error is induced by the fact that in the

Table 6
Examples of recognized entities. The alternation of black and red words is used to emphasize the output of the tokenization process. Best viewed in color.

i2b2 Category: Subcategory	Entity	Recognized by
AGE	47aa	EN–IT
CONTACT: PHONE	118	EN–IT
DATE	12.02.2020	EN–IT
LOCATION: CITY	Milano	EN–IT
	Marcianise	EN–IT
	Vibo Valentia	EN–IT
	Wuhan	EN–IT
LOCATION: HOSPITAL	Unità di terapia intensiva (*intensive care unit*)	IT
LOCATION: OTHER	Veneto	EN–IT
NAME: DOCTOR	Liu, Bin	EN–IT

Table 7
Challenging entities. The alternation of black and red words is used to emphasize the output of the tokenization process. Best viewed in color.

i2b2 Category: Subcategory	Entity	Motivation
CONTACT: URL	http://yzs.satcm.gov.cn/zhengcewenjian/2020-02-19/13221.html	AMB
DATE	domenica (*sunday*)	S
ID: ID NUMBER	10.1186/s40779-020-00240-0	AMB
LOCATION: CITY	Fabrizia	S
	Melito	S
	VV	AB
	CE	AB
LOCATION: HOSPITAL	reparto dedicato ai pazienti COVID-19 (*COVID-19 patients department*)	AMB, D
	HUB di riferimento Covid (*Covid reference HUB*)	AMB, D
LOCATION: OTHER	vibonese	AMB, S
	lodigiano	AMB, S
PROFESSION	clinici (*clinician*)	S
	dipendente di industria chimica (*chemical industry employee*)	S
	medico di Pronto Soccorso (*Emergency Room medical doctor*)	S

Italian language the names of cities or places are often used also to give the name to the hospital that oversees the city or place.

Finally, the third case considers those entities of type AGE, CONTACT: PHONE, DATE which, although not present in large numbers, are expressed through recurrent patterns also in other languages such as English: some examples can be *47aa* "47yo", *118*, *12.02.2020*.

5.2. Challenging entities

In this section the focus is on the entities that are difficult to identify for both systems, with the aim of providing some explanation. As already mentioned by [6], the main sources of error are generally due to (1) abbreviations, whose brevity and variety contribute to confuse the learning system, (2) ambiguities, due to polysemic tokens used in unclear contexts, (3) debatable annotations, i.e. annotation errors, shortcomings or variations with respect to the guidelines and (4) both scarcity and sparsity of certain types of entities within the data sets. These error sources have been indicated by the abbreviations AB, AMB, D and S, respectively, and used in the *Motivation* column of Table 7. In detail, this table shows the entities that are most difficult to identify and the alternation of the colors black and red indicates how tokenization works.

Some examples are LOCATION: CITY entities such as *Fabrizia* or *Melito*, scarcely present, or LOCATION: CITY abbreviations widely present as *VV* and *CE* to indicate the cities of Vibo Valentia and Caserta respectively, or LOCATION: OTHER entities such as *vibonese* and *lodigiano*, which represent unusual ways of identifying the provinces Vibo Valentia and Lodi.

Moreover, the systems under analysis are not able to successfully identify those complexly structured entities such as *reparto dedicato ai pazienti COVID-19* "COVID-19 patient department" or *HUB di riferimento Covid* "Covid Reference HUB", labeled as LOCATION: HOSPITAL but not predicted in any way, probably because of the too ambiguous way of identifying specific places without even using capital letters.

While ID: ID NUMBER or CONTACT: URL entities such as *10.1186/s40779-020-00240-0* and http://yzs.satcm.gov.cn/zhengcewenjian/2020-02-19/13221.html that the tokenizer tends to break into several sub-tokens are never correctly recognized by either system and, in addition, also ambiguities contribute to lower the score: for example, within the second entity it might be easy to confuse the *2020-02-19* part with an entity of type DATE.

Furthermore, those entities of type PROFESSION, such as *clinici* "clinicians", *dipendente di industria chimica* "chemical industry employee" or *medico di Pronto Soccorso* "Emergency Room doctor", are not detected by the systems because of the scarcity, as the number of entities in training is too small.

Likewise, entities that do not recur in the training set but that also present a completely different morphology such as *domenica* "sunday" (type DATE) are not detected at all.

6. Conclusion

In this study two cutting-edge NER architectures, Bi-LSTM+CRF and BERT, suitable for de-identification, were analyzed in order to understand their behavior on COVID-19 medical records with respect to a low-resource language scenario like the Italian one. For this purpose, an additional data set was built in Italian from publicly available raw data, called SIRM COVID-19 data set. Additionally, four strategies were tested to pinpoint the best to apply in this particular context. Performed tests showed that the best strategy to adopt was a double training, before in English then in Italian, exploiting a Bi-LSTM+CRF architecture in combination with MultiBPEmb and Flair Multilingual Fast embeddings. The results obtained leave further room for improvement, although they have allowed to highlight how, in this situation, it is desirable to proceed with clinical de-identification given the low-resources language problem. An interesting future development

could be the comparison of different architectures even among those not available for multilingual purposes, to understand if at the moment the results obtained are the best possible. The real limitation of this research area remains the size of the data sets available for clinical de-identification: it would be appropriate to increase the availability of de-identification data sets of the same size as the English i2b2 2014, so as to allow a fair comparison with monolingual systems and provide strong baselines of reference before attempting necessary approaches to low resources case studies.

CRediT authorship contribution statement

Rosario Catelli: Writing - original draft, Writing - review & editing, Methodology, Data Curation, Software, Validation. **Francesco Gargiulo:** Methodology, Writing - review & editing, Formal analysis, Software, Validation. **Valentina Casola:** Supervision, Writing - review & editing. **Giuseppe De Pietro:** Supervision, Writing - review & editing, Funding acquisition. **Hamido Fujita:** Supervision, Writing - review & editing. **Massimo Esposito:** Project administration, Writing - review & editing, Formal analysis, Conceptualization, Methodology, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Hernandez-Matamoros, H. Fujita, T. Hayashi, H. Perez-Meana, Forecasting of COVID19 per regions using ARIMA models and polynomial functions, Appl. Soft Comput. (2020) 106610.

[2] T.B. Røst, L. Slaughter, Ø. Nytrø, A.E. Muller, G. Vist, Using deep learning to support high-quality Covid-19 evidence mapping, 2020.

[3] V. Vincze, R. Farkas, De-identification in natural language processing, in: 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO, IEEE, 2014, pp. 1300–1303.

[4] A. Stubbs, Ö. Uzuner, Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus, J. Biomed. Inform. 58 (2015) S20–S29.

[5] D. Nadeau, S. Sekine, A survey of named entity recognition and classification, Lingvist. Investig. 30 (1) (2007) 3–26.

[6] F. Dernoncourt, J.Y. Lee, O. Uzuner, P. Szolovits, De-identification of patient notes with recurrent neural networks, J. Am. Med. Inform. Assoc. 24 (3) (2016) 596–606, http://dx.doi.org/10.1093/jamia/ocw156.

[7] S.M. Meystre, F.J. Friedlin, B.R. South, S. Shen, M.H. Samore, Automatic de-identification of textual documents in the electronic health record: a review of recent research, BMC Med. Res. Methodol. 10 (1) (2010) 70.

[8] A. Stubbs, C. Kotfila, Ö. Uzuner, Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1, J. Biomed. Inform. 58 (2015) S11–S19.

[9] J.P. Chiu, E. Nichols, Named entity recognition with bidirectional LSTM-CNNs, Trans. Assoc. Comput. Linguist. 4 (2016) 357–370, http://dx.doi.org/10.1162/tacl_a_00104.

[10] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 260–270, http://dx.doi.org/10.18653/v1/N16-1030.

[11] X. Ma, E. Hovy, End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1064–1074, http://dx.doi.org/10.18653/v1/P16-1101.

[12] R. Alzaidy, C. Caragea, C.L. Giles, Bi-LSTM-CRF sequence labeling for keyphrase extraction from scholarly documents, in: The World Wide Web Conference, in: WWW '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 2551–2557, http://dx.doi.org/10.1145/3308558.3313642.

[13] Z. Liu, B. Tang, X. Wang, Q. Chen, De-identification of clinical notes via recurrent neural network and conditional random field, J. Biomed. Inform. 75 (2017) S34 – S42, http://dx.doi.org/10.1016/j.jbi.2017.05.023, A Natural Language Processing Challenge for Clinical Records: Research Domains Criteria (RDoC) for Psychiatry.

[14] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.

[15] J.L. Elman, Finding structure in time, Cogn. Sci. 14 (2) (1990) 179–211.

[16] C. Goller, A. Kuchler, Learning task-dependent distributed representations by backpropagation through structure, in: Proceedings of International Conference on Neural Networks, ICNN'96, vol. 1, IEEE, 1996, pp. 347–352.

[17] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[18] Y. Wu, M. Jiang, J. Lei, H. Xu, Named entity recognition in Chinese clinical text using deep neural network, Stud. Health Technol. Inform. 216 (2015) 624.

[19] Y. Wu, J. Xu, M. Jiang, Y. Zhang, H. Xu, A study of neural word embeddings for named entity recognition in clinical text, in: AMIA Annual Symposium Proceedings, vol. 2015, American Medical Informatics Association, 2015, p. 1326.

[20] Y. Wu, M. Jiang, J. Xu, D. Zhi, H. Xu, Clinical named entity recognition using deep learning models, in: AMIA Annual Symposium Proceedings, vol. 2017, American Medical Informatics Association, 2017, p. 1812.

[21] Y. Wu, X. Yang, J. Bian, Y. Guo, H. Xu, W. Hogan, Combine factual medical knowledge and distributed word representation to improve clinical named entity recognition, in: AMIA Annual Symposium Proceedings, vol. 2018, American Medical Informatics Association, 2018, p. 1110.

[22] F. Dernoncourt, J.Y. Lee, P. Szolovits, NeuroNER: an easy-to-use program for named-entity recognition based on neural networks, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 97–102, http://dx.doi.org/10.18653/v1/D17-2017.

[23] Y.-S. Zhao, K.-L. Zhang, H.-C. Ma, K. Li, Leveraging text skeleton for de-identification of electronic medical records, BMC Med. Inform. Decis. Making 18 (1) (2018) 18.

[24] Y. Kim, P. Heider, S. Meystre, Ensemble-based methods to improve de-identification of electronic health record narratives, in: AMIA Annual Symposium Proceedings, vol. 2018, American Medical Informatics Association, 2018, p. 663.

[25] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/N19-1423.

[26] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly available clinical BERT embeddings, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019, pp. 72–78.

[27] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[29] L. Luo, Z. Yang, P. Yang, Y. Zhang, L. Wang, H. Lin, J. Wang, An attention-based bilstm-CRF approach to document-level chemical named entity recognition, Bioinformatics 34 (8) (2018) 1381–1388.

[30] A. Hu, Z. Dou, J.-Y. Nie, J.-R. Wen, Leveraging multi-token entities in document-level named entity recognition, in: AAAI, 2020, pp. 7961–7968.

[31] M. Marimon, A. Gonzalez-Agirre, A. Intxaurrondo, H. Rodriguez, J.L. Martin, M. Villegas, M. Krallinger, Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results, in: IberLEF@ SEPLN, 2019, pp. 618–638.

[32] B. Tang, D. Jiang, Q. Chen, X. Wang, J. Yan, Y. Shen, De-identification of clinical text via bi-LSTM-CRF with neural language models, in: AMIA Annual Symposium Proceedings, vol. 2019, American Medical Informatics Association, 2019, p. 857.

[33] J.M. Giorgi, G.D. Bader, Towards reliable named entity recognition in the biomedical domain, Bioinformatics 36 (1) (2020) 280–286.

[34] N. Mehrabi, T. Gowda, F. Morstatter, N. Peng, A. Galstyan, Man is to person as woman is to location: Measuring gender bias in named entity recognition, in: Proceedings of the 31st ACM Conference on Hypertext and Social Media, 2020, pp. 231–232.

[35] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2009) 1345–1359.

[36] D. Zeman, P. Resnik, Cross-language parser adaptation between related languages, in: Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, 2008, pp. 34–42.

[37] J.-K. Kim, Y.-B. Kim, R. Sarikaya, E. Fosler-Lussier, Cross-lingual transfer learning for pos tagging without cross-lingual resources, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2832–2838.

[38] J. Xie, Z. Yang, G. Neubig, N.A. Smith, J. Carbonell, Neural cross-lingual named entity recognition with minimal resources, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 369–379, http://dx.doi.org/10.18653/v1/D18-1034.

[39] W. Ahmad, Z. Zhang, X. Ma, E. Hovy, K.-W. Chang, N. Peng, On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 2440–2452, http://dx.doi.org/10.18653/v1/N19-1253.

[40] S. Ruder, I. Vulić, A. Søgaard, A survey of cross-lingual word embedding models, J. Artificial Intelligence Res. 65 (2019) 569–631.

[41] G. Lample, A. Conneau, M. Ranzato, L. Denoyer, H. Jégou, Word translation without parallel data, 2018, in: International Conference on Learning Representations.

[42] M. Artetxe, G. Labaka, E. Agirre, A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 789–798, http://dx.doi.org/10.18653/v1/P18-1073.

[43] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Trans. Assoc. Comput. Linguist. 5 (2017) 135–146.

[44] B. Heinzerling, M. Strube, BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018.

[45] R. Sennrich, B. Haddow, A. Birch, Neural machine translation of rare words with subword units, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (vol. 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1715–1725, http://dx.doi.org/10.18653/v1/P16-1162.

[46] J. Bingel, J. Bjerva, Cross-lingual complex word identification with multitask learning, in: Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, 2018, pp. 166–174.

[47] S.M. Yimam, C. Biemann, S. Malmasi, G. Paetzold, L. Specia, S. Štajner, A. Tack, M. Zampieri, A report on the complex word identification shared task 2018, in: Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 66–78, http://dx.doi.org/10.18653/v1/W18-0507.

[48] M. Zhao, H. Schütze, A multilingual bpe embedding space for universal sentiment lexicon induction, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 3506–3517.

[49] Y. Zhu, I. Vulić, A. Korhonen, A systematic study of leveraging subword information for learning word representations, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 912–932, http://dx.doi.org/10.18653/v1/N19-1097.

[50] G.G. Şahin, C. Vania, I. Kuznetsov, I. Gurevych, Linspector: Multilingual Probing Tasks for Word Representations, MIT Press, 2019.

[51] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2227–2237, http://dx.doi.org/10.18653/v1/N18-1202.

[52] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, R. Vollgraf, FLAIR: An easy-to-use framework for state-of-the-art NLP, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 54–59, http://dx.doi.org/10.18653/v1/N19-4010.

[53] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543, http://dx.doi.org/10.3115/v1/D14-1162.

[54] A. Akbik, D. Blythe, R. Vollgraf, Contextual string embeddings for sequence labeling, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 1638–1649.

[55] A. Akbik, T. Bergmann, R. Vollgraf, Pooled contextualized embeddings for named entity recognition, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 724–728, http://dx.doi.org/10.18653/v1/N19-1078.

[56] A. Johnson, P. Karanasou, J. Gaspers, D. Klakow, Cross-lingual transfer learning for Japanese named entity recognition, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 2 (Industry Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 182–189, http://dx.doi.org/10.18653/v1/N19-2023.

[57] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (vol. 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 328–339, http://dx.doi.org/10.18653/v1/P18-1031.

[58] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training, 2018.

[59] A. Conneau, G. Lample, Cross-lingual language model pretraining, in: Advances in Neural Information Processing Systems, 2019, pp. 7059–7069.

[60] T. Schuster, O. Ram, R. Barzilay, A. Globerson, Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 1599–1613, http://dx.doi.org/10.18653/v1/N19-1162.

[61] P. Mulcaire, J. Kasai, N.A. Smith, Polyglot contextual representations improve crosslingual transfer, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 3912–3918, http://dx.doi.org/10.18653/v1/N19-1392.

[62] M. Arkhipov, M. Trofimova, Y. Kuratov, A. Sorokin, Tuning multilingual transformers for language-specific named entity recognition, in: Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, 2019, pp. 89–93.

[63] R. Vaishya, M. Javaid, I.H. Khan, A. Haleem, Artificial intelligence (AI) applications for COVID-19 pandemic, Diabetes Metab. Syndr.: Clin. Res. Rev. (2020).

[64] Y. Mohamadou, A. Halidou, P.T. Kapen, A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of COVID-19, Appl. Intell. (2020) 1–13.

[65] B.S. Santos, I. Silva, M. da Câmara Ribeiro-Dantas, G. Alves, P.T. Endo, L. Lima, COVID-19: A scholarly production dataset report for research analysis, Data Brief (2020) 106178.

[66] J.S. Suri, A. Puvvula, M. Biswas, M. Majhail, L. Saba, G. Faa, I.M. Singh, R. Oberleitner, M. Turk, P.S. Chadha, et al., COVID-19 pathways for brain and heart injury in comorbidity patients: A role of medical imaging and artificial intelligence-based COVID severity classification: A review, Comput. Biol. Med. (2020) 103960.

[67] C. Coombs, Will COVID-19 be the tipping point for the intelligent automation of work? A review of the debate and implications for research, Int. J. Inf. Manage. (2020) 102182.

[68] M.H. Shakil, Z.H. Munim, M. Tasnia, S. Sarowar, COVID-19 and the environment: A critical review and research agenda, Sci. Total Environ. (2020) 141022.

[69] A. Hernandez-Matamoros, H. Fujita, T. Hayashi, H. Perez-Meana, Forecasting of covid19 per regions using ARIMA models and polynomial functions, Appl. Soft Comput. 96 (2020) 106610, http://dx.doi.org/10.1016/j.asoc.2020.106610, http://www.sciencedirect.com/science/article/pii/S1568494620305482.

[70] B.B. Hazarika, D. Gupta, Modelling and forecasting of COVID-19 spread using wavelet-coupled random vector functional link networks, Appl. Soft Comput. 96 (2020) 106626, http://dx.doi.org/10.1016/j.asoc.2020.106626, http://www.sciencedirect.com/science/article/pii/S1568494620305640.

[71] G. calo Marques, D. Agarwal, I. de la Torre Díez, Automated medical diagnosis of COVID-19 through efficientnet convolutional neural network, Appl. Soft Comput. 96 (2020) 106691, http://dx.doi.org/10.1016/j.asoc.2020.106691, http://www.sciencedirect.com/science/article/pii/S1568494620306293.

[72] A. Arora, A. Shrivastava, M. Mohit, L.S.-M. Lecanda, A. Aly, Cross-lingual transfer learning for intent detection of Covid-19 utterances, 2020.

[73] İ. Kırbaş, A. Sözen, A.D. Tuncer, F.Ş. Kazancıoğlu, Comperative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches, Chaos Solitons Fractals (2020) 110015.

[74] W.L. Taylor, "Cloze procedure": A new tool for measuring readability, Journalism Quart. 30 (4) (1953) 415–433.

[75] M. Schuster, K. Nakajima, JapanEse and korean voice search, in: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2012, pp. 5149–5152.
[76] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual BERT? in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4996–5001, http://dx.doi.org/10.18653/v1/P19-1493.
[77] S. Wu, M. Dredze, Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 833–844, http://dx.doi.org/10.18653/v1/D19-1077.
[78] K. Karthikeyan, Z. Wang, S. Mayhew, D. Roth, Cross-lingual ability of multilingual bert: An empirical study, in: International Conference on Learning Representations, 2019.
[79] B. Heinzerling, M. Strube, Sequence tagging with contextual and non-contextual subword representations: A multilingual evaluation, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 273–291, http://dx.doi.org/10.18653/v1/P19-1027.
[80] R. Hvingelby, A.B. Pauli, M. Barrett, C. Rosted, L.M. Lidegaard, A. Søgaard, DaNE: A named entity resource for danish, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 4597–4604.
[81] A. Neuraz, I. Lerner, W. Digan, N. Paris, R. Tsopra, A. Rogier, D. Baudoin, K.B. Cohen, A. Burgun, N. Garcelon, et al., Natural language processing for rapid response to emergent diseases: Case study of calcium channel blockers and hypertension in the COVID-19 pandemic, J. Med. Internet Res. 22 (8) (2020) e20773.
[82] M. Haider Syed, S. Khan, M. Raza Rabbani, Y.E. Thalassinos, An Artificial Intelligence and NLP Based Islamic FinTech Model Combining Zakat and Qardh-Al-Hasan for Countering the Adverse Impact of COVID 19 on SMEs and individuals, Eleftherios Thalassinos, 2020.
[83] E. Strubell, P. Verga, D. Belanger, A. McCallum, Fast and accurate entity recognition with iterated dilated convolutions, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2670–2680, http://dx.doi.org/10.18653/v1/D17-1283.
[84] J. Li, A. Sun, J. Han, C. Li, A survey on deep learning for named entity recognition, IEEE Trans. Knowl. Data Eng. (2020).
[85] P.J. Liu*, M. Saleh*, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, N. Shazeer, Generating Wikipedia by summarizing long sequences, in: International Conference on Learning Representations, 2018.
[86] N. Kitaev, D. Klein, Constituency parsing with a self-attentive encoder, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (vol. 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2676–2686, http://dx.doi.org/10.18653/v1/P18-1249.
[87] T. Kudo, J. Richardson, Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 66–71, http://dx.doi.org/10.18653/v1/D18-2012.
[88] Ž. Agić, I. Vulić, JW300: A wide-coverage parallel corpus for low-resource languages, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3204–3210, http://dx.doi.org/10.18653/v1/P19-1310.
[89] Y. Si, J. Wang, H. Xu, K. Roberts, Enhancing clinical concept extraction with contextual embeddings, J. Amer. Med. Inform. Assoc. 26 (11) (2019) 1297–1304, http://dx.doi.org/10.1093/jamia/ocz096.
[90] M. Jiang, T. Sanger, X. Liu, Combining contextualized embeddings and prior knowledge for clinical named entity recognition: Evaluation study, JMIR Med. Inform. 7 (4) (2019) e14850.
[91] K.S. Kalyan, S. Sangeetha, SECNLP: A survey of embeddings in clinical natural language processing, J. Biomed. Inform. 101 (2020) 103323, http://dx.doi.org/10.1016/j.jbi.2019.103323.
[92] J. Lafferty, A. McCallum, F.C. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, 2001.
[93] D. Hendrycks, K. Gimpel, Bridging nonlinearities and stochastic regularizers with gaussian error linear units, 2016.
[94] S. Mayhew, T. Tsygankova, D. Roth, Ner and pos when nothing is capitalized, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 6256–6261, http://dx.doi.org/10.18653/v1/D19-1650.
[95] L. Ramshaw, M. Marcus, Text chunking using transformation-based learning, in: Third Workshop on Very Large Corpora, 1995.