applied sciences

MDPI

Article

Iterative Annotation of Biomedical NER Corpora with Deep Neural Networks and Knowledge Bases

Stefano Silvestri *,† **, Francesco Gargiulo** † **and Mario Ciampi**

Institute for High Performance Computing and Networking of National Research Council, ICAR-CNR, Via Pietro Castellino 111, 80131 Naples, Italy; francesco.gargiulo@icar.cnr.it (F.G.); mario.ciampi@icar.cnr.it (M.C.)

* Correspondence: stefano.silvestri@icar.cnr.it

† These authors contributed equally to this work.

Abstract: The large availability of clinical natural language documents, such as clinical narratives or diagnoses, requires the definition of smart automatic systems for their processing and analysis, but the lack of annotated corpora in the biomedical domain, especially in languages different from English, makes it difficult to exploit the state-of-art machine-learning systems to extract information from such kinds of documents. For these reasons, healthcare professionals lose big opportunities that can arise from the analysis of this data. In this paper, we propose a methodology to reduce the manual efforts needed to annotate a biomedical named entity recognition (B-NER) corpus, exploiting both active learning and distant supervision, respectively based on deep learning models (e.g., Bi-LSTM, word2vec FastText, ELMo and BERT) and biomedical knowledge bases, in order to speed up the annotation task and limit class imbalance issues. We assessed this approach by creating an Italian-language electronic health record corpus annotated with biomedical domain entities in a small fraction of the time required for a fully manual annotation. The obtained corpus was used to train a B-NER deep neural network whose performances are comparable with the state of the art, with an F1-Score equal to 0.9661 and 0.8875 on two test sets.

Keywords: biomedical NER; corpus annotation; distant supervision; active learning; deep learning

Citation: Silvestri, S.; Gargiulo, F.; Ciampi, M. Iterative Annotation of Biomedical NER Corpora with Deep Neural Networks and Knowledge Bases. *Appl. Sci.* **2022**, *12*, 5775. https://doi.org/10.3390/app12125775

Academic Editor: Federico Divina

Received: 10 May 2022
Accepted: 4 June 2022
Published: 7 June 2022

1. Introduction

Nowadays, a huge amount of digitised information is produced in clinical and healthcare domains. A large part of this data is formed by or contains natural language (NL) texts, such as electronic health records (EHRs), diagnoses, medical reports, or patient summaries. Extracting and analysing the information in these documents has a great potential for caregivers and policy makers, making possible to support and improve the quality of the healthcare [1,2]. On the other hand, this huge amount of NL text can be processed only through Natural Language Processing (NLP) systems able to automatically extract the required information. An essential NLP task for the Information Extraction (IE) from clinical and biomedical NL documents is the biomedical named entity recognition (B-NER) [3], namely the identification and the classification of words and multi-word expressions belonging to the biomedical domain. The information through NER can be leveraged for many purposes, ranging from primary and secondary use analyses [4] to the support for the standardisation and interoperability of clinical data [5].

Deep learning (DL)-based NER methodologies are actually the best performing approaches in terms of realising NER systems [6–8], but they actually have two main limits: they are strictly language- and domain-dependent and they need a large annotated corpus to train a deep neural network (DNN) with optimal results. The lack of annotated corpora is one of the open issues related to automatic clinical document analysis [9]. An annotated corpus can be obtained only through laborious and costly work performed by domain experts, who must manually analyse and annotate a large number of documents, following precise

guidelines in order to produce a high-quality corpus [10,11]. Thus, not many annotated corpora are freely available, especially in the clinical domain and in languages different from English. Some methods have been proposed in the literature trying to overcome the lack of these important resources by using unsupervised machine-learning (ML) [12–14] or rule-based (RB) approaches [15–17], but in both cases the quality of the results is not comparable with that obtained through the manual efforts of domain experts. Other recent works have leveraged cross-language approaches [18,19], but in these cases annotated training and test sets in at least one language are required, in addition to knowledge bases or multi-lingual language models. Methodologies able to ease the work of the experts in the realisation of annotated corpora are required to narrow the gap between automatic and manual annotation, to the end of speeding up the manual annotation process, lowering its cost and reducing the needed efforts [20].

Interesting approaches for the annotation of corpora in an easier and less costly way are based on active learning (AL) and distant supervision (DS). Active learning [21] is an iterative annotation process supported by an ML model. In the first step of this approach, a small dataset extracted from a bigger corpus must be manually annotated. This set is then used to train a machine-learning classifier, to the end of annotating automatically the rest of the corpus. Among these automatic annotations, a human oracle must select the samples with presumably high utility to improve the classifier training, eventually correcting wrong predictions caused by an incomplete or small available dataset. More complex methodologies have been also proposed to improve the selection of the new samples [22]. The selected new samples are then added to the annotated training set and the ML model is retrained, improving the overall classification results in the prediction phase of the unannotated corpus. This process can be iterated until stop criteria or optimal performances are reached. AL methods can generate annotated corpora with less human efforts, but often the data are biased, depending on the method used for the new samples' selection during each iteration and on the content of the original corpus [23].

Distant supervision [24] is a completely automatic approach and exploits the knowledge extracted from knowledge bases (KBs) such as thesauri or a dictionary, assuming that if a string in text is included in a KB, then that string can be automatically annotated as an entity. This approach has no human cost, but the resulting corpus usually suffers from incomplete and noisy annotations. Incomplete annotations are named entities not listed in the KB, which will not be automatically annotated in the training corpus. On the other hand, a noisy annotation is a partial identification of a named entity, due to the presence in the KB of only an entity part (e.g., missing some words of that entity) or due to slight differences between the entity listed in the thesaurus and the one in the corpus (e.g., the use of a synonym of one of the words in multi-word entity, or a plural version of the same word).

In this paper a methodology that leverages both AL and DS for the annotation of B-NER clinical corpora is proposed, addressing some of the issues of both approaches to improve the quality and the speed of the annotation process. Firstly, an AL-based annotation is performed, exploiting a deep-learning NER architecture as an automatic classifier. Then, biomedical KBs are used for DS annotation and dataset expansion through data augmentation, with the purpose of mitigating the class imbalance problems [25] that could affect the annotations obtained through AL. In the experimental assessment the contribution of different pretrained Word Embedding (WE) models trained on a closed biomedical domain corpus as input of the DNN is also analysed, in particular comparing the contribution of word2vec [26], FastText [27] and ELMo [28] with a fine-tuned BERT model [29] pretrained on a general domain corpus. The proposed approach was used to easily and rapidly create an Italian language B-NER annotated corpus with very little effort with respect to a fully manual annotation procedure. The obtained corpus was evaluated on the aforementioned B-NER task, achieving performances comparable with the state of the art, as demonstrated in the experimental assessment.

In summary, the main contributions of this paper are:

- An automatic annotation methodology for B-NER corpora based on AL and DS techniques;
- An analysis of the contribution of different clinical closed-domain WE models (including word2vec, FastText and ELMo models), compared to a fine-tuned BERT model trained on a general-domain document collection;
- The annotation of an Italian clinical B-NER corpus.

The paper is organised as follows: in the next Section 2, an overview of the recent related works is presented, mainly focusing on methods for the annotation of texts from clinical and biomedical domains. Then, the details of the proposed approach are described in Section 3. In Section 4, the experimental assessment and the obtained results are shown and discussed and, finally, in Section 5 the final considerations, conclusions and future works are highlighted.

2. Related Works

Many methodologies devoted to the support of the annotation of an NER corpus have been proposed in recent years. Some studies are related to the guidelines for manual annotation of large corpora [10,11], which are very important for ensuring that the domain experts will follow the same approach during the annotation process. Besides them, many automatic and semi-automatic methods based on active learning and distant supervision have been presented. In [30], several AL algorithms were implemented to produce and assess corpora for a clinical text classification task in detail to determine the assertion status of clinical concepts. The results demonstrated that AL strategies are able to generate better classification models than the passive learning method such as random sampling. In [20,22], different sample selections for AL methods devoted to the clinical concept extraction task were proposed and evaluated, demonstrating their effectiveness in terms of building effective and robust ML models, reducing the time and the efforts involved in manual annotation. The authors of [31] described an AL method for the annotation of a corpus formed by MEDLINE abstracts annotated with pathological named entities. They proposed two different annotations, namely a short annotation that maps well defined diseases, and a long annotation that describes longer statements related to pathological phenomena and observations. Then, they defined an AL approach, which introduces a sampling bias by focusing on the most uncertain annotation samples, generating the annotated corpus. A clustering-based AL approach for B-NER is described in [32]. A document vector representation is obtained through TF-IDF; shared nearest neighbour (SNN) clustering is used to select documents with higher informative content during the iterations of AL, following the assumption that documents sharing similar named entities provide less information to the ML classifier. This AL method achieved a sensible improvement compared with random selection.

The authors of [33] presented a method to support the annotation of proteins, leveraging and ensemble learning together with WE, recurrent convolutional neural network, logistic regression and support vector machine models to effectively classify whether the title of a journal publication provides the information needed to show that experimental evidence of protein function for a given protein annotation is presented in the publication, reducing the manual effort only to a simple final confirmation. Their approach proved to outperform the transformer-based BioBERT model [34] fine-tuned on the same data.

The work described in [35] investigates whether conditional random fields (CRF) can be efficiently trained for NER in German texts, by means of an iterative procedure combining self-learning with a manual annotation—active learning—component, which leverages a CRF-based annotation and a manual correction to iteratively increase and improve the available dataset. Their results showed that their approach enabled the training of more accurate models with the annotation of fewer, more relevant data points, which are most helpful for modelling training.

In [36], the authors described an approach to deploy an annotated corpus for NER with minimal data and a light effort from experts combining both statistical and rule-based

approaches. The authors of [24] proposed a novel approach to mitigate the incomplete and noisy annotations obtained from automatic annotation through DS. This approach is based on an instance selector, exploiting reinforcement learning. The selector chooses sentences from a candidate dataset to expand training data, improving the performances of a DL NER architecture. The instance selector is trained on a reward provided by the NER tagger. The authors of [37] provided a tool which is able to leverage and integrate the information from many available biomedical knowledge bases with the purpose, among the other things, of creating and annotating new corpora. In [38], the authors presented a method to reduce human efforts for the annotation of a clinical text classification corpus, exploiting weak supervision and deep representation. In detail, they annotated training data using KBs and a rule-based approach, and then they used WEs as deep representation features as input to different ML models. They proved that this approach is very effective when used to train a convolutional neural network, but needs many training samples and suffers when applied in multi-class problems. Other methods to annotate a corpus through DS using domain KBs and rule-based approaches are discussed in [15,16]. In these latter cases, the results are strongly dependent on the predefined rule set and the considered KBs.

In [39], a semi-supervised self-learning technique is presented to extend an Arabic sentiment annotated corpus with unlabeled data. In detail, a long short term memory (LSTM) neural network is used to train a set of models on a manually labeled dataset. These models were then used to extend the original corpus, ensuring an improvement in the Arabic sentiment classification task. In [40], an approach to automatically annotate EHRs is described. First, a DS based on KBs is used to create an annotated training set. Then, a weighted function of WEs was used to create a sentence-level vector representation of relevant expressions, which are used to train an ML classifier, with the purpose of assessing the presence, absence, or risk of urinary incontinence and bowel dysfunction. The resulting model outperformed a other rule-based models for annotation with a significant margin. In [41], the authors described an approach for the annotation of a B-NER corpus, exploiting an automatic translator and knowledge bases, such as UMLS or ICD9, which contain lists of medical domain terms. They first used automatic translators to convert the English language annotated corpus into Italian. Then KBs were used to address the limits of the machine translations when applied to the specific lexicon from the biomedical domain, improving in this way the quality of the obtained corpus. In [42], the authors proposed a method to enhance the performance of a DL biGRU-CRF model devoted to clinical named-entity recognition in the French language, exploiting medical terminologies. Regardless, we also compared the results of the proposed approach with a fine-tuned BERT model pretrained on a generic domain Italian corpus, leveraging it for both the AL phase, as well as for the analysis of the performances of the annotated biomedical NER corpus.

3. Methodology

The proposed annotation methodology can be split into two main phases: an iterative active learning phase, followed by a distant supervision phase.

3.1. Active Learning

In the preliminary step of the methodology, human experts have manually annotated a small number of documents extracted from an unannotated corpus. A small part of these annotated documents is used as a training set of a DL model, whereas the remaining annotated samples are used as a test set during all iterations of the AL phase, with the purpose of assessing the improvement obtained in each step and providing a stop criterion when no more performance increment is observed. The few samples of the training set can lead to poor performances in the DL model; on the other hand, the reduced time and efforts for the annotation of a small fraction of the whole corpus make this process affordable. At this point, experts will not annotate more documents, but they must simply review a subset of new documents from the whole dataset automatically annotated through DL, eventually correcting the wrong or missing predictions. These new annotated samples are then added

to the training set, in order to retrain the ML model with higher precision thanks to a larger training set. The same procedure, namely the selection and review/correction of new AL-annotated samples and the retraining of the DL model, must be iterated until no further improvements of the ML results are observed. Figure 1 illustrates a schematic representation of the proposed AL-based annotation procedure.

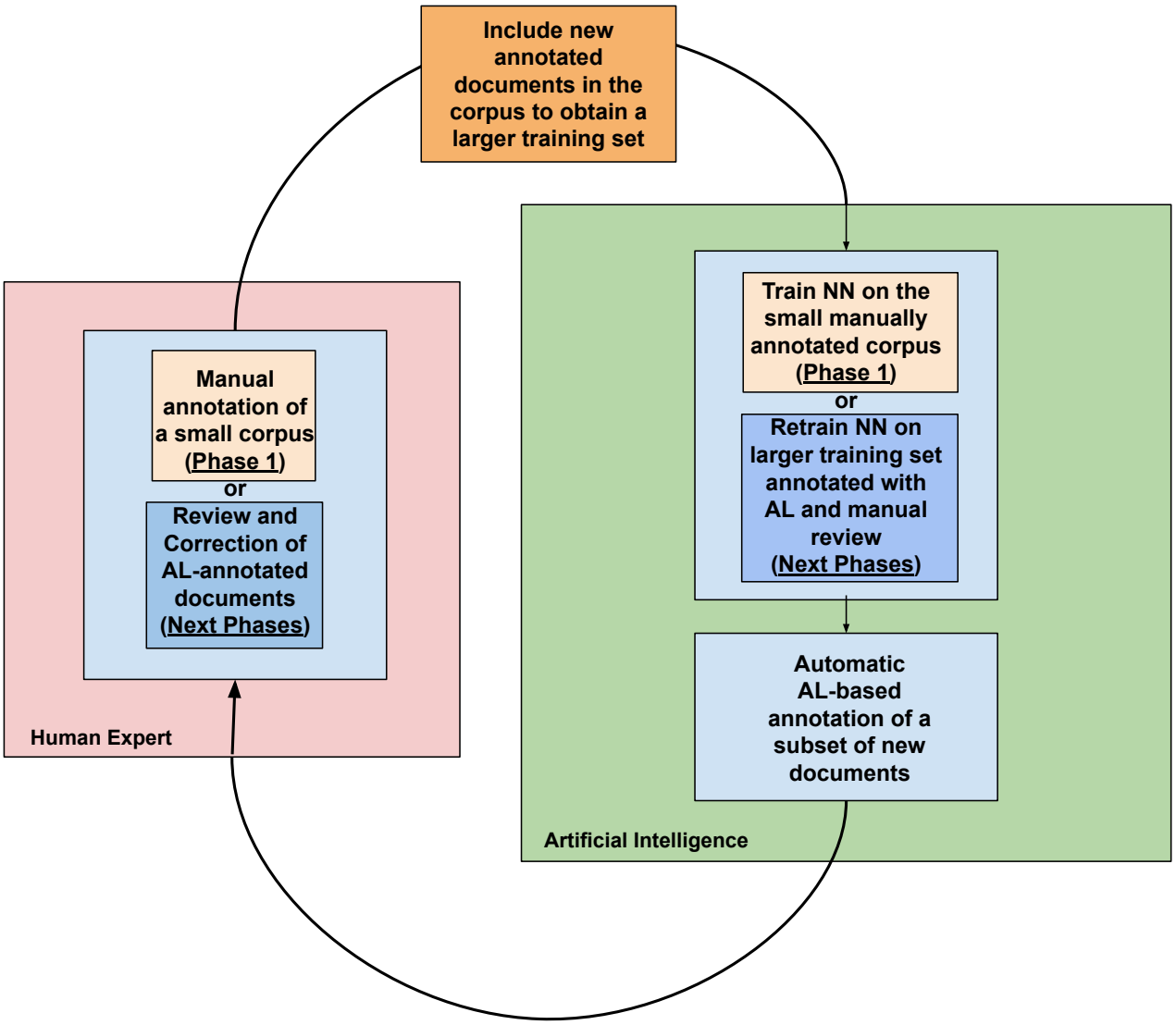

Figure 1. Schematic representation of the active learning annotation procedure.

The iterative AL annotation, followed by a manual review of the data, improves the quality of the obtained results with respect to a single-step AL annotation, because the effort of the human experts allows to correct any missing or wrong annotation obtained after each AL phase.

The selection of new samples from the dataset that will be annotated by the ML system is demanded of the domain experts without further support of automatic algorithms, such as those done in more complex AL approaches [22]. An improvement of the performance of the automatic annotation system is obtained using WE models trained on a biomedical closed domain corpus, as explained in Section 3.1.1. Deep neural network architectures are actually the state-of-the-art approaches for the B-NER task [3]. Thus, a DNN architecture for NER is used as an automatic ML classifier in the AL procedure. We adopted the classic DNN model presented in [43], known as Bi-LSTM CRF. This architecture is formed by the following layers: a bidirectional long short term memory (Bi-LSTM) character embedding

layer, concatenated with a pretrained WE layer, a Bi-LSTM layer for words and a conditional random field (CRF) layer, counting in total 166,082,553 parameters. The Bi-LSTM CRF model offers both good performance and reasonable training times. Moreover, the BERT model [29,34] pretrained on a general domain was also considered, in comparison with the Bi-LSTM CRF architecture.

3.1.1. Closed Domain Embedding Models

As mentioned above, the proposed methodology requires the preliminary training of an ML model in order to start the iterative AL process. In this first step, a manually annotated training set that counts few examples is used. While it does not require a long time to be manually annotated, its limited number of samples limits the performances of the ML system trained on it. In order to mitigate this issue, we represented the input text using WE models pretrained on biomedical-domain document collections [44], improving in this way the performance of the NER DNN. A higher precision of the results during the AL phase can provide a substantial help to the experts, further reducing the efforts required for the selection and correction of new samples. In particular, following the results described in [44–46], we conducted experiments with several WE models specifically trained on a biomedical closed-domain corpus. For this purpose, a further collection of documents related to the biomedical domain were collected in order to train the embedding models (see Section 4.3 for further details on this corpus). Five different WE models are tested: two word2vec (W2V) models [26], two fastText (FT) models [27], considering in both cases skip-gram and cbow algorithms, and ELMo [28], a contextual embedding model, pretrained on the Italian language biomedical domain, following the same approach presented for the BioELMo model in English [47].

We analysed the performance of these embedding models when used to represent the text in the first layer of the adopted DNN architecture, during the training of the AL model in the preliminary step of the proposed method, when only a small manually annotated training set is available. All embedding models during the subsequent steps of the proposed methodology are also tested to better underline their contribution when a larger training set is available. Finally, the results are compared with models trained on a very large Italian language general domain corpora: a word2vec model [48], provided by ISTI-CNR (the model is publicity available at https://github.com/MartinoMensio/it_vectors_wiki_spacy, accessed on 6 June 2022), and a BERT model [29], fine tuned on the B-NER task, as better explained in Section 3.1.2.

3.1.2. Fine-Tuned BERT Model

As explained above, we also adopted in our experimental assessment a BERT model [29], with the main purpose of comparing the performance of the Bi-LSTM CRF model with WEs trained on a biomedical closed-domain corpus, with a fine-tuned BERT model pretrained on a general domain corpus. In particular, we adopted the *bert-base-italian-xxl-uncased* model from the MDZ Digital Library team (dbmdz) BERT Italian model (https://huggingface.co/dbmdz/bert-base-italian-cased, accessed on 6 June 2022). This model is based on the *BERT-base* architecture, which is formed by a stack of 12 layers of decoder-only transformers [49], 768 hidden dimensional states and 12 attention heads. This model was pretrained on a very large general domain Italian corpus, whose size is 81 GB and counts 13,138,379,147 tokens, exploiting the masked language modelling approach, which consists in randomly applying a mask on a fraction of the words in the training corpus, encoding in this way information of the sentences from both directions and training at the same time the model to predict the masked words.

The transformer-based language models, such as BERT, allow for the transfer learning of the knowledge acquired through the pretraining on large corpora, as well as for the fine-tuning of the model on other tasks. Several pretrained BERT models are available in the literature due to the long time and computational resources required for the pretraining phase, as well as due to the need for collecting sufficiently large document collections. For

these reasons, we were not able to pretrain the BERT model on a biomedical closed-domain document corpus, neither a biomedical domain Italian language pretrained BERT model is available.

3.2. Distant Supervision Dataset Augmentation

Corpora annotated using ML-based methods are often affected by the problem of skewed class distribution [23]. An imbalanced class in the training set could limit the performance of a DNN trained with such corpora [50]. Undersampling or oversampling can help to mitigate the class imbalance problem [51], but undersampling can also lower the overall performances, deleting samples of all classes. With the purpose of improving the quality of the annotated corpus and resolving some of the problems related to class imbalance, a distant supervised annotation and augmentation after the AL phase is proposed. In detail, the annotated corpus is augmented with new samples belonging to the imbalanced classes, obtained through DS-exploiting domain KBs. The KB must contain a list of entities of the same class that must be augmented.

The dataset augmentation after the AL annotation is performed as follows. The sentences containing at least one entity belonging to the imbalanced classes are extracted from the corpus. Then, new sentences are obtained by substituting the named entities in these sentences with new entities of the same class randomly extracted from the respective KB. The process is iterated until a sufficient number of new sentences is obtained; that is, the respective class is less imbalanced, and, at the same time, all the entities from the KBs have been considered. In this way, we also include new entities in the dataset, in addition to reducing the class imbalance. Moreover, the augmentation process also oversamples the entities belonging to not imbalanced classes, providing in general more samples for all classes. As demonstrated by the results described in Section 4, this improves the overall performance, not only in the cases of imbalanced classes. Finally, the obtained new sentences are randomly reinserted in the corpus.

The whole proposed annotation methodology, including both iterative AL and DS phases, is represented in the block diagram depicted in Figure 2.

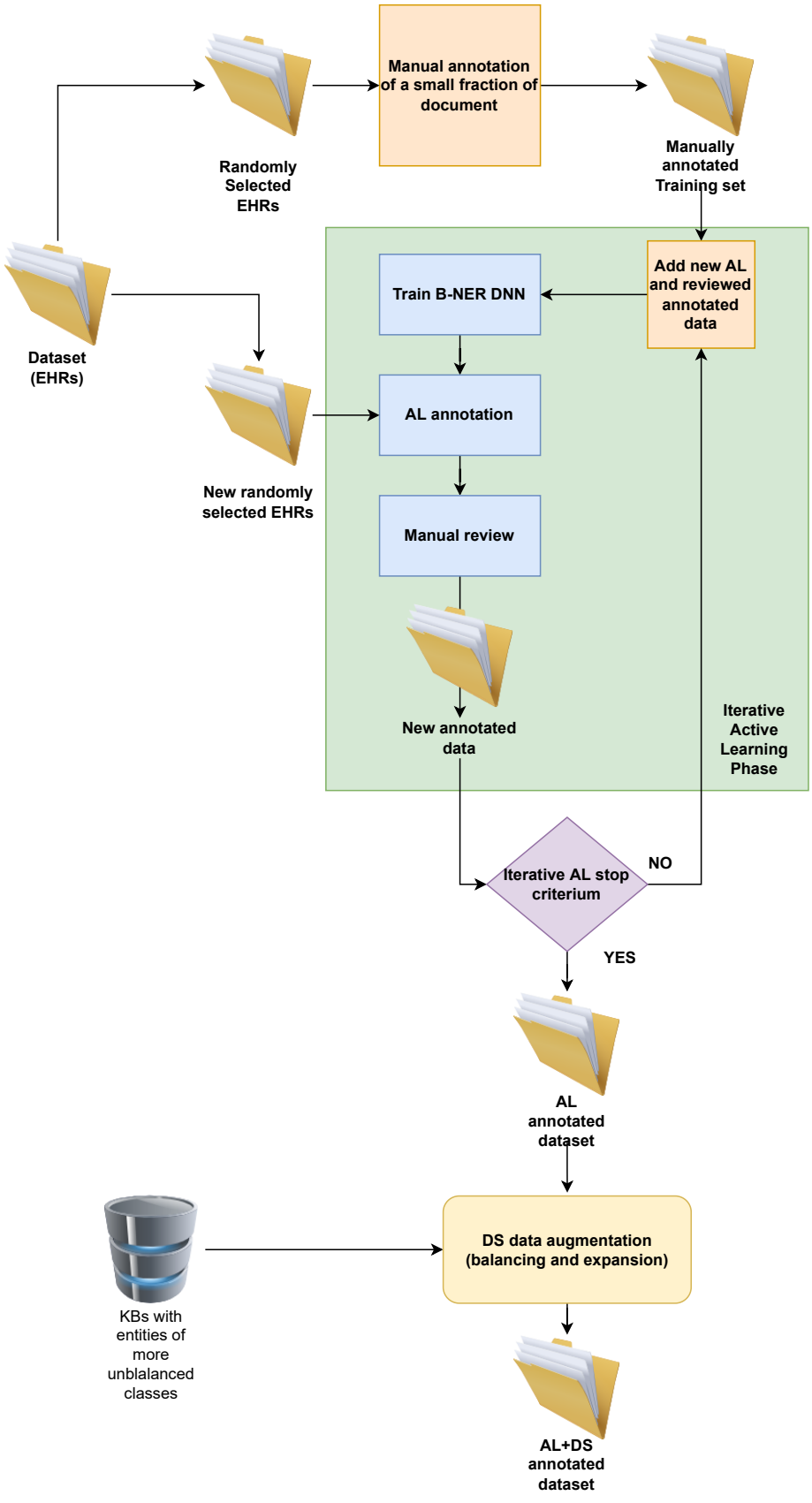

Figure 2. Schematic representation of the active learning annotation procedure.

4. Experimental Assessment and Discussion

In this section, in the following, after a description of the features of the original dataset and the details of the obtained annotated corpus, the performances of the DNN trained using the self-made corpus will be discussed, measured in terms of precision, recall and F1-score and considering also the contribution of the closed-domain WE models.

4.1. B-NER Annotated Corpora

The original unannotated dataset is formed by the narrative parts of NL text extracted from a set of 1011 anonymized EHRs in the Italian language, which has a total word count of 1,657,970. In detail, the dataset contains EHRs acquired from the eHealth systems of some different hospitals in Italy. As mentioned above, the EHRs had been previously anonymized and they are related to patients admitted to different departments of the hospitals. The content of these documents is relatively homogeneous, containing the clinical diary of the patients, where the causes of the admission to the hospitals, the diseases, the prognosis, the follow-ups, the exams, the procedures and the prescriptions are described. Some sample sentences extracted from two different EHRs (translated into English) are reported below.

- *Found sub-capital fracture and dislocation of left shoulder and contusion of right hip caused by accidental fall at home.*
- *Tomorrow follow-up exams.*
- *Patient admitted to cardiology from 9 February to 19 February due to episodes of arrhythmia, likely secondary to chronic renal failure.*

Eight different named-entity classes are identified, as shown in Table 1, following UMLS semantic types [52] and considering at the same time possible real-world applications of the trained ML models [2].

Table 1. Entity classes with corresponding acronyms and examples. The English translation of the examples is in italics between parentheses.

Class Type	Acronym	Examples
Diseases and Symptoms	DIS	Febbre (*Fever*), pressione alta (*High blood pressure*), cirrosi epatica (*liver cirrhosis*)
Drug names	DRU	Paracetamolo (*Paracetamol*), antibiotico (*antibiotic*)
Departments	DEP	Ortopedia (*Orthopedics*), pronto soccorso (*emergency room*)
Therapeutic procedures and Medical Instruments	THE	Ecografo (*ultrasound scanner*, profilassi antitrombotica (*thrombosis prophilaxis*), stent (*stent*)
Body Parts	BOD	Piede destro (*right foot*), testa dell'omero (*humeral head*), fegato (*liver*)
Measures	MEA	30 cc, 12 mm, 120 bpm
Dates	DAT	23 giugno 2012 (*23 June 2012*), oggi (*today*), ore 12:30 (*12:30*)
Diagnostic procedures or lab tests	ANA	Radiografia (*radiography*), valutazione cardiologica (*cardiac assessment*), glicemia (*glycaemia*), coronarografia (*angiography*)

As explained in Section 3.1, in the preliminary step of the annotation process a small set of documents formed by the text extracted from 25 randomly selected EHRs was manually annotated by two domain experts. The annotation procedure was conducted according to predefined guidelines, which describe general and specific annotation rules. The labelling

process followed the *IOB* notation [53], i.e., each token belonging to an entity is labelled with the corresponding class adding the prefix *B* (Begin) if it is the first token of the entity, the prefix *I* for all subsequent tokens of the same multi-word entity and the tag *O* (Outside) if the token does not belong to an entity. The result of the manual annotation is a small dataset, which counts 7421 tokens and 1963 named entities, as shown in the first row of Table 2. The experts worked for approximately eight hours to produce this dataset, including the discussion about conflicts and disambiguation of the conflicting annotations. To the end of providing a stop criterion for the iterative AL phase (see Section 3.1), a further test set, which counts 21,133 tokens, was also manually annotated.

Table 2. Number of words and annotated entities in each step of the iterative AL annotation procedure.

Step	Word Number	Entity Number
1	7421	1963
2	20,083	5621
3	32,856	9285
4	78,449	21,914
5	133,200	37,029
6	201,956	55,601
7	304,797	60,669

This small dataset is used to train the DNN Bi-LSTM-CRF [43] architecture. This DL model has been used to automatically annotate new documents randomly extracted from the whole dataset, starting the iterative AL phase. In each iteration, the human experts had to review the correctness of the annotations produced by the DL model, eventually correcting the wrong or the missing ones. They worked each step for approximately eight hours, but, in this case, they were able to annotate wider datasets, thanks to the reduced effort provided by the partial annotation of the data, as shown in Table 2. The new data obtained in each iteration were added to the training set, producing a larger dataset, which was used to retrain the DL model. The same process was iterated and at each step the experts were able to speed up the annotation process, producing at the same time an increasing number of annotations thanks to the higher precision of the DL model trained on a larger and more complete dataset (see Table 2 for the details). The iterative AL process was stopped after seven iterations (see Section 4.3) when no more notable performance improvements of the ML model were observed. At the end of the AL phase, a corpus counting 304,798 words and 60,669 entities was annotated.

The results shown in Table 2 demonstrate that the proposed approach allows one to obtain a sensitive improvement of the time required for the annotation, with respect to a fully manual process. In the first preliminary step, a human expert was able to annotate a document collection formed by almost 8000 words in about 8 h, with a rate of 1000 words per hour. The dataset obtained through the AL phase counts 304,797 words: considering the same annotation rate of the preliminary step, the fully manual annotation of this dataset would have required about 300 h. The proposed AL process required seven steps where the experts reviewed and corrected the annotations of the new data obtained from the DNN for about 8 h for each step, with a total manual effort of 56 h. Moreover, the process required an average training time of the DNN equal to 1.5 h for each iteration (the training time increases with larger training sets) on the hardware used for the experimental assessment (see Section 4.2). In summary, the proposed iterative AL phase required in total about 66 h, allowing one to obtain an annotated dataset in almost 1/5 of the time required by a fully manual annotation.

Table 3 shows the distribution of the classes in the dataset obtained at the end of the AL phase. We note that there are very few examples of DEP (Departments) and DRU (Drugs) classes. This skewed class distribution can limit the performances of the ML systems, in particular for these two specific classes (see next Table 7). Then, in order to mitigate the

skewed class distribution, the annotated corpus was automatically augmented using DS with our proposed approach, exploiting knowledge sources related to the more imbalanced classes, such as a complete list of drugs and pharmaceutical substances extracted from the Pharmaceutical Reference Book officially maintained by the Agenzia Italiana del Farmaco (https://farmaci.agenziafarmaco.gov.it/bancadatifarmaci/cerca-farmaco, accessed on 6 June 2022), the Italian government agency in charge for drug administration, and a list of medical departments was obtained from the main Italian medical centre (hospitals, clinical facilities, etc.) websites. These two KBs were used to expand the corpus, applying the data augmentation/oversampling, as described in Section 3.2. The final resulting annotated corpus has a total word count equal to 1,699,028 and a total entity count equal to 424,776. In Table 3, it is shown that the distribution of the samples after the DS augmentation clearly reduces the original skewness.

Table 3. Number of entities in the annotated corpus before and after the application of DS entity expansion.

Class Type	Entity Number	
	No Expansion	Expansion
MEA	12,168	65,668
DRU	2046	45,336
DEP	1099	25,469
THE	8170	46,900
BOD	11,423	33,203
DIS	31,179	125,059
DAT	4933	34,263
ANA	12,258	48,878
Total	60,669	424,776

The final corpus was split into a training set and a test set, randomly selecting about 15% of the data for the test and the remaining data for the training. In this way, the entity classes, respectively, in the training set and the test set are distributed as shown in Table 4. The test set was used to assess the performance of the DNN with the annotated corpus.

Table 4. Number of entities in the final annotated corpus, split into test set and training set.

Class Type	Entity Number	
	Test Set	Training Set
MEA	9458	56,210
DRU	6624	38,712
DEP	3860	21,609
THE	6859	40,041
BOD	4539	28,664
DIS	17,354	107,705
DAT	4920	29,343
ANA	7055	41,823
Total	60,669	364,107

Finally, a further test set was also manually annotated by the domain experts, extracting documents from a different medical domain document collection, with the purpose of assessing the quality of the corpus obtained with the proposed methodology. The aforementioned document collection, named hereinafter *out-of-corpus*, is formed by short medical notes and diagnoses from various medical departments and counts 15,728 words and 3816 entities. A common problem of the Bi-LSTM CRF B-NER architecture is that it often fails to generalise to out of vocabulary words, namely words that do not appear in the training

set [54]. Thus, we tested the DL model also on the out-of-corpus test set, which contains many named entities not present in the original dataset.

4.2. Hardware

The AL phase requires the availability of hardware equipped with GPUs capable of training the DNN in a reasonable time. The hardware used in our experiments was a dual CPU Intel Xeon E5-2630, clocked at 2.2 GHz, with 256 GB of RAM and 1TB SDD, equipped with four Nvidia Titan X 1080 GPU with 11 GB of VRAM. With this system, the average time required to train the DNN during each iteration of the AL phase was about 1.5 h, considering that the training time increases with the size of the dataset.

4.3. Performances

To verify the effectiveness of the annotated corpus, we evaluated the performance of the same DNN used in the AL phase, trained on the obtained corpus. As explained above, we also tested different WE models to represent the input of the DNN, whose details are reported below.

Firstly, we considered a word2vec model [26] trained on a general domain Italian language corpus, hereinafter called *W2V ISTI*, formed by a Wikipedia dump and a collection of 31,432 novels [48]. This document collection is very large (242,261,172 sentences and 2,534,600,769 words), and its content is related to many knowledge fields. The training parameters used for this model are: skip-gram algorithm, vector size 300, window size 10 and negative samples 10.

Then, a more specific biomedical closed domain text corpus, hereinafter *BIO-Corpus*, was used to train the embedding models. This corpus was created considering different biomedical sources, in detail: (i) a dump of a selection of Italian Wikipedia pages related to medicine, biology, healthcare and other similar domains, following the procedure and the tools described in [45]; (ii) the text extracted from the package leaflets of all drugs available in Italy, downloading all pdf files from Agenzia Italiana del Farmaco (AIFA) and extracting the corresponding text exploiting Apache Tika (https://tika.apache.org/, accessed on 6 June 2022) and some specific Python scripts; (iii) the text extracted from the Italian Medical Dictionary of the *Corriere della Sera* (https://www.corriere.it/salute/dizionario/, accessed on 6 June 2022) through a set of custom web scraping Python scripts; and (iv) the text extracted from other Italian biomedical documents freely available online, such as scientific papers, presentations, technical reports and other things, exploiting also in this case Tika pipelines and Python scripts. The BIO-corpus is made up of 2,160,704 sentences and 511,649,310 words and it was used as a training set for five different WE models: two word2vec (W2V) [26] models and two FastText (FT) models [27], considering in both cases skip-gram and cbow algorithms and setting the vector size equal to 300, the window size equals to 10, the negative samples equals to 10 and, in the case of FastText embeddings, the char n-gram size varying from 3 to 6, as well as one contextual embedding model based on ELMo [28]. These latter models trained on the BIO-Corpus were called, respectively, *W2V cbow*, *W2V skip*, *FT cbow*, *FT skip* and *ELMo*.

Finally, we also tested the obtained annotated corpus by fine-tuning the BERT model pretrained on a very large general domain corpus, previously described in Section 3.1.2.

Table 5 shows the results obtained on the manually annotated test set (see Section 3.1) in the preliminary step of the AL phase when the DNN has been trained with few manually annotated data. The results are in terms of F1-Score, precision and recall averaged over all classes. It is possible to observe that *ELMo* embeddings trained on the BIO-Corpus and used to represent the input to the DNN obtain performances sensibly higher than the other cases, despite a training set with few samples. This can provide substantial help to the experts during the next steps of the AL phase, further reducing the effort in the correction of wrong predictions.

Thus, this model was selected for further steps of the AL phase for the annotation of the B-NER corpus, as well as the input layer of the DNN used to test the effectiveness of the

annotated corpus. Moreover, the experiments also considered the fine tuning of the BERT model pretrained on a general domain document collection, that being the current reference model for NER tasks in the literature and because it obtained performances comparable to the ELMo case in the preliminary step of the proposed approach.

Table 5. Results in terms of F1-Score, precision and recall averaged on all classes obtained during the first training of the DNN of the AL phase, using different pretrained embedding models.

WE Model	F1-Score	Precision	Recall
W2V ISTI	0.4520	0.5274	0.4624
W2V cbow	0.3905	0.4764	0.3738
W2V skip	0.4734	0.6062	0.4131
FT cbow	0.3758	0.3976	0.4478
FT skip	0.4611	0.4438	0.4913
ELMo	**0.6900**	**0.6758**	**0.7078**
BERT	0.6787	0.6557	0.7034

The AL-based iterative annotation stopped when no further improvements to the results were obtained. Seven iterations are considered empirically sufficient to produce in the AL phase an annotated corpus with 304,977 words and 60,669 entities. Table 6 shows the performance improvements obtained in the test set at each step of the iterative AL procedure, using the ELMo model with BiLSTM CRF and the BERT model. As shown in Table 6, increasing the size of the annotated corpus during the steps of the iterative AL phase improved the performances of both the ELMo and the BERT experiments. We also note that the ELMo model pretrained on the biomedical domain corpus performs slightly better when fewer data in the training set are available during the first iterations of the procedure, while, when larger training sets are obtained during the AL phases, the BERT model pretrained on a general domain corpus obtains slightly better results. In any case, both models obtain comparable performances, demonstrating that a simpler neural language model, such as ELMo, pretrained on the biomedical domain corpus obtains performances comparable with the ones produced by a more complex DNN, such as BERT, pretrained on a general domain corpus. Then, we focused the next phase of the experimental assessment only on the ELMo model, investigating the contribution of the DS data augmentation phase.

Table 6. Performance of the best performing DNNs (ELMo Bi-LSTM CRF and BERT fine tuned) at each step of the AL phase of the annotation procedure, in terms of precision, recall and F1-Score averaged over all classes.

	ELMo BiLSTM-CRF			**BERT Fine Tuned**		
Iteration Step	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
1	0.6758	0.7078	0.6900	0.6557	0.7034	0.6787
2	0.7195	0.7504	0.7346	0.7187	0.7517	0.7349
3	0.7269	0.7567	0.7406	0.7252	0.7638	0.7440
4	0.7364	0.7697	0.7522	0.7449	0.7738	0.7590
5	0.7552	0.7743	0.7646	0.7581	0.7849	0.7712
6	0.7629	0.7767	0.7695	0.7639	0.7915	0.7775
7	0.7635	0.7889	0.7760	0.7790	0.8001	0.7893

In Table 7, the results of the ELMo experiment obtained in the last step of the AL phase are highlighted, showing the precision, recall and F1-Score obtained for each class of the dataset. Observing at the same time the left column of Table 3, where the number of entities of each class are shown, and Table 7, with the results obtained by the DNNs trained on the corpus obtained at the end of the AL phase, it is possible to note that the worst

performances were obtained in the cases of the entities belonging to the more imbalanced classes, namely DRU (Drugs) and DEP (Departments), also limiting the average results.

Table 7. Performance of the ELMo Bi-LSTM CRF at the last step of the AL phase of the annotation procedure, in terms of precision, recall and F1-Score for each class.

Entity Type	Precision	Recall	F1-Score
MEA	0.8436	0.8599	0.8517
DRU	0.8085	0.3576	0.4959
DEP	0.1845	0.1404	0.1595
THE	0.5668	0.8459	0.6788
BOD	0.8283	0.8949	0.8603
DIS	0.8316	0.9125	0.8702
DAT	0.8905	0.9492	0.9189
ANA	0.8145	0.9137	0.8612
Average	0.7635	0.7889	0.7760

We introduced the DS data augmentation phase in order to limit this issue. After the expansion and the balancing of the training set using the second part of the proposed approach, where new sentences are obtained leveraging DS with domain KBs containing lists of entities of two more imbalanced classes, the performance of the ELMo DNN trained on the training set obtained with both the AL and DS phases are sensibly improved, as shown in Table 8. In this case, we reported only the results obtained by the best performing model, which was the Bi-LSTM CRF architecture with the ELMo embeddings. This behaviour is expected, due to overfitting issues of the BERT model trained on very large datasets [55].

In particular, comparing the obtained results for DRU and DEP classes in Table 8, where the DS augmentation for balancing and expansion were applied for the annotation of the training set after the AL, with the results achieved in the same class types shown in Table 7, where only the AL is performed, it is possible to observe that the DS augmentation applied to the most unbalanced classes DEP and DRU provided a sensible performance boost. Moreover, we can also note an improvement in all the other classes thanks to the oversampling performed during the DS data augmentation.

Table 8. Results obtained by the ELMo Bi-LSTM CRF trained with the final annotated corpus (AL and DS augmentation) in terms of precision, recall and F1-Score for each class.

Entity Type	Precision	Recall	F1-Score
MEA	0.9636	0.9675	0.9655
DRU	0.9863	0.9893	0.9878
DEP	0.9878	0.9860	0.9869
THE	0.9609	0.9636	0.9622
BOD	0.9203	0.9262	0.9232
DIS	0.9595	0.9634	0.9615
DAT	0.9783	0.9809	0.9796
ANA	0.9647	0.9718	0.9682
Average	0.9642	0.9679	0.9661

To the end of further verifying the effectiveness of the final obtained annotated corpus, we also tested the DNN models on the out-of-corpus test set, previously described in Section 4.1. This additional manually annotated test set was extracted from a different document collection, which contains many entities not present in the dataset used to build and annotate the training set. Table 9 shows the results obtained by the ELMo BiLSTM CRF architecture trained on the final annotated corpus and tested on the out-of-corpus test set. It is worth noting that, despite a slight performance drop, the DNN model still performs at a good level, assessing the effectiveness of the obtained annotated training corpus.

Table 9. Results in terms of precision, recall and F1-Score averaged on all classes obtained with the DNN with ELMo embeddings trained on the final annotated corpus and tested on the out-of-corpus test set.

Entity Type	Precision	Recall	F1-Score
MEA	0.9374	0.9051	0.9210
DRU	0.6429	0.8824	0.7438
DEP	0.9000	1.000	0.9474
THE	0.7983	0.8559	0.8261
BOD	0.9112	0.8701	0.8902
DIS	0.8475	0.9278	0.8858
DAT	0.5608	0.9222	0.6975
ANA	0.9364	0.8805	0.9076
Average	0.8809	0.8986	0.8875

In summary, these results demonstrate that the DS data augmentation phase is capable of further improving the quality of the dataset obtained from the previous iterative AL phase, mitigating the issues of the AL related to unbalanced classes and out-of-corpus named entities.

Finally, the next Table 10 reports the metrics averaged on all classes obtained by each considered DNN model, namely the Bi-LSTM CRF with the various considered WE models as input layer and the fine-tuned BERT model, trained on the final annotated dataset (AL and DS) and tested on the out-of-corpus test set. The purpose of this last experiment is to evaluate the contribution of different neural language models on a corpus containing many named entities not present in the training set. The results in Table 10 show that the WE model trained on a biomedical closed-domain document collection (W2V cbow, W2V skip, FT cbow and FT skip) provides sensible improvements with respect to the W2V ISTI model, trained on a general domain corpus. We also note that the WEs trained using the skipgram algorithm provide improved performance with respect to the cbow algorithm. The ELMo model produces the best performance, but the simpler W2V skip model also obtains good results, although it does not reach the performance obtained by more complex ELMo and BERT architectures. As in the previous case, the performances of the BERT model are limited by the overfitting issues, although we adopted a drop-out rate equal to 0.7 to limit them, following the literature [55].

Table 10. Results in terms of F1-Score, precision and recall obtained by the DNN on the out-of-corpus test set, using different pretrained WE models.

WE Model	Precision	Recall	F1-Score
W2V ISTI	0.7794	0.7714	0.7714
W2V cbow	0.8047	0.8000	0.8010
W2V skip	0.8676	0.8464	0.8545
FT cbow	0.8164	0.8143	0.8125
FT skip	0.8367	0.8107	0.8213
ELMo	**0.8809**	**0.8986**	**0.8875**
BERT	0.7356	0.7246	0.7301

5. Conclusions

This paper presented an approach based on both active learning and distant supervision, which makes the manual annotation of a corpus for biomedical named entity recognition (B-NER) a less costly process, reducing the efforts needed by human experts. In detail, the method is based on a first AL phase, where a DNN architecture for NER composed of a BiLSTM-CRF is used to support the manual annotation. When no further improvements are achieved by the AL-based process, the corpus is augmented using DS, exploiting domain KBs, in order to mitigate the class imbalance. Finally, an assessment of the utility of using a WE model trained on a closed domain document collection as input

for the DNN was carried out, considering word2vec, FastText and ELMo embeddings, and also comparing the obtained results with the fine tuned BERT model pretrained on a very large general domain document collection.

The approach was tested by creating an Italian language B-NER corpus used to train different B-NER DNNs. The experiments demonstrated that the obtained corpus is capable of training a B-NER DNN with very good performance, allowing one to annotate an NER corpus in a fraction of the time required for a fully manual annotation. Moreover, they showed that the pretraining of the ELMo contextual embedding model on a biomedical closed domain corpus allows one to obtain results comparable with the more complex BERT architecture pretrained on a very large general domain document collection, which demands more computational resources.

The proposed annotation methodology can facilitate the development and the implementation of AI-powered information extraction and indexing systems, improving the management of large natural language document collections, as well as supporting the analysis and the extraction of knowledge from such documents. On the other hand, a limit of the proposed approach is that KBs in the domain and the language of the annotations must be available to apply the DS phase. Moreover, the method is not fully automatic, requiring in any case human supervision, as well as a fully manual annotation in the preliminary phase. It also requires the availability of DL-dedicated hardware to carry out the AL phase in a reasonable time. Finally, the training of the NLM (in particular, the BERT-based models) requires the collection of a very large closed-domain unannotated document corpus, which in some cases may not be easy to obtain.

In future work, the B-NER DL model trained on the obtained annotated corpus on more out-of-corpus documents, such as medical tweets or scientific papers, assessing the effectiveness of the proposed annotation methodology will be evaluated. Moreover, we want to collect a very large biomedical closed domain corpus in order to pretrain a domain-specific Italian biomedical BERT model, following the BioBERT [34] approach, in order to further test the proposed annotation approach.

Finally, the presented annotation methodology could be applied to other languages and domains in order to demonstrate its general validity. In particular, the same approach was also developed, tailored and tested for the annotation of a cyber security (CS) English NER corpus, exploited for an innovative ML-based threat assessment methodology [56] proposed in the EC-funded AI4HEALTHSEC project (https://www.ai4healthsec.eu, accessed on 6 June 2022). In this case, a large document collection was previously extracted from a CS news website, allowing for the creation of an unannotated training set for the neural language models, while CAPEC (https://capec.mitre.org, accessed on 6 June 2022) and CPE (https://nvd.nist.gov/products/cpe, accessed on 6 June 2022) KBs were used in the DS phase for the annotation of CS threats and the corresponding assets. Moreover, a CS closed-domain BERT model was also exploited, confirming the effectiveness of the use of a closed-domain transformer-based NLM.

Author Contributions: Conceptualization, S.S. and F.G.; methodology, S.S. and F.G.; software, S.S. and F.G.; validation, S.S., F.G. and M.C.; formal analysis, S.S. and F.G.; investigation, S.S. and F.G.; resources, S.S., F.G. and M.C. ; data curation, S.S. and F.G.; writing—original draft preparation, S.S. and F.G.; writing—review and editing, S.S., F.G. and M.C.; visualization, S.S. and F.G.; supervision, M.C.; project administration, M.C.; funding acquisition, M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the European Commission, grant number 883273, AI4HEALTHSEC—A Dynamic and Self-Organized Artificial Swarm Intelligence Solution for Security and Privacy Threats in Healthcare ICT Infrastructures.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Simona Sada and Giuseppe Trerotola for the technical and administrative support.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Yadav, P.; Steinbach, M.; Kumar, V.; Simon, G.J. Mining Electronic Health Records (EHRs): A Survey. *ACM Comput. Surv.* **2018**, *50*, 1–40. [CrossRef]
2. Silvestri, S.; Esposito, A.; Gargiulo, F.; Sicuranza, M.; Ciampi, M.; De Pietro, G. A Big Data Architecture for the Extraction and Analysis of EHR Data. In Proceedings of the 2019 IEEE World Congress on Services (SERVICES), Milan, Italy, 8–13 July 2019; Volume 2642-939X; pp. 283–288. [CrossRef]
3. Shickel, B.; Tighe, P.; Bihorac, A.; Rashidi, P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE J. Biomed. Health Inform.* **2018**, *22*, 1589–1604. [CrossRef] [PubMed]
4. Abadeer, M. Assessment of DistilBERT performance on Named Entity Recognition task for the detection of Protected Health Information and medical concepts. In Proceedings of the 3rd Clinical Natural Language Processing Workshop, Online, 16–20 November 2020; pp. 158–167. [CrossRef]
5. Oemig, F.; Blobel, B. Natural Language Processing Supporting Interoperability in Healthcare. In *Text Mining: From Ontology Learning to Automated Text Processing Applications*; Biemann, C., Mehler, A., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 137–156. [CrossRef]
6. Yadav, V.; Bethard, S. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In Proceedings of the 27th International Conference on Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 2145–2158.
7. Lewis, P.; Ott, M.; Du, J.; Stoyanov, V. Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art. In Proceedings of the 3rd Clinical Natural Language Processing Workshop, Online, 16–20 November 2020; pp. 146–157. [CrossRef]
8. Weber, L.; Sänger, M.; Münchmeyer, J.; Habibi, M.; Leser, U.; Akbik, A. HunFlair: An easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics* **2021**, *37*, 2792–2794. [CrossRef]
9. Xiao, C.; Choi, E.; Sun, J. Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review. *JAMIA* **2018**, *25*, 1419–1428. [CrossRef] [PubMed]
10. Patel, P.; Davey, D.; Panchal, V.; Pathak, P. Annotation of a Large Clinical Entity Corpus. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 2033–2042.
11. Xia, F.; Yetisgen-Yildiz, M. Clinical corpus annotation: Challenges and strategies. In Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM'2012) in conjunction with the International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey, 21–27 May 2012.
12. Alicante, A.; Corazza, A.; Isgrò, F.; Silvestri, S. Unsupervised entity and relation extraction from clinical records in Italian. *Comput. Biol. Med.* **2016**, *72*, 263–275. [CrossRef] [PubMed]
13. Wangpoonsarp, A.; Shimura, K.; Fukumoto, F. Unsupervised Predominant Sense Detection and Its Application to Text Classification. *Appl. Sci.* **2020**, *10*, 6052. [CrossRef]
14. Nadif, M.; Role, F. Unsupervised and self-supervised deep learning approaches for biomedical text mining. *Briefings Bioinform.* **2021**, *22*, 1592–1603. [CrossRef]
15. Ghiasvand, O.; Kate, R.J. Learning for clinical named entity recognition without manual annotations. *Inform. Med. Unlocked* **2018**, *13*, 122 – 127. [CrossRef]
16. Diomaiuta, C.; Mercorella, M.; Ciampi, M.; Pietro, G.D. A novel system for the automatic extraction of a patient problem summary. In Proceedings of the 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, Greece, 3–6 July 2017; pp. 182–186. [CrossRef]
17. Hammami, L.; Paglialonga, A.; Pruneri, G.; Torresani, M.; Sant, M.; Bono, C.; Caiani, E.G.; Baili, P. Automated classification of cancer morphology from Italian pathology reports using Natural Language Processing techniques: A rule-based approach. *J. Biomed. Inform.* **2021**, *116*, 103712. [CrossRef]
18. Silvestri, S.; Gargiulo, F.; Ciampi, M.; De Pietro, G. Exploit Multilingual Language Model at Scale for ICD-10 Clinical Text Classification. In Proceedings of the 2020 IEEE Symposium on Computers and Communications (ISCC), Rennes, France, 7–10 July 2020; pp. 1–7. [CrossRef]
19. Suárez-Paniagua, V.; Dong, H.; Casey, A. A multi-BERT hybrid system for Named Entity Recognition in Spanish radiology reports. In Proceedings of the Working Notes of CLEF 2021—Conference and Labs of the Evaluation Forum, Bucharest, Romania, 21–24 September 2021; Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F., Eds.; CEUR-WS.org: Bucharest, Romania, 2021; Volume 2936, *CEUR Workshop Proceedings*; pp. 846–856.
20. Kholghi, M.; Sitbon, L.; Zuccon, G.; Nguyen, A. Active learning reduces annotation time for clinical concept extraction. *Int. J. Med. Inform.* **2017**, *106*, 25 – 31. [CrossRef]
21. Cohn, D.A.; Ghahramani, Z.; Jordan, M.I. Active Learning with Statistical Models. *J. Artif. Intell. Res.* **1996**, *4*, 129–145. [CrossRef]

22. Kholghi, M.; Sitbon, L.; Zuccon, G.; Nguyen, A.N. Active learning: A step towards automating medical concept extraction. *JAMIA* **2016**, *23*, 289–296. [CrossRef]
23. Tomanek, K.; Hahn, U. Reducing class imbalance during active learning for named entity annotation. In Proceedings of the 5th International Conference on Knowledge Capture (K-CAP 2009), Redondo Beach, CA, USA, 1–4 September 2009; pp. 105–112. [CrossRef]
24. Yang, Y.; Chen, W.; Li, Z.; He, Z.; Zhang, M. Distantly Supervised NER with Partial Annotation Learning and Reinforcement Learning. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 2159–2169.
25. Li, Q.; Mao, Y. A review of boosting methods for imbalanced data classification. *Pattern Anal. Appl.* **2014**, *17*, 679–693. [CrossRef]
26. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. In Proceedings of the International Conference on Learning Representations ICLR 2013, Scottsdale, AZ, USA, 2–4 May 2013.
27. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
28. Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep Contextualized Word Representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 2227–2237. [CrossRef]
29. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. 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, Volume 1 (Long and Short Papers), Minneapolis, MI, USA, 2–7 June 2019; pp. 4171–4186. [CrossRef]
30. Chen, Y.; Mani, S.; Xu, H. Applying active learning to assertion classification of concepts in clinical text. *J. Biomed. Inform.* **2012**, *45*, 265–272. [CrossRef]
31. Hahn, U.; Beisswanger, E.; Buyko, E.; Faessler, E. Active Learning-Based Corpus Annotation—The PathoJen Experience. In Proceedings of the AMIA 2012, American Medical Informatics Association Annual Symposium, Chicago, IL, USA, 3–7 November 2012.
32. Han, X.; Kwoh, C.K.; Kim, J. Clustering based active learning for biomedical Named Entity Recognition. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 1253–1260. [CrossRef]
33. Tao, J.; Brayton, K.A.; Broschat, S.L. Automated Confirmation of Protein Annotation Using NLP and the UniProtKB Database. *Appl. Sci.* **2021**, *11*, 24. [CrossRef]
34. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2019**. btz682, [CrossRef] [PubMed]
35. Alves-Pinto, A.; Demus, C.; Spranger, M.; Labudde, D.; Hobley, E. Iterative Named Entity Recognition with Conditional Random Fields. *Appl. Sci.* **2022**, *12*, 330. [CrossRef]
36. Gabbard, R.; DeYoung, J.; Lignos, C.; Freedman, M.; Weischedel, R. Combining rule-based and statistical mechanisms for low-resource named entity recognition. *Mach. Transl.* **2018**, *32*, 31–43. [CrossRef]
37. Kanterakis, A.; Kanakaris, N.; Koutoulakis, M.; Pitianou, K.; Karacapilidis, N.; Koumakis, L.; Potamias, G. Converting Biomedical Text Annotated Resources into FAIR Research Objects with an Open Science Platform. *Appl. Sci.* **2021**, *11*, 9648. [CrossRef]
38. Wang, Y.; Sohn, S.; Liu, S.; Shen, F.; Wang, L.; Atkinson, E.J.; Amin, S.; Liu, H. A clinical text classification paradigm using weak supervision and deep representation. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 1. [CrossRef] [PubMed]
39. Al-Laith, A.; Shahbaz, M.; Alaskar, H.F.; Rehmat, A. AraSenCorpus: A Semi-Supervised Approach for Sentiment Annotation of a Large Arabic Text Corpus. *Appl. Sci.* **2021**, *11*, 2434. [CrossRef]
40. Banerjee, I.; Li, K.; Seneviratne, M.; Ferrari, M.; Seto, T.; Brooks, J.D.; Rubin, D.L.; Hernandez-Boussard, T. Weakly supervised natural language processing for assessing patient-centered outcome following prostate cancer treatment. *JAMIA Open* **2019**. [CrossRef] [PubMed]
41. Attardi, G.; Cozza, V.; Sartiano, D. Annotation and Extraction of Relations from Italian Medical Records. In Proceedings of the 6th Italian Information Retrieval Workshop, Cagliari, Italy, 25–26 May 2015.
42. Lerner, I.; Paris, N.; Tannier, X. Terminologies augmented recurrent neural network model for clinical named entity recognition. *J. Biomed. Inform.* **2020**, *102*, 103356. [CrossRef]
43. Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; Dyer, C. 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, San Diego, CA, USA, 12–17 June 2016; pp. 260–270. [CrossRef]
44. Silvestri, S.; Gargiulo, F.; Ciampi, M. Improving Biomedical Information Extraction with Word Embeddings Trained on Closed-Domain Corpora. In Proceedings of the 2019 IEEE Symposium on Computers and Communications (ISCC), Barcelona, Spain, 29 June–3 July 2019; pp. 1129–1134. [CrossRef]
45. Alicante, A.; Corazza, A.; Isgrò, F.; Silvestri, S. Semantic Cluster Labeling for Medical Relations. In Proceedings of the third International Conference Innovation in Medicine and Healthcare 2016, Puerto de la Cruz, Spain, 15–17 June 2016; pp. 183–193. doi: [CrossRef]

46. Kameswara Sarma, P.; Liang, Y.; Sethares, B. Domain Adapted Word Embeddings for Improved Sentiment Classification. In Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP, Melbourne, Australia, 15–20 July 2018; pp. 51–59.
47. Jin, Q.; Dhingra, B.; Cohen, W.; Lu, X. Probing Biomedical Embeddings from Language Models. In Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP, Minneapolis, MN, USA, 2–7 June 2019; pp. 82–89. [CrossRef]
48. Berardi, G.; Esuli, A.; Marcheggiani, D. Word Embeddings Go to Italy: A Comparison of Models and Training Datasets. In Proceedings of the 6th Italian Information Retrieval Workshop, Cagliari, Italy, 25–26 May 2015.
49. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the Annual 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
50. Buda, M.; Maki, A.; Mazurowski, M.A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* **2018**, *106*, 249–259. [CrossRef]
51. Han, W.; Huang, Z.; Li, S.; Jia, Y. Distribution-Sensitive Unbalanced Data Oversampling Method for Medical Diagnosis. *J. Med. Syst.* **2019**, *43*, 39:1–39:10. [CrossRef]
52. Bodenreider, O. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Res.* **2004**, *32*, D267–D270. [CrossRef]
53. Tjong, E.F.; Sang, K.; Veenstra, J. Representing Text Chunks. In Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics, Bergen, Norway, 8–12 June 1999.
54. Wang, X.; Zhang, Y.; Ren, X.; Zhang, Y.; Zitnik, M.; Shang, J.; Langlotz, C.; Han, J. Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics* **2019**, *35*, 1745–1752. [CrossRef]
55. Wang, Y.; Liu, F.; Verspoor, K.; Baldwin, T. Evaluating the Utility of Model Configurations and Data Augmentation on Clinical Semantic Textual Similarity. In Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, Online, 9 July 2020; pp. 105–111. [CrossRef]
56. Islam, S.; Papastergiou, S.; Silvestri, S. Cyber Threat Analysis Using Natural Language Processing for a Secure Healthcare System. In Proceedings of the 27th IEEE Symposium on Computers and Communications (ISCC 2022), Rhodes Island, Greece, 29 June–3 July 2022, to be published.