Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

Tecnica per l'Annotazione di un Corpus per l'Addestramento di un Sistema Deep Learning per Biomedical Named Entity Recognition

Francesco Gargiulo, Stefano Silvestri, Mario Ciampi

RT-ICAR-NA-2019-03 ***Data:*** **Maggio 2019**

Consiglio Nazionale delle Ricerche, Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR) – Sede di Napoli, Via P. Castellino 111, I-80131 Napoli, Tel: +39-0816139508, Fax: +39-0816139531, e-mail: napoli@icar.cnr.it, URL: *www.na.icar.cnr.it*

Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

Tecnica per l'Annotazione di un Corpus per l'Addestramento di un Sistema Deep Learning per Biomedical Named Entity Recognition

Francesco Gargiulo, Stefano Silvestri, Mario Ciampi

Rapporto Tecnico N: **RT-ICAR-NA-2019-03** ***Data:*** **Maggio 2019**

Tecnica per l'Annotazione Assistita da Intelligenza Artificiale di un Corpus per l'Addestramento di un Sistema Deep Learning per Biomedical Named Entity Recognition

Francesco Gargiulo, Stefano Silvestri, Mario Ciampi
Istituto di Calcolo e Reti ad Alte Prestazioni del Consiglio Nazionale delle Ricerche (ICAR-CNR)
Via Pietro Castellino, 111 – 80131 Napoli, Italia
{francesco.gargiulo, stefano.silvestri, mario.ciampi}@icar.cnr.it

Abstract

In questo technical report è descritta una tecnica atta a facilitare l'annotazione di un corpus specifico per il training di sistemi di intelligenza artificiale supervisionati per l'estrazione di entità di dominio (Named Entity Recognition - NER) da testi non strutturati e scritti in linguaggio naturale. Tale tecnica è basata sull'utilizzo congiunto di una rete neurale di tipo Deep Learning, di Word Embeddings di dominio, di strumenti NLP e di basi di conoscenza di dominio ed ha lo scopo di facilitare e supportare il compito di un esperto nel complicato e lungo task di annotazione di un corpus, riducendone i tempi e l'effort richiesti. La tecnica proposta è stata applicata nel caso del dominio biomedicale e, in particolare, su cartelle cliniche elettroniche (Electronic Health Records - EHR). In tale ambito applicativo, infatti, la necessità di sistemi intelligenti per l'estrazione di informazioni si scontra con la mancanza di corpora annotati, soprattutto in lingue differenti dall'inglese, richiedendo, quindi, lo sviluppo di metodologie del tipo di quella proposta di seguito. I risultati ottenuti mostrano l'utilità della metodologia descritta, la quale può essere anche facilmente declinata in ulteriori domini, oltre che in differenti lingue.

Keywords: Named Entity Recognition, Deep Learning, Annotated Biomedical Corpus, Italian Language, Class Imbalance

1. Introduzione

Recentemente, visto il successo di metodologie supervised per l'addestramento di sistemi supervisionati di Intelligenza Artificiale (IA) per applicazioni nel campo del Natural Language Processing (NLP) e, in particolare, di sistemi per l'estrazione di entità di dominio (Named Entity Recognition - NER) e di relazioni (Relation Extraction), è cresciuta la necessità di avere disponibili training set specifici per questi determinati task. Un training set per sistemi supervisionati di intelligenza artificiale (chiamato in letteratura anche *corpus*) per NLP è costituito un insieme di documenti in linguaggio naturale dotati delle delle rispettive etichette, o annotazioni, la cui estensione sia sufficiente per addestrare un sistema di IA con il necessario livello di precisione. Sebbene siano disponibili alcuni dataset etichettati in lingua inglese, di contro è difficile reperire training set in linguaggi differenti, come ad esempio l'italiano, proprio a causa della difficoltà, del tempo e del costo di un processo di etichettatura. Tale mancanza di corpus richiede da tempo soluzioni efficienti da parte della comunità scientifica [26]. Anche nel caso in cui il contenuto dei documenti oggetto di NER appartenga a domini specifici, come ad esempio il caso del dominio biomedicale [1] (il task dell'estrazione di entità appartenenti al tale dominio prende il nome di *B-NER - Biomedical NER*) è forse ancora più sentita la necessità di disporre di documenti specifici annotati, data la sostanziale differenza del linguaggio e dei termini utilizzati all'interno di testi di questo tipo, in cui un sistema addestrato su testo di tipo generico otterrebbe prestazioni molto deludenti [38], [51].

Il processo di annotazione, detto anche di etichettatura, è un processo lungo, laborioso e tedioso, che richiede la collaborazione con esperti di dominio, quali ad esempio medici nel caso dell'annotazione di documenti appartenenti nel campo biomedicale. Per questo, la realizzazione di corpus annotati si scontra anche con la difficoltà di reperire esperti di dominio disponibili a fornire la propria professionalità, dato l'impegno di tempo e di risorse richieste. Allo scopo di sopperire alla mancanza di training set e di facilitare il task di annotazione di nuovi corpora supportando il lavoro di esperti che possono dedicarsi a questo tipo di attività, sono state sviluppate alcune metodologie automatiche, o semi-automatiche per l'annotazione.

Focalizzandosi ai casi specifici di corpora per task di B-NER, in [2] sono stati ottenuti training set per NER formati da cartelle cliniche in lingua italiana a partire da documenti in lingua inglese pre-etichettati. Per ottenere dataset omologhi a quelli in inglese, ma in italiano, sono stati utilizzati strumenti di traduzione automatica, supportati da dizionari/thesauri specifici multilingua, come il Metathesaurus di UMLS[1], utilizzando questi ultimi con lo scopo di migliorare la precisione della traduzione dei termini specifici del dominio biomedicale. Sebbene il risultato finale di tale processo abbia portato ad ottenere un gold standard funzionale per scopi di ricerca, l'uso delle stesse metodologie applicato a sistemi in grado di analizzare reali cartelle cliniche, o referti medici non produrrebbe risultati soddisfacenti a causa di numerosi problemi, causati da un lato dai limiti dei traduttori automatici e dall'altro lato dal contenuto stesso di tale classe di documenti. Infatti, l'ampio uso che il personale medico addetto alla compilazione di cartelle cliniche fa di acronimi, abbreviazioni e terminologia non standard (e quindi non presenti nei dizionari), la presenza di un linguaggio prettamente di dominio (con locuzioni e costrutti specifici), la diffusa presenza di errori di scrittura, oltre alla variabilità e complessità delle entità stesse, rendono necessario adottare un differente approccio, che permetta di ottenere un training set per la realizzazione di sistemi di intelligenza artificiale per il B- NER in maggiormente specializzato sul tipo di documenti che il sistema di IA dovrà essere in grado successivamente di analizzare.

In letteratura si è cercato di ovviare alla mancanza di corpus per Machine Learning annotati in italiano (o in altre lingue differenti dall'inglese) in maniera alternativa, utilizzando ad esempio metodologie per il B-NER basate su regole predeterminate e basi di conoscenza [7], [8], oppure sfruttando sistemi di machine learning di tipo unsupervised, quali il clustering [9], [10], [11], con risultati variabili. Infatti, nel primo caso le prestazioni sono limitate dalla difficoltà di generalizzazione di regole fisse, mentre nel secondo caso, l'uso di un addestramento non supervisionato rende complicata la successiva fase di classificazione dei risultati ottenuti. Inoltre, i risultati più recenti mostrano che le reti neurali di tipo Deep Learning (DL) sono in grado di raggiungere performance allo stato dell'arte nel task di etichettatura automatica di entità [3] e sono state applicate anche con successo nel caso specifico considerato, ossia a cartelle cliniche e a narrativa di tipo medico [4], [5], [6] per la soluzione di problemi di tipo B-NER, migliorando ulteriormente i risultati raggiunti dai precedenti sistemi allo stato dell'arte, basati su machine learning e Conditional Random Fields (CRF), come ad esempio quello descritto in [12]. Per questi motivi, diviene di fondamentale importanza disporre di corpora annotati in differenti linguaggi e appartenenti a domini specifici, al fine di poter addestrare reti neurali di tipo DL con l'adeguato livello di precisione, riuscendo così a sfruttarne i risultati per la realizzazione di sistemi cognitivi per il supporto di professionisti, come è avvenuto per esempio nel caso di *IBM Watson Health*[2].

Nel seguito del presente rapporto tecnico sarà descritta la tecnica semi-automatica proposta, attraverso la quale è possibile applicare l'etichettatura delle entità al testo di un dataset formato da reali referti e cartelle cliniche anonimizzati in italiano. Il metodo richiede un effort da parte di esperti di dominio molto ridotto, soprattutto se comparato con l'approccio di annotazione manuale classico e, quindi, è in grado di permettere l'annotazione di training set per il NER di documenti clinici in italiano in maniera molto rapida e semplice. L'efficacia del metodo e dei training set ottenuti grazie alla sua applicazione sono stati valutati attraverso un sistema Deep Learning (DL) per NER in dominio biomedicale, testato su cartelle cliniche (Electronic Health Records - EHRs) in italiano [49], [50], tramite l'uso del corpus prodotto come training set, dimostrandone l'efficacia nell'addestramento del sistema di IA basato su un'architettura di rete DL che segue quanto

[1] https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/
[2] https://www.ibm.com/watson/health/

proposto in [13].

2. Annotazione di documenti per il NER

Un'entità di dominio consiste in una parola o in una locuzione (detta anche *multiword expression*) appartenenti ad una classe specifica del dominio considerato. Ad esempio, nel dominio biomedicale possono essere considerate named entity i termini *insufficienza renale cronica*, *antibiotico*, *pronto soccorso*, etc.. Annotare un documento perché possa essere utilizzato come parte di un training set per un sistema automatico di apprendimento supervisionato per il task di NER consiste nell'etichettare ogni parola del testo, indicando se faccia parte di una entità e di che tipo sia quella determinata entità. In applicazioni di B-NER, tipici esempi di classi di entità possono essere *farmaco*, *malattia*, *parte del corpo*, etc. Nel caso in esame, visto anche il tipo e il contenuto dei documenti da analizzare, sono stati individuati i seguenti tipi di entità, mostrati in Tabella 1.

Tipo di Entità	**Sigla**	**Esempi**
Malattie e Sintomi	DIS	*Febbre , pressione alta, cirrosi epatica, frattura*
Misure	MEA	*30 cc, 12 mm, 120 bpm*
Date	DAT	*15/12/2010, lunedì, ore 20:55, 3 gennaio*
Dipartimenti	DEP	*Pronto Soccorso, Ortopedia, Chirurgia Generale*
Analisi di laboratorio, esami ed indagini cliniche	ANA	*Creatinina, rx, ecografia, glicemia*
Farmaci	DRU	*Antibiotico, voltaren retard, acido acetilsalicilico*
Parti del corpo	BOD	*Stomaco, piede destro, quinta vertebra, testa dell'omero*
Terapie e strumenti medici	THE	*Ecografo, appendicectomia, bypass, profilassi, protesi*

Tabella 1 - Tipi di entità ed esempi

L'annotazione delle entità può essere eseguita seguendo alcune codifiche standard: tra di esse, nel caso in esame è stata scelta la cosiddetta notazione *IOB* (*Inside Outside Begin*) [14], in cui il prefisso B sta per *Begin* (prima parola di un'entità), I sta per *Inside* (etichetta tutti i successivi token successivi al primo etichettato con B facenti parte della stessa entità) e O sta per *Outside* (token non appartenente ad una entità). Più precisamente, ricordando che che una entità di dominio può essere formata sia da una singola parola (es. cefalea), ma anche da un gruppo di parole consecutive (es. diabete mellito di tipo 2), è necessaria una notazione capace di identificare correttamente anche le entità formate da più parole: la notazione IOB risolve proprio questa necessità. Di seguito, in Figura 1, si riporta un esempio di testo annotato seguendo le regole sopra descritte.

Lo scopo della tecnica proposta e descritta nel presente rapporto tecnico è quello di ottenere un insieme sufficientemente esteso di documenti annotati come mostrato nella Figura 1, a partire da documenti clinici di tipo narrativo scritti in linguaggio naturale, come EHR. L'innovatività della tecnica proposta consiste nell'essere in grado di ridurre al minimo l'applicazione che l'utente esperto deve dedicare all'etichettatura

manuale e, al contempo, di mitigare alcune problematiche tipiche dei training set, quali il *class imbalancing* (lo sbilanciamento del numero di esempi di una determinata classe rispetto ad un altra) e il label noise (il rumore, ossia la variabilità, dell'etichettatura di elementi uguali, dovuta ad errori) e il riconoscimento di entità *out-of-corpus*, ossia entità non viste dal sistema IA in fase di training. In sintesi, gli obiettivi principali che tale metodologia vuole raggiungere sono:

- Ottenere un modello di classificazione il più affidabile possibile;
- Risolvere il problema delle classi sbilanciate (*class imbalancing*);
- Superare i limiti delle architetture IA in fase di testing su entità *out of corpus*;
- Ridurre al minimo il lavoro umano necessario per l'etichettatura del dato.

Di seguito saranno descritte nel dettaglio le caratteristiche del dataset iniziale che sarà oggetto di annotazione ed il procedimento proposto, mostrando anche i tool e gli strumenti di supporto necessari a riprodurre i risultati e ad applicarli a differenti contesti e/o lingue.

```
[...]
in  O
data    O
14  B-DAT
/   I-DAT
02  I-DAT
/   I-DAT
2013    I-DAT
in  O
seguito O
a   O
caduta  O
accidentale O
riscontro   O
di  O
frattura    B-DIS
amielica    I-DIS
di  O
l1  B-BOD
(   O
sottoposto  O
a   O
artrodesi   B-THE
strumentata I-THE
t12 B-BOD
-   O
l1  B-BOD
-   O
l2  B-BOD
tramite O
5   B-MEA
viti    B-THE
,   O
2   B-MEA
barre   B-THE
di  I-THE
titanio I-THE
e   I-THE
fosfato I-THE
silicato    I-THE
di  I-THE
calcio  I-THE
[...]
```

Figura 1 - Esempio di testo annotato

3. Tecnica di annotazione assistita da intelligenza artificiale: Active Learning

Dato lo scopo di realizzare un corpus annotato per il training di sistemi intelligenti specifici per task B-NER in italiano, il dataset originale da annotare considerato è composto dalla parte narrativa in linguaggio naturale in italiano estratta da un corpus di 989 cartelle cliniche elettroniche (EHR). Il corpus in esame conta in tutto un totale di 1.355.867 parole. Per permettere un addestramento corretto e, allo stesso tempo, elevate performance in fase di testing di una rete neurale DL sono necessari corpora di elevate dimensioni (ad esempio il dataset per il training di sistemi NER di dominio generico relativo allo shared task per la lingua inglese CONLL-2003 [26] contiene 203.621, parole distribuite in 14.987 frasi) e, pertanto, uno o più esperti di dominio dovrebbero analizzare con cura le singole frasi relative a oltre 200.000 parole, individuando, classificando ed annotando manualmente tutte le entità di dominio presenti. Tale approccio, chiaramente, richiede molto tempo e lavoro da parte di esperti di dominio e di informatici e, quindi, ha anche un costo elevato: a causa di ciò, vi è ancora oggi una scarsità di corpus annotati, soprattutto in lingue differenti dall'inglese. La tecnica proposta e testata di seguito ha lo scopo di sfruttare una rete DL per semplificare e velocizzare il processo di annotazione, supportando l'esperto in questo gravoso e tedioso compito.

Per potere ridurre in maniera sensibile l'apporto manuale necessario per l'annotazione di un corpus di dimensioni sufficienti, è stata definita una procedura ricorsiva semiautomatica basata su *active learning* [48], il cui schema è illustrato nella seguente Figura 2. Tale tipo di approccio sfrutta il supporto di una rete neurale DL, allo scopo di facilitare il lavoro dell'esperto nell'annotazione. In particolare, si basa sul supporto di sistemi di Machine Learning (ML), i quali permettono di ottenere una annotazione parziale in maniera automatica, rendendo il lavoro manuale dell'esperto più focalizzato alla revisione che alla annotazione attiva. Approcci basati su active learning sono stati già utilizzati con successo in vari campi, come ad esempio il supporto degli esperti di PubMed per l'annotazione dei documenti indicizzati tramite la codifica delle MeSH (Medical Subject Headings) attraverso sistemi intelligenti basati su ML [27], [28], [29], [30], [31].

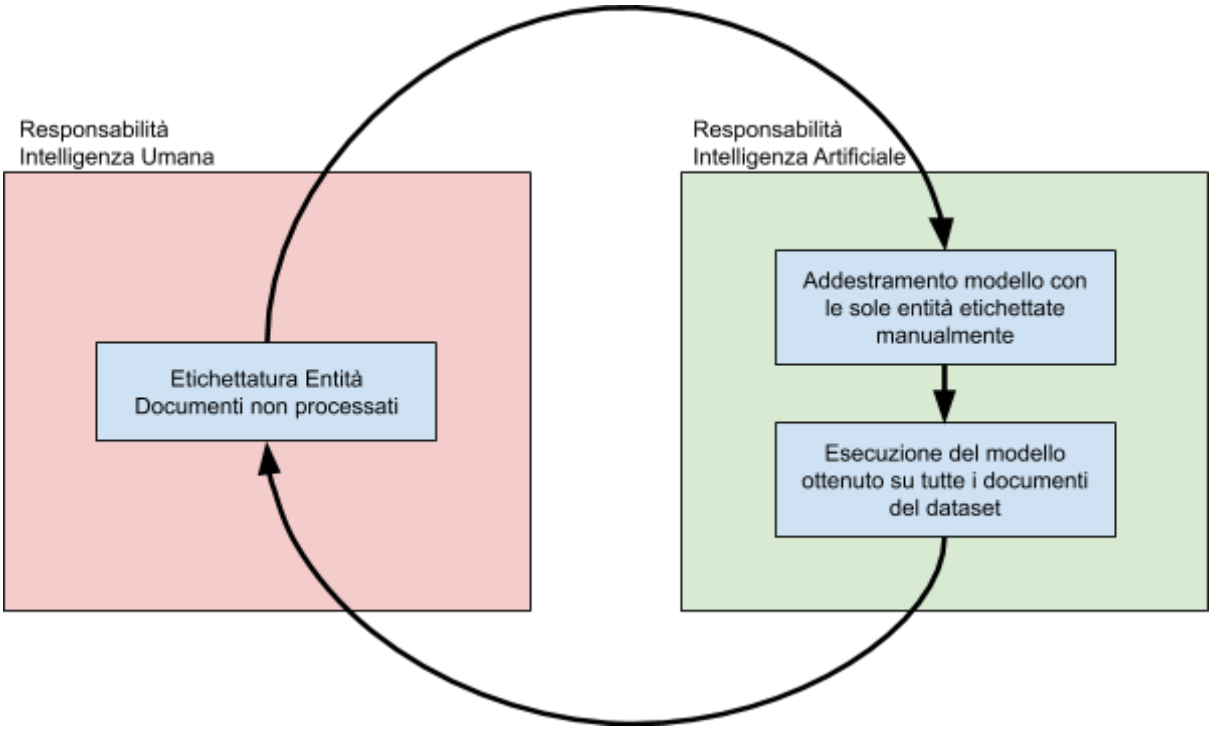

Figura 2 - Rappresentazione schematica del processo semi-automatico di annotazione

Di seguito è spiegata nel dettaglio la metodologia basata su active learning proposta. Il processo di annotazione inizia con l'etichettatura manuale esclusivamente di un numero estremamente esiguo di documenti: nel caso in esame sono stati etichettate 10 EHR estratte casualmente dal corpus, che contano un totale di 12.632 parole. Tale operazione non richiede certamente lo sforzo dell'annotazione di un intero dataset, di dimensioni almeno 20 volte maggiori di quello considerato inizialmente. Questo primo piccolo insieme di dati deve essere utilizzato per addestrare la rete DL descritta nel successivo Paragrafo 4. Nel caso in esame il sistema DL usato negli esperimenti dopo 50 epoche è riuscito a raggiungere una precisione del 99,9% delle etichette sullo stesso insieme di dati usato per l'addestramento.

Quindi, come mostrato in Figura 2, il modello DL così ottenuto va utilizzato per annotare automaticamente tutti restanti documenti del corpus, sfruttando così l'intelligenza artificiale per una operazione di *pre-etichettatura* dell'intero dataset. I risultati ottenuti sull'intero dataset dopo questo primo step non sono di certo eccellenti, a causa del limitato training set, ma, di contro, permettono di ottenere un corpus con una serie di pre-annotazioni, sebbene con una precisione non elevata (la Precision sul test set, come mostrato nel successivo paragrafo 7, è dell'ordine del 60%). Seguendo l'approccio iterativo illustrato in Figura 2, nel passo successivo vengono estratti casualmente ulteriori 20 documenti dal corpus (un numero maggiore rispetto al primo step), i quali, in questo caso, contengono già la pre-etichettatura iniziale. L'esperto, quindi, nel secondo step avrà solamente il compito di verificare la correttezza delle etichette fornite dal sistema DL e, in questo modo, si riduce il tempo per completare l'annotazione dei nuovi documenti, all'interno dei quali è presente già una certa percentuale di entità correttamente individuate dal sistema automatico. I nuovi documenti annotati, uniti a quelli relativi al primo step, vengono nuovamente utilizzati come training corpus per riaddestrare la rete neurale, la quale, sarà poi usate per etichettare di nuovo tutti i documenti restanti del corpus. In questo secondo caso, disponendo di un numero maggiore di esempi in fase di training, i risultati della nuova pre-etichettatura automatica operata dal sistema DL saranno sicuramente migliorati, permettendo così nello step successivo un ancora più rapido lavoro di revisione da parte dell'esperto, grazie alla presenza di una maggiore percentuale di entità già correttamente annotate, frutto della migliorata precisione del sistema neurale che ha visto in fase di training un maggior numero di esempi. Il processo viene quindi ripetuto iterativamente nello stesso modo, fino a che non si ottiene un corpus di dimensioni sufficienti e, parimenti, non si osservano ulteriori miglioramenti dei risultati in fase di testing del modello. Grazie al continuo miglioramento del modello DL ottenuto, inoltre, si riduce progressivamente l'effort manuale per la correzione dei nuovi documenti analizzati in ogni successivo step, incrementando il numero di elementi che l'esperto riesce ad analizzare a parità di tempo. Il processo iterativo sopra descritto basato su active learning viene terminato quando la performance del modello DL B-NER non mostrano più sostanziali miglioramenti.

Un corpus così ottenuto, sebbene abbia richiesto meno tempo al gruppo di esperti per l'annotazione manuale, grazie al supporto dell'intelligenza artificiale che, inoltre, migliora la precisione dei risultati prodotti ad ogni step, potrebbe essere, di contro, afflitto da alcune problematiche, come quelle dovute al *class imbalance* [15]. La scelta casuale dei documenti da annotare dal corpus iniziale di cartelle cliniche non annotate, nonché il contenuto stesso dei documenti, possono comportare un forte sbilanciamento degli esempi di determinate classi rispetto ad altre. Questo fenomeno, detto appunto class imbalance, può causare a sua volta un degrado delle prestazioni della rete neurale NER se addestrata con corpora detti "sbilanciati" [44], [45]. In letteratura alcuni metodi per la soluzione di questo problema sono stati proposti [16], [17], [18], [43] , utilizzando oversampling o undersampling di campioni, oppure metodi più sofisticati basati su machine learning. Nel caso in esame, la presenza di basi di conoscenza di dominio permette di mitigare gli effetti negativi del class imbalance sulla rete neurale per il B-NER. Infatti, successivamente alla fase di active learning sopra descritta, la tecnica proposta consiste nell'applicare anche una seconda fase basata di annotazione, in questo caso completamente automatizzata e basata su *distant supervision* [47]. La tecnica di distant supervision permette di annotare un corpus in maniera totalmente automatica, sfruttando basi di conoscenza di dominio preesistenti, come dizionari, thesauri ed altro. La metodologia, descritta nel dettaglio nel successivo paragrafo 5, permette in questo modo di incrementare gli esempi relativi alle classi maggiormente sbilanciate, ossia con meno campioni, espandendo il corpus seguendo il seguente approccio. Tutte le frasi contenenti almeno un'entità appartenente alle classi sbilanciate vengono estratte dal corpus;

successivamente, si generano nuove frasi, sostituendo all'interno delle frasi precedentemente selezionate le entità delle suddette classi con una nuova e dello stesso tipo, estratta a sua volta casualmente dalle rispettive KB. Questo processo viene iterato finché non si ottiene un numero sufficiente di nuove frasi e, parimenti, sono state usate tutte le entità delle KB, in modo da fornire anche esempi di entità eventualmente non presenti originariamente nel corpus, ponendo rimedio anche al problema delle entità *out-of-corpus*, ossia non viste dalla IA in fase di training. Infine, le nuove frasi generate secondo questo approccio, si inseriscono nuovamente nel corpus, in maniera casuale. Si fa notare che questa tecnica di espansione sovracampiona anche le entità appartenenti a classi non sbilanciate che sono presenti nelle frasi utilizzate per l'espansione, fornendo così ulteriori esempi anche per altre classi. In questo modo si vanno a migliorare ulteriormente le prestazioni complessive, come dimostrano i risultati mostrati nel successivo paragrafo 7.

4. Modello di rete neurale Deep Learning per il NER

La rete neurale utilizzata per applicare la pre-annotazione dei documenti come supporto all'esperto, nonché per testare la qualità del corpus annotato definitivo ottenuto applicando in toto la tecnica proposta è basata sulla baseline allo stato dell'arte descritta in [13], la quale è stata opportunamente modificata per poter essere applicata al problema in esame.

La seguente Figura 3 mostra l'architettura della rete DL per il NER. Nello schema in Figura vanno distinti tre differenti livelli. Il primo di questi, relativo all'ingresso della rete (parte in basso), ha il compito di fornire la rappresentazione delle parole in ingresso. Più nel dettaglio, le parole delle frasi in ingresso vengono rappresentate attraverso character embeddings e Word Embeddings (WEs), i quali possono essere pre-addestrati (nel successivo Paragrafo 4.1 sono mostrati i dettagli dei modelli sperimentati). Tale rappresentazione è ottenuta dalla concatenazione di character-based word embeddings [33], ottenuti attraverso una rete Bi-LSTM (Bidirectional Long Short Term Memory) direttamente dal testo utilizzato come training set e di Word Embeddings (WEs) pre-addestrati su corpus differenti. Nel caso in esame, come spiegato nel successivo Paragrafo 4.1, sono stati sperimentati differenti corpus e differenti modelli di WEs (word2vec [22], [23] e fastText [24], [25]).

Il livello successivo ha il compito di estrarre la rappresentazione delle parole all'interno delle frasi, attraverso l'uso anche in questo caso di una Bi-LSTM, i cui output delle rappresentazioni forward e backward vengono concatenati, ottenendo in questo modo una rappresentazione efficace delle parole in un contesto [34]. L'ultimo layer ha il compito della classificazione delle parole, assegnando il corretto tag relativo alla NER. In questo caso, al posto di invece di modellare le decisioni di tagging in maniera indipendente, vengono modellate congiuntamente, utilizzando un layer Conditional Random Field [35]. Attraverso l'uso della CRF in uscita, si massimizza la probabilità che la rete produca una sequenza valida di etichette di uscita, minimizzando ad esempio la probabilità che ad una etichetta B-DRU possa succedere una etichetta di tipo I-DEP, che rappresenterebbe una sequenza errata.

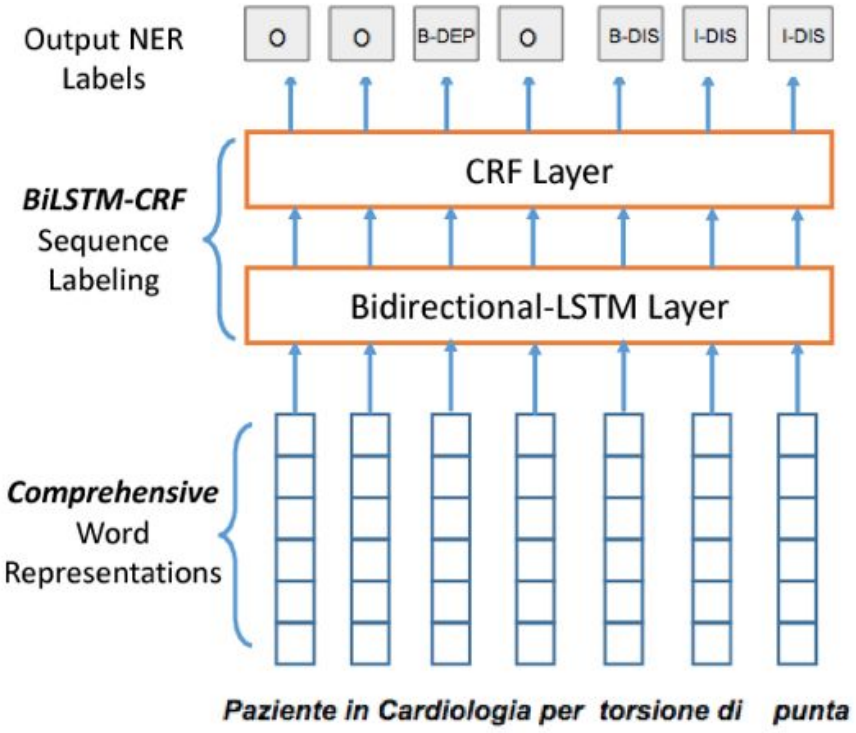

Figura 3 - Architettura delle rete DL utilizzata per il supporto al tagging del corpus

La rete DL utilizzata per poter essere addestrata prende in input un corpus annotato con il NER e un modello di embedding pre-addestrato e, quando usata in fase di testing, produce per ogni parola la corrispondente etichetta secondo la notazione IOB. Per questo, sono stati anche sperimentati differenti modelli WEs, allo scopo di incrementare ulteriormente le prestazioni della tecnica proposta.

4.1. Modelli di embeddings

Come spiegato sopra, la rete neurale DL utilizzata necessita anche di un modello di WEs pre-addestrato in input. Negli esperimenti eseguiti, sono stati testati sia modelli WEs in italiano preaddestrati su testo generico estratto da Wikipedia disponibili in rete, sia modelli custom addestrati per il task in esame su testo di dominio. I dettagli dei modelli di embeddings considerati e del corpus utilizzato per il training sono descritti nel seguito di questo Paragrafo, mentre la seguente Tabella 2 riassume i parametri di training di ogni singolo modello sperimentato.

I primi esperimenti sono stati eseguiti utilizzando il modello di word2vec in italiano fornito dall'ISTI-CNR [37] e disponibile pubblicamente[3]. Tale modello è stato addestrato su di un dump di Wikipedia in italiano e su testo estratto da libri in italiano, senza l'applicazione di preprocessing NLP e, pertanto, il dizionario ottenuto contiene rumore, causato dai numerosi termini presenti, oltre a non annoverare numerosi token specifici di dominio, non presenti nel testo utilizzato per il training. Come visibile in Tabella 2, il modello è stato addestrato con l'algoritmo skip-gram [23] e considerando una window size pari a 10 e un negative samples pari a 10, producendo un modello la cui vector size è pari a 300.

[3] http://hlt.isti.cnr.it/wordembeddings/

Allo scopo di disporre di un modello di embeddings specifico per il task in esame, sono stati successivamente addestrati ulteriori modelli di WEs utilizzando esclusivamente testo in linguaggio naturale di dominio biomedico. Per questo motivo. è stato prodotto un ulteriore corpus non annotato contenente esclusivamente testo di dominio biomedicale, per permettere l'addestramento dei modelli WEs specifici. Nel dettaglio, il corpus per il training di questi ulteriori WEs è stato ottenuto dalle seguenti fonti:

1. Dump di Wikipedia in italiano, selezionando solamente le pagine relative alla categoria "Scienze della Salute". Seguendo la pipeline definita in [10], per ottenere questo dump selettivo è stato utilizzato il tool di Wikipedia PetScan[4], attraverso il quale è possibile ottenere una lista delle pagine di Wikipedia appartenenti ad una o più categorie, oltre alle pagine le cui categorie sono correlate. Wikipedia è categorizzato in un grafo ciclico e PetScan permette di selezionare la profondità di sottocategorie da includere nella lista richiesta in output. Nel caso in esame, è stata scelta una profondità pari a 6, allo scopo non introdurre troppo rumore (ossia pagine il cui argomento fosse molto scorrelato dal dominio scelto). Una volta ottenuta la lista di pagine il cui topic è correlato a quello biomedicale, è possibile ottenere un dump di Wikipedia contenente soltanto le pagine scelte attraverso lo strumento Export di Wikipedia (disponibile in italiano[5], inglese[6], oltre che per tutte le altre lingue in cui è disponibile Wikipedia). Il file ottenuto da tale processo è in formato standard xml; per estrarre da quest'ultimo soltanto il testo delle corrispondenti pagine Wikipedia è stato utilizzato il tool WikiExtractor[7], sviluppato dal Medialab dell'Università di Pisa.
2. Testo dei bugiardini dei medicinali, ossia il contenuto del cosiddetto "foglietto illustrativo". I bugiardini sono ampiamente disponibili online, anche, per esempio, sul sito ufficiale dell'Agenzia Italiana del Farmaco (AIFA)[8]. Nel dettaglio, sono disponibili per tutti i farmaci in commercio sia il testo sia dei foglietti illustrativi, sia il riassunto delle caratteristiche del prodotto, in formato pdf. Attraverso un software di web-scraping realizzato in Python specificamente per il sito in esame, è stato possibile ottenere tutti i file in formato pdf; successivamente, il testo contenuto nel file pdf è stato estratto e normalizzato attraverso le librerie in Python Apache Tika[9], insieme a script Python specifici per il tipo di documenti in esame.
3. Dizionario della Salute online in italiano del Corriere della Sera[10]. Così come altri dizionari disponibili sul web, in questo dizionario fornito dal Corriere della Sera sono presenti termini medici corredati dalla corrispondente descrizione. Il testo del dizionario è stato ottenuto, così come nel caso precedente, attraverso sistemi di web-scraping custom, scritti in linguaggio Python.
4. Documenti di dominio biomedico ad accesso open in rete, come materiale universitario, presentazioni, documentazione tecnica, pubblicazioni open access. Tali documenti sono disponibili in diversi formati, come pdf, docx, ppt, rtf e, per questo, è stato necessario estrarre e normalizzare il testo utilizzando, come nel caso dei bugiardini, una pipeline specifica basata sulle librerie Apache Tika.

Il testo ottenuto dalle fonti sopra elencate è stato normalizzato dopo la fase di estrazione, eliminando i caratteri non standard, oltre che eventuali intestazioni, numeri di pagina e così via. Successivamente, attraverso TextPro [19], uno strumento NLP allo stato dell'arte per la lingua italiana sviluppato dall'Istituto di Ricerca FBK di Trento, il testo è stato oggetto di tokenizzazione e sentence splitting, ottenendo le singole parole che compongono il testo, suddivise per frase. In ultimis, allo scopo di ridurre ulteriormente il rumore presente nel modello di word embeddings, tutte le parole sono state convertite in minuscolo. La lista delle frasi così ottenute (un totale di 2.160.704 sentences per un totale di 511.649.310 parole) è stata utilizzata per

[4] https://petscan.wmflabs.org/
[5] https://it.wikipedia.org/wiki/Speciale:Esporta
[6] https://en.wikipedia.org/wiki/Special:Export
[7] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor
[8] https://farmaci.agenziafarmaco.gov.it/bancadatifarmaci/cerca-farmaco
[9] https://tika.apache.org/0.7/parser.html
[10] https://www.corriere.it/salute/dizionario/?fr=correlati

addestrare i modelli WEs. Tale corpus è stato utilizzato per addestrare quattro differenti modelli word2vec, utilizzando rispettivamente sia l'algoritmo skip-gram che il cbow (continuous bag of words) [22], [23], insieme al metodo negative sampling e rispettivamente attivando e disattivando il metodo hierarchical softmax, come visibile in Tabella 2.

Inoltre, è stato considerato un ulteriore modello di embeddings per la rappresentazione del testo in sistemi Deep Learning, ossia FastText [24], [25]. Tale modello aggiunge in fase di apprendimento alle informazioni sugli n-grammi di parole del classico modello word2vec, anche le informazioni relative alle subword, ossia ad n-grammi di caratteri, costruendo così un modello di embeddings basato contemporaneamente sulle occorrenze degli n-grammi di caratteri e degli n-grammi di parole contenuti nel training set. Tale modello è in grado di migliorare le prestazioni di determinati task [25], [9] rispetto al classico modello word2vec. Pertanto, utilizzando lo stesso training set sopra descritto, sono stati addestrati e testati anche i modelli FastText, i cui parametri di training sono sempre riassunti nella successiva Tabella 2.

	Word2vec (ISTI-CNR)	**Word2vec (ICAR-CNR)**	**FastText (ICAR-CNR)**
Testo usato per il training	Wikipedia	Corpus biomedico	Corpus biomedico
Dimensione dizionario	733.392	552.291	552.291
Vector size	300	300	300
Parametri di addestramento	iter = 5 algorithm = skipgram window = 10 size = 300 neg-samples = 10	iter = 5 algorithms = [skipgram,cbow] window = 10 size = 300 neg-samples = 10 hs = [0,1] sample = 1e-5	iter = 10 algorithms = [skipgram,cbow] window = 10 size = 300 neg-samples = 10 min_n = 3 max_n = 6 hs = 0 alpha = 0.025 alpha_min = 0.0001 word_ngrams = 1
Numero di modelli ottenuti	1	4	2

Tabella 2 - Parametri di addestramento dei modelli di embeddings testati

In totale sono stati generati 2 differenti modelli word2vec e 2 differenti modelli FastText addestrati su testo di dominio biomedicale in italiano. Questi modelli sono stati testati, insieme al modello word2vec addestrato su testo generico in italiano fornito dal gruppo di ricerca dell'INSTI-CNR, nel processo di supporto alla annotazione del corpus. Come sarà spiegato nel successivo Paragrafo 7, i risultati migliori sono stati ottenuti con l'uso del modello word2vec addestrato con l'algoritmo skip-gram e, pertanto, quest'ultimo tipo di embeddings sono stati utilizzati per la generazione finale del dataset annotato.

5. Espansione del Training Set

Un problema che spesso affligge i dataset che vengono utilizzati per il training di reti neurali e di sistemi di intelligenza artificiale per la classificazione è quello del *class imbalancing* [15], ossia la presenza di esempi di una determinata classe in un numero molto più basso rispetto ad altre nel corpus. Ciò causa la polarizzazione dei risultati del sistema di IA verso quelle classi con il maggior numero di esempi e riduce le performance del sistema in fase di testing. Affrontare tale problema non è banale e una prima soluzione potrebbe essere l'undersampling dei campioni del dataset, ossia l'eliminazione di esempi più numerosi. Questo tipo di approccio, però, non sempre è facilmente applicabile, poiché, soprattutto in un classificatore multiclasse, il ribilanciamento euristico delle classi tramite undersampling può risultare non semplice e, inoltre, nel caso di riduzione troppo spinta del numero di documenti che compongono il dataset, si potrebbe ottenere un corpus di dimensioni non sufficienti a garantire un training corretto, oppure eliminare esempi importanti per il corretto apprendimento della IA. Per questo, recentemente in letteratura sono stati proposti approcci più sofisticati per il ribilanciamento delle classi, come ad esempio in quelli descritti in [16], [17], oltre a metodologie di apprendimento specifiche per dataset con classi non bilanciate [18].

Nel caso di riconoscimento e classificazione di entità all'interno di testo in linguaggio naturale, l'applicazione di undersampling per risolvere il problema del class imbalancing è difficilmente applicabile, poiché eliminare frasi o singole entità potrebbe anche causare la perdita della semantica del documento, oltre che di esempi importanti ai fini dell'apprendimento automatico. Nel caso in esame, però, grazie alla disponibilità di basi di conoscenza di dominio, è possibile applicare oversampling tramite l'espansione del dataset in maniera artificiale ed automatica, riducendo in questo modo l'effetto del class imbalancing e, allo stesso tempo, aggiungendo ulteriori esempi utili al sistema per migliorare ulteriormente le performance. Come visibile nella seguente Tabella 3 il dataset risultante dal processo iterativo di annotazione assistito da IA presenta effettivamente un marcato fenomeno di class imbalancing, dovuto al contenuto stesso dei documenti, che a sua volta inficia il risultato finale dell'addestramento della rete neurale con il training set annotato ottenuto. Quindi, lo step successivo della metodologia proposta ha il fine di estendere il training set creando nuovi documenti ottenuti sostituendo in maniera casuale le entità in numero minore con altre appartenenti alla stessa classe, sfruttando basi di conoscenza di dominio disponibili, per esempio sostituendo *"paracetamolo"* (DRU) con *"antibiotico"* (DRU), oppure, *"ortopedia"* (DEP) con *"unità di chirurgia vascolare"* (DEP), riuscendo così a migliorare i risultati ottenuti dal punto di vista della precisione del sistema per il B-NER.

Più nel dettaglio, la procedura da seguire per l'espansione del dataset è la seguente. Una volta ottenuto il dataset annotato attraverso active learning, così come descritto nel paragrafo 4, è necessario analizzare la distribuzione delle etichette, per verificare quali classi posseggono, in percentuale, il minor numero di esempi. Parimenti, è necessario reperire, o, nel caso, costruire ad hoc, basi di conoscenza contenenti entità appartenenti alle classi che necessitano di espansione, poiché sbilanciate verso il basso. A questo punto si devono estrarre dai documenti tutte le frasi che contengono tali tipi di entità e sostituire le entità suddette con nuove entità estratte casualmente dalle basi, iterando tale procedimento fino al raggiungimento del numero di nuove entità desiderato, stimato in base all'analisi delle occorrenze delle classi eseguito in fase preliminare e, contemporaneamente, all'utilizzo di tutte le entità presenti nelle KB. Infine, le frasi nuove contenenti le entità estratte dalle basi di conoscenza vanno reinserite nel corpus sempre in maniera casuale, mitigando in questo modo in maniera sostanziale così il class imbalancing.

Questo approccio, oltre a ridurre lo sbilanciamento delle classi, comporta anche ulteriori vantaggi, quali: i) estensione del training set, grazie all'incremento di ulteriori esempi; ii) incremento del numero di entità che il sistema vede in fase di addestramento, dovuto all'inserimento di nuove entità provenienti dalla basi di conoscenza, non presenti nel dataset iniziale; iii) incremento del numero di possibili sinonimi che il sistema riesce a interpretare, grazie alle varianti ottenute con le nuove frasi. Di contro, tale approccio basato su di una sostituzione casuale delle entità potrebbe causare la perdita parziale di relazioni semantiche e funzionali tra entità, rendendo lo stesso training set non adatto all'addestramento di modelli IA per il riconoscimento delle relazioni, in uno step successivo. Ad esempio, si potrebbe ottenere a partire dalla frase *"ricoverato in*

cardiologia (DEP) per infarto (DIS)" una nuova frase del tipo *"ricoverato in ortopedia (DEP) per infarto (DIS)"*: benché non cambi la relazione tra una entità di tipo DEP e una di tipo BOD, in questo caso viene modificato il contenuto semantico, oltre che realizzare un esempio di relazione tra entità non plausibile. Applicando tale tecnica di espansione solo al training set da utilizzare per l'addestramento del riconoscimento delle entità, la problematica relativa alle relazioni non si presenta. Il dataset ottenuto tramite active learning, prima dell'applicazione della metodologia di espansione basata su distant supervision, contiene la distribuzione di entità mostrata nella seguente Tabella 3.

Tipo di Entità	**Numero di Entità (Prima dell'Espansione)**
DIS	31.179
MEA	12.168
DAT	4.933
DEP	1.099
ANA	12.258
DRU	2.046
BOD	11.423
THE	8.170

Tabella 3 - Tipi di entità e numero relativo presente all'interno del dataset prima dell'espansione

Come è possibile vedere nella precedente Tabella 3, le classi con il minor numero di esempi, il cui ordine di grandezza differisce da quello delle altre classi, sono del tipo DEP (Dipartimenti) e DRU (Farmaci). Per poter applicare l'espansione è stato necessario quindi costruire due basi di conoscenza contenenti entità di questi due tipi. Nel caso dei farmaci, è possibile sfruttare le liste dei farmaci e delle sostanze in italiano fornite dall'AIFA (Associazione Italiana del Farmaco)[11], da cui è possibile estrarre una lista di tutti i farmaci e dei principi attivi da utilizzare per l'espansione delle entità di tipo DRU. Analogamente, è possibile reperire in rete liste dei dipartimenti presenti nelle strutture cliniche, dai quali, dopo una semplice elaborazione è stato possibile costruire una lista completa di tutti i dipartimenti clinici . Nella seguente Tabella 4 sono mostrate il numero di nuove entità ottenute dalle basi di conoscenza relative alle classi DRU e DEP. Tale approccio, inoltre, permette di superare anche un ulteriore limite dell'architettura Bi-LSTM CRF per il B-NER, relativo ai termini *out-of-corpus*, ossia i termini non presenti nel training set. Infatti, è noto che tale rete neurale, sebbene presenti prestazioni allo stato dell'arte, abbia dei limiti nel classificare correttamente parole out-of-corpus [51]: con l'espansione del dataset, è possibile anche mostrare in fase di training un più vasto insieme di entità di dominio, migliorando anche la successiva classificazione in fase di testing del modello ottenuto.

Iterando il processo di espansione sopra descritto, è possibile incrementare il numero di esempi delle classi sbilanciate, ottenendo la nuova distribuzione mostrata in Tabella 5. Si nota in quest'ultima Tabella, dal confronto con la precedente Tabella 3 contenente la distribuzione originale delle classi di entità, che la metodologia proposta comporta l'aumento anche del numero di entità delle altre classi oltre DEP e DRU,

[11] http://www.agenziafarmaco.gov.it/content/elenco-medicinali-di-fascia-e-h

poiché le frasi che vengono utilizzate per l'espansione spesso contengono entità anche di altri tipi, che, naturalmente, non vengono modificate.

Classe	**Numero di elementi nella base di conoscenza per l'espansione**
DRU	5627
DEP	1554

Tabella 4 - Numero di entità presenti nelle basi di conoscenza da utilizzare nel processo di espansione

Tipo di Entità	**Numero di Entità (Dopo l'Espansione)**
DIS	125.059
MEA	65.668
DAT	34.263
DEP	25.469
ANA	48.878
DRU	45.336
BOD	33.203
THE	46.900

Tabella 5 - Tipi di entità e numero relativo presente all'interno del dataset dopo l'espansione

6. Strumenti per l'implementazione

Il testo originale, al fine di per poter essere annotato e, inoltre, elaborato dalla rete neurale, necessita di essere pre-processato attraverso strumenti classici di NLP. Il testo dei documenti che andranno ad essere processati secondo la metodologia descritta nel precedente paragrafo, infatti, deve essere in primo luogo tokenizzato, ossia ogni singola parola deve essere individuata; inoltre, si devono separare le singole frasi, operazione detta di *sentence splitting*, poiché la rete neurale ricorrente alla base del modello DL utilizzato come supporto per l'annotazione apprende il contesto relativo a singole frasi. Per eseguire tali operazioni è stato utilizzato lo strumento NLP TextPro [19], il quale è specializzato, nonché altamente performante, per operazioni NLP in italiano. Inoltre, allo scopo di ridurre il rimore nel corpus, il testo è stato successivamente posto interamente in lettere minuscole attraverso degli script Python.

I modelli di word embeddings sperimentati (word2vec [22], [23], fastText [24], [25]) sono stati pre-addestrati utilizzando le rispettive implementazioni degli algoritmi fornite in Gensim [20], un package

Python specializzato nell'addestramento di embeddings. Per la costruzione del corpus testuale di dominio biomedicale non annotato per l'addestramento degli embeddings, l'estrazione e il preprocessing sono stati ottenuti attraverso l'uso di TextPro, di script Python e dei tool forniti in Apache Tika, oltre che degli strumenti già descritti nel precedente Paragrafo 4.1 La rete neurale usata e descritta nel precedente Paragrafo 4 è stata implementata in Keras [21], un framework Python per la definizione, l'addestramento e il test di reti neurali di tipo Deep Learning. Infine, le basi di conoscenza e l'espansione del dataset sono state ottenute attraverso la definizione di script Python specifici.

7. Risultati

Attraverso la tecnica proposta è stato possibile annotare un corpus per B-NER in italiano in tempi molto minori rispetto a quelli che sarebbero stati richiesti se si fosse adottato un approccio manuale. Inoltre, l'uso congiunto di active learning e distant supervision ha permesso di mitigare problematiche rispettive che i singoli metodi avrebbero presentato. Il corpus, formato da cartelle cliniche annotate, possiede le caratteristiche riassunte nelle seguenti Tabella 6, che riporta rispettivamente il numero totale di token e di entità nel corpus. Il dettaglio del numero di entità per classe è mostrato nella precedente Tabella 5. Inoltre, come spiegato in precedenza, il corpus è stato usato come training set per l'architettura B-NER descritta nel paragrafo 4. Per testare l'efficacia del modello DL B-NER addestrato, il corpus è stato splittato in un test set e un training set, selezionando casualmente circa il 15% dei token per il test set ed il restante 85% dei token per il training set, i cui dettagli sono riportati nella successive Tabelle 6 e 7.

	Numero totale di parole corpus	**Numero totale di entità**
Totale	1.699.028	424.776
Training Set	1.456.674	364.107
Test Set	242.354	60.669

Tabella 6 - Caratteristiche generali del corpus annotato ottenuto

Tipo di Entità	**Training Set**	**Test Set**
DIS	107.705	17.354
MEA	56.210	9.458
DAT	29.343	4.920
DEP	21.609	3.860
ANA	41.823	7.055
DRU	38.712	6.624

BOD	28.664	4.539
THE	40.041	6.859

Tabella 7 - Numero di entità nel test set e nel training set

Gli esperimenti preliminari hanno avuto il fine di valutare l'efficacia dei modelli WEs addestrati sull'ulteriore corpus non annotato di dominio biomedicale, in termini di performance ottenute durante il primo step della fase di annotazione assistita da active learning, in termini di Precision, Recall e F1-Score, sul test set finale. Come si vede nella successiva Tabella 8, il modello più performante è quello basato sull'algoritmo word2vec addestrato con il metodo skip-gram: pertanto, è stato scelto come modello di input per la rete neurale sia per la fase di annotazione assistita da active learning, sia, successivamente, per la fase di testing delle prestazioni del modello B-NER.

Modello Word Embeddings	**Precision**	**Recall**	**F1-Score**
word2vec (ICAR) skipgram	**0.6062**	0,4131	**0,4734**
word2vec (ICAR) cbow	0,3905	0,3738	0,3905
fastText (ICAR) skipgram	0,4438	**0,4913**	0,4611
fastText (ICAR) cbow	0,3976	0,4478	0,3758
word2vec (INSTI) skipgram	0,5274	0,4624	0,4520

Tabella 8 - Performance del modello DL per B-NER nel primo step di active learning. I modelli ICAR sono stati addestrati su testo di dominio biomedico, mentre quello INSTI è stato addestrato su testo di dominio generale

Partendo dal modello DL addestrato con il corpus frutto del primo step di annotazione manuale, eseguendo 7 iterazioni di annotazione assistita da active learning, inclusa quella iniziale, le performance del modello DL sono state incrementate, fino a che i risultati dell'ultima iterazione non hanno prodotto miglioramenti delle performance sensibili: quindi, la prima fase è stata terminata. Il risultato ottenuto è un corpus contenente un totale di 327.854 token e 60.669 entità. La seguente Tabella 9 mostra i risultati ottenuti dal corpus frutto solo della prima fase basata su active learning, in termini di Precision, Recall e F1-Score. Come è possibile vedere nella Tabella, i peggiori risultati si ottengono proprio per le classi di entità più sbilanciate e con il minore numero di esempi, ossia DRU e DEP.

Tipo di Entità	Precision	Recall	F1-Score
DIS	0,8316	0,9125	0,8702
MEA	0,8436	0,8599	0,8517
DAT	0,8905	0,9492	0,9198
DEP	0,1845	0,1404	0,1595
ANA	0,8145	0,9137	0,8612
DRU	0,8085	0,3576	0,4959
BOD	0,8283	0,8949	0,8603
THE	0,5668	0,8459	0,6788
Media	0,7624	0,7889	0,7618

Tabella 9 - Performance sul test set ottenute dal modello B-NER addestrato sul corpus annotato frutto solo della prima fase di active learning

Quindi, a partire dal corpus ottenuto da active learning, si è applicata la tecnica di di espansione del dataset basata su distant supervision descritta nel paragrafo 5, la quale ha proprio il fine di mitigare i limiti causati dal class imbalance e la scarsa precisione dell'architettura Bi-LSTM CRF quando si trova ad analizzare parole non presenti nel training set, ottenendo il corpus finale oggetto dello studio presentato in questa relazione tecnica, le cui caratteristiche sono state descritte nel dettaglio sempre nel precedente paragrafo 5. Il suddetto corpus annotato è stato suddiviso in un test set ed un training set; quest'ultimo è stato utilizzato per addestrare l'architettura B-NER, la quale ha ricevuto in input il modello word2vec skipgram preaddestrato su testo di dominio biomedico. La successiva Tabella 10 mostra le performance ottenute dalla architettura B-NER sul test set, le quali dimostrano l'efficacia della tecnica di espansione basata su distant supervision, in particolare nel caso di parole appartenenti a classi sbilanciate e di cui ci sono pochi, o addirittura nessun esempio nel training set. In particolare, si nota,confrontando la Tabella 9 e la Tabella 10 che l'applicazione dell'espansione del training set permette di migliorare sensibilmente le performance per tutti i tipi di entità.

Tipo di Entità	Precision	Recall	F1-Score
DIS	0,9595	0,9634	0,9615
MEA	0,9636	0,9675	0,9655
DAT	0,9783	0,9809	0,9796
DEP	0,9878	0,9860	0,9869
ANA	0,9642	0,9679	0,9682
DRU	0,9863	0,9893	0,9878

BOD	0,9203	0,9262	0,9232
THE	0,9609	0,9336	0,9622
Media	0,9642	0,9679	0,9661

Tabella 10 - Performance sul test set ottenute dal modello B-NER addestrato sul corpus annotato ottenuto con la successiva esecuzione della fase basata su distant supervision

In sintesi, i risultati mostrati convalidano l'efficacia del training set ottenuto attraverso la tecnica descritta in questa relazione tecnica, basata sull'uso combinato di active learning e distant supervision, oltre che con il supporto di modelli WEs addestrati su dominio biomedicale, facilitando il lavoro di annotazione manuale da parte degli esperti e, al contempo, superando anche i limiti della architettura DL per quello che riguarda le classi sbilanciate e le parole out-of-corpus.

8. Front-End

Allo scopo di poter utilizzare ed interrogare il modello DL per il B-NER addestrato attraverso il corpus annotato realizzato con la procedura iterativa assistita da intelligenza artificiale sopra descritta, è stato anche prodotto un front-end in applicazione web, mostrato nella successiva Figura 4. Il back-end corrispondente mette a disposizione un servizio RESTful per le richieste, implementato utilizzando il framework Python Flask[12]. L'applicazione web permette sia l'inserimento diretto del testo, ottenendo l'evidenziazione delle entità individuate con un colore differente a seconda della classe corrispondente (vedi Figura 5), sia il caricamento di file contenenti il testo in linguaggio naturale da etichettare.

[12] http://flask.pocoo.org

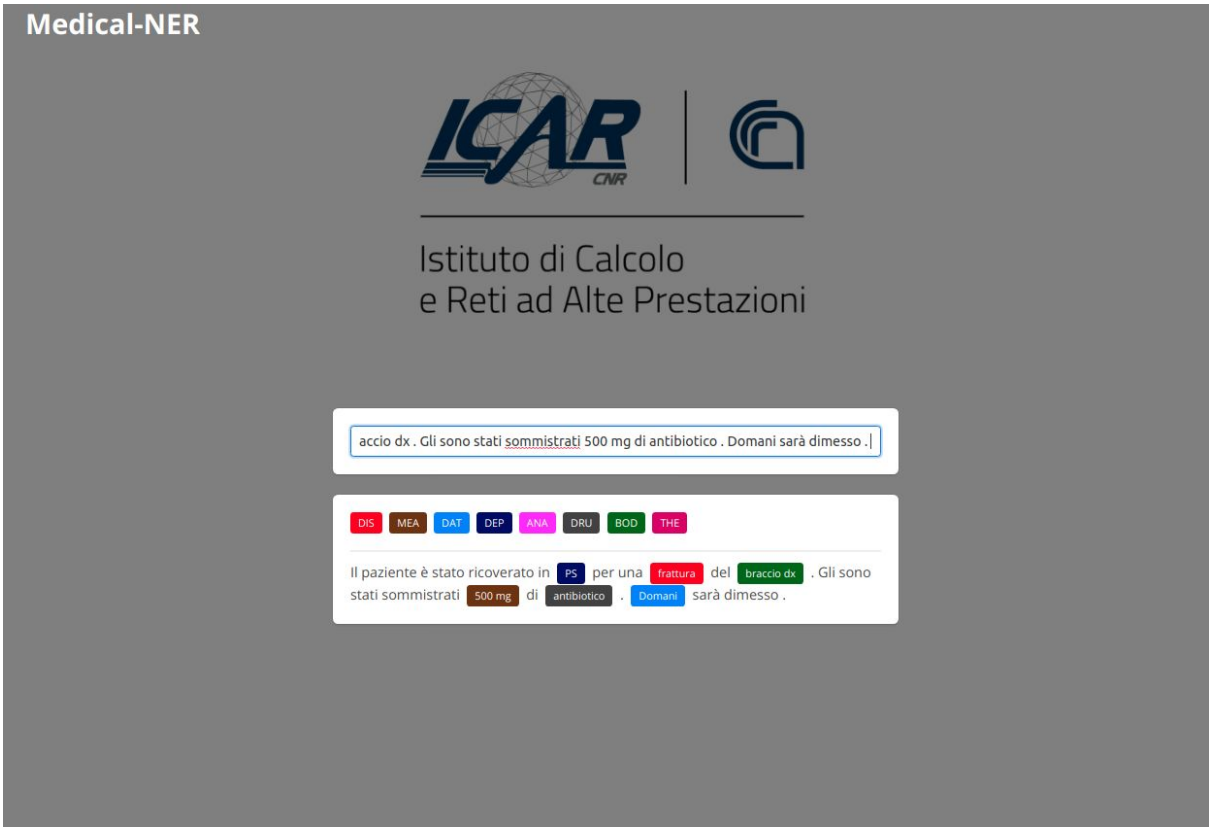

Figura 4 - Front-end dell'applicazione B-NER realizzato

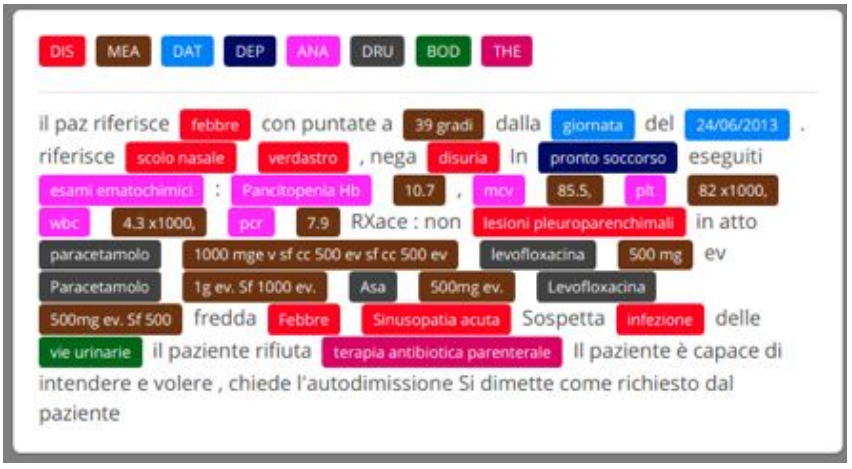

Figura 5 - Dettaglio dell'esempio di utilizzo del front-end

9. Conclusioni e sviluppi futuri

Nel presente rapporto tecnico è descritta una tecnica di annotazione per corpora in linguaggio naturale di campo biomedicale basata su intelligenza artificiale, atta a ridurre l'effort manuale richiesto ad esperti di dominio per lo svolgimento di questo complesso, laborioso e tedioso task. Nel dettaglio, utilizzando word embeddings addestrati su corpus specifico di dominio e una rete neurale allo stato dell'arte per il B-NER, è

possibile assistere e velocizzare il lavoro di annotazione attraverso l'approccio dell'*active learning*. Inoltre, allo scopo di migliorare ulteriormente le prestazioni e, parimenti, risolvere le problematiche legate al class imbalance e alla precisione del modello B-NER su entità out of corpus, è stata proposta una tecnica per l'espansione del dataset basata su *distant supervision* e sull'uso di basi di conoscenza di dominio. I risultati ottenuti mostrano in primo luogo l'utilità della tecnica proposta come supporto all'annotazione, rendendo possibile la riduzione dei tempi necessari a completare tale task. Inoltre, l'efficacia del corpus come training set per B-NER è stata dimostrata attraverso l'addestramento di una rete neurale DL basata sulla architettura allo stato dell'arte proposta da [13], la quale ha prodotto risultati più che soddisfacenti. Inoltre, è stata implementato un front-end web per l'uso del sistema B-NER.

Come sviluppi futuri, in primo luogo è possibile testare differenti approcci Deep Learning nella fase di active learning, in particolare per quello che riguarda la rappresentazione del testo in input. Sono infatti attualmente in fase di sperimentazione e di addestramento una serie di modelli di embeddings basati su ELMo [32] ed altri recenti modelli di tipo *bidirectional Language Model* (bi-LM) ([41], [42], [43]) i quali, a differenza dei classici vector space model ottenuti attraverso autoencoders [22], [24], vengono calcolati come una funzione lineare degli stati interni di una rete neurale DL molto più complessa, costituita, nel caso di ELMo, da una bi-LSTM con incluso in ingresso un livello per la convoluzione dei caratteri e un input iniziale di embeddings standard. Tali modelli hanno dimostrato la capacità di fornire un boost di prestazioni rispetto allo stato dell'arte per molti task del Natural Language Processing, come il NER, il Question Answering, la Sentiment Analysis ed altri. La maggiore complessità di questo tipo di rappresentazioni neurali delle parole, però, comporta un tempo di calcolo molto più elevato sia per il training degli ELMo Embeddings, sia, successivamente, per l'addestramento e il test della rete neurale per il B-NER.

Inoltre, I risultati ottenuti e mostrati nel precedente paragrafo, oltre a verificare l'efficacia e l'utilità della metodologia proposta, ne suggeriscono l'applicazione in differenti domini e diversi task. Pertanto, la stessa pipeline, opportunamente modificata, potrebbe essere utilizzata per la realizzazione di corpora in lingua italiana specifici per ulteriori task NLP, per i quali anche vi è una scarsezza di risorse annotate e di gold standard, come ad esempio il *Text Classification*, il *Relation Extraction* o il *Question Answering*. Parimenti, potrebbe essere usata per l'annotazione di corpora per domini come quello dei beni culturali, o giuridico, in cui recentemente è attivo l'interesse della comunità scientifica sull'applicazione di metodologie basate su intelligenza artificiale.

Infine, sfruttando anche architetture Big Data come ad esempio quella mostrata in [39] e [40] è possibile applicare i risultati ottenuti con il corpus applicato al B-NER per procedere alla strutturazione automatica e all'indexing delle informazioni estratte dai testi biomedici in linguaggio naturale, come ad esempio la parte narrativa di EHR o Personal Health Record (PHR).

10. Ringraziamenti

Si ringrazia Simona Sada per il supporto tecnico fornito nella verifica del dataset per la sperimentazione della metodologia descritta nel presente rapporto tecnico.

Bibliografia

[1] Giuseppe Attardi, Vittoria Cozza, Daniele Sartiano. (2015) *"*Annotation and Extraction of Relations from Italian Medical Records*"*. In: *IIR* 2015. CEUR.

[2] G. Attardi, V. Cozza and D.Sartiano. (2014). "Adapting Linguistic Tools for the Analysis of Italian Medical Records". In: *Vol. I: First Italian Conference on Computational Linguistics CLiC-it 2014*. Pisa University Press.

[3] Yadav, V., & Bethard, S. (2018). "A survey on recent advances in named entity recognition from deep learning models". In: *Proceedings of the 27th International Conference on Computational Linguistics,* pp. 2145-2158. ACL.

[4] Yadav, P., Steinbach, M., Kumar, V., & Simon, G. (2018). Mining electronic health records (EHRs): a survey. *ACM Computing Surveys (CSUR)*, *50*(6), 85. ACM.

[5] Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association,* pp. 1419-1428.

[6] Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE journal of biomedical and health informatics*, *22*(5), pp. 1589-1604. IEEE.

[7] Diomaiuta, C., Mercorella, M., Ciampi, M., & De Pietro, G. (2017, July). A novel system for the automatic extraction of a patient problem summary. In *Computers and Communications (ISCC), 2017 IEEE Symposium on* (pp. 182-186). IEEE.

[8] Diomaiuta, C., Mercorella, M., Ciampi, M., & De Pietro, G. (2017, June). Medical Entity and Relation Extraction from Narrative Clinical Records in Italian Language. In *International Conference on Intelligent Interactive Multimedia Systems and Services* (pp. 119-128). Springer, Cham.

[9] Alicante, A., Corazza, A., Isgrò, F., & Silvestri, S. (2014). Unsupervised information extraction from Italian clinical records. *Proceeding of Innovation in Medicine and Healthcare 2014*, pp. 340-349. IOS Press.

[10] Alicante, A., Corazza, A., Isgrò, F., & Silvestri, S. (2016, June). Semantic cluster labeling for medical relations. In *International Conference on Innovation in Medicine and Healthcare*, pp. 183-193. Springer, Cham.

[11] Alicante, A., Corazza, A., Isgrò, F., & Silvestri, S. (2016). Unsupervised entity and relation extraction from clinical records in Italian. *Computers in biology and medicine*, *72*, 263-275. Elsevier.

[12] Alicante, A., Benerecetti, M., Corazza, A., & Silvestri, S. (2016). A distributed architecture to integrate ontological knowledge into information extraction. *International Journal of Grid and Utility Computing*, *7*(4), 245-256. Indescience.

[13] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, Chris Dyer (2016). Neural Architectures for Named Entity Recognition. In: *HLT-NAACL* 2016, pp. 260-270. ACL.

[14] Erik F. Tjong Kim Sang, Jorn Veenstra. (1999). Representing Text Chunks. In: EACL 1999: pp. 173-179. ACL.

[15] Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis*, *6*(5), 429-449. IOS Press.

[16] Peng, M., Zhang, Q., Xing, X., Gui, T., Huang, X., Jiang, Y. G., Ding, K. & Chen, Z. (2019). Trainable Undersampling for Class-Imbalance Learning. AAAI.

[17] Raghuwanshi, B. S., & Shukla, S. (2019). Class imbalance learning using UnderBagging based kernelized extreme learning machine. *Neurocomputing*, *329*, 172-187. Elsevier.

[18] Mathew, J., Pang, C. K., Luo, M., & Leong, W. H. (2018). Classification of imbalanced data by oversampling in kernel space of support vector machines. *IEEE transactions on neural networks and learning systems*, *29*(9), 4065-4076. IEEE.

[19] Pianta, E., Girardi, C., & Zanoli, R. (2008, May). The TextPro Tool Suite. In *LREC*.

[20] Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45-50. ELRA.

[21] Chollet, F. (2015). Keras.

[22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*, pp. 3111–3119.

[23] T. Mikolov, K. Chen, G. Corrado, J. Dean (2013). Efficient estimation of word representations in vector space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*.

[24] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jegou, T. Mikolov (2016). Fasttext.zip: Compressing text classification models, arXiv preprint arXiv:1612.03651.

[25] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov (2017). Enriching word vectors with sub-word information. *Transactions of the Association for Computational Linguistics,* 5 (2017) 135–146. ACL.

[26] Erik F. Tjong, Kim Sang, Fien De Meulder (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *CoNLL 2003*. pp. 142-147.

[27] Mork, J. G., Demner-Fushman, D., Schmidt, S., & Aronson, A. R. (2014). Recent Enhancements to the NLM Medical Text Indexer. *CLEF (Working Notes)*, pp. 1328-1336. CEUR.

[28] Zavorin, I., Mork, J., & Demner-Fushman, D. (2016). Using Learning-To-Rank to Enhance NLM Medical Text Indexer Results. *Proceedings of the Fourth BioASQ workshop*, pp. 8-15.

[29] Mork, J. G., Jimeno-Yepes, A., & Aronson, A. R. (2013, September). The NLM Medical Text Indexer System for Indexing Biomedical Literature. In *BioASQ@ CLEF*.

[30] S. Peng, R. You, H. Wang, C. Zhai, H. Mamitsuka, S. Zhu (2016). Deepmesh: deep semantic representation for improving large-scale mesh indexing, *Bioinformatics* 32 (12), 70–79. Elsevier.

[31] S. Peng, H. Mamitsuka, S. Zhu (2018). MeSHLabeler and DeepMeSH: Recent progress in large-scale MeSH indexing. *Data Mining for Systems Biology. Methods in Molecular Biology*, Vol. 1807, pp. 203–209. Springer.

[32] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of NAACL-HLT 2018*, pp. 2227–2237. ACL.

[33] Wang Ling, Chris Dyer, Alan W. Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Luís Marujo, Tiago Luís (2015). Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation. *EMNLP 2015*, pp. 1520-1530. ACL.

[34] Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM networks. In *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint conference on* (Vol. 4, pp. 2047-2052). IEEE.

[35] Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *ICML*.

[36] Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). Advances in pre-training distributed word representations. *LREC 2018*, pp. 52-55.

[37] Giacomo Berardi, Andrea Esuli, and Diego Marcheggiani (2015). Word embeddings go to Italy: A comparison of models and training datasets. *Proceedings of the 6th Italian Information Retrieval Workshop*, Cagliari, Italy. CEUR-WS.org.

[38] A. Gupta, P. Goyal, S. Sarkar, and M. Gattu (2019). "Fully contextualized biomedical NER". *Information Retrieval*, L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff, and D. Hiemstra, Eds. Cologne, Germany, pp. 117–124. Springer International Publishing.

[39] Francesco Gargiulo, Stefano Silvestri and Mario Ciampi (2017). "A Big Data architecture for knowledge discovery in PubMed articles." *2017 IEEE Symposium on Computers and Communications (ISCC)*, Heraklion, Greece, pp. 82-87. IEEE.

[40] Stefano Silvestri, Angelo Esposito, Mario Sicuranza, Mario Ciampi and Giuseppe De Pietro (2019). "A Big Data Architecture for the Extraction and Analysis of EHR Data". *IEEE Services 2019*, Milan, Italy. IEEE.

[41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova (2018). "BERT: Pre-training of deep bidirectional transformers for language understanding," *Computing Research Repository*, vol. arXiv:1810.04805.

[42] A.Akbik, D.Blythe and R.Vollgraf (2018). "Contextual string embeddings for sequence labeling,". *COLING 2018*, Santa Fe, New Mexico, USA, pp. 1638–1649. ACL.

[43] A. Akbik, T. Bergmann and R. Vollgraf. (2019). "Pooled contextualized embeddings for named entity recognition". *NAACL-HLT2019*, Minneapolis, MN, USA. ACL.

[44] Katrin Tomanek and Udo Hahn (2009). "Reducing class imbalance during active learning for named entity annotation". *Proceedings of the 5th International Conference on Knowledge Capture (K-CAP 2009)*, pp. 105–112. ACM.

[45] Mateusz Buda, Atsuto Maki and Maciej A. Mazurowski (2018). "A systematic study of the class imbalance problem in convolutional neural networks". *Neural Networks*, 106:249–259.

[46] Weihong Han, Zizhong Huang, Shudong Li, and Yan Jia (2019). "Distribution-sensitive unbalanced data oversampling method for medical diagnosis". *J. Medical Systems*, 43(2):39:1–39:10.

[47] Yaosheng Yang, Wenliang Chen, Zhenghua Li, Zhengqiu He, and Min Zhang (2018). "Distantly supervised NER with partial annotation learning and reinforcement learning". *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2159–2169. ACL.

[48] Yukun Chen, Subramani Mani, and Hua Xu (2012). "Applying active learning to assertion classification of concepts in clinical text". *Journal of Biomedical Informatics*, 45(2):265–272.

[49] Maria Teresa Chiaravalloti, Mario Ciampi, Erika Pasceri, Mario Sicuranza, Giuseppe De Pietro, Roberto Guarasci (2015). "A model for realizing interoperable EHR systems in Italy". *Proceedings of the 15th International HL7 Interoperability Conference*, pp. 13-22.

[50] Mario Ciampi, Giuseppe De Pietro, Christian Esposito, Mario Sicuranza, Paolo Donzelli (2013). "A federated interoperability architecture for health information systems". *International Journal of Internet Protocol Technology*, vol. 7, no. 4, pp. 189-202.

[51] Jiawei Han, Jingbo Shang, Yu Zhang, Xuan Wang, Xiang Ren, Curtis Langlotz, Yuhao Zhang, and Marinka Zitnik (2018). "Cross-type Biomedical Named Entity Recognition with Deep Multi-Task Learning". *CoRR*, abs/1801.09851