UNIVERSITÀ DI PISA

Seminario di Cultura Digitale

# Artificial Neural Networks e Riconoscimento Entità Nominate in una prospettiva storica

**Relazione**

Alessandro Bondielli, matricola 466951

# Abstract

La presente relazione si propone di esaminare e approfondire alcuni degli argomenti trattati nell'ambito del Seminario di Cultura Digitale organizzato dall'Università Di Pisa. In particolare, si cercherà di evidenziare come le tecnologie più all'avanguardia nel campo delle scienze linguistico-computazionali, nello specifico le Reti Neurali Artificiali, possano essere in futuro un eventuale compagno o sostituto di tecniche per l'elaborazione naturale della lingua ad oggi ben consolidate ma che richiedono una conoscenza linguistica a priori molto più ampia, soprattutto in contesti storici in cui il mutamento linguistico, per quanto riguarda sia le strutture sintattiche, sia semantico lessicali dei testi, gioca un ruolo estremamente importante.

# 1. Introduzione

Con l'avvento dell'era digitale, che ha avuto inizio ormai più di due decadi fa, si è assistito a un progressivo interesse per la digitalizzazione di fonti storiche scritte di ogni genere, soprattutto riguardanti il ventesimo secolo, complici anche la presenza di un quantitativo di materiale più massiccio rispetto ai secoli precedenti, la creazione di strumenti standardizzati atti a tale scopo e la crescita esponenziale del Web e di risorse di consultazione più facilmente raggiungibili. Tali contenuti digitali [1] possono essere infatti utilizzati per un gran numero di scopi, che spaziano dalla didattica per le future generazioni alla ricerca di nuove prospettive storiche e, come si vedrà, linguistiche.

Data la mole di testi ormai disponibili anche liberamente ci si potrebbe quasi spingere a parlare di *Big Historical Data*. Considerando anche solamente l'ambito delle due guerre mondiali, oltre ai bollettini ufficiali, è comunque sufficiente pensare alla stampa ed alle comunicazioni personali della popolazione, per ogni nazione coinvolta nei conflitti, per avere un'idea di quanto ampio possa essere il bacino da cui attingere informazioni. La documentazione inoltre, se non ci si limita alle guerre ma si vuole guardare all'intero secolo scorso, appare chiaramente sterminata, ed in continua crescita volgendo lo sguardo alle generazioni future.

## 1.1 Memorie di guerra

Un esempio di utilizzo di una tale mole di risorse può senza dubbio essere *Memorie di Guerra*, un progetto realizzato dal Laboratorio di Linguistica Computazionale dell'Università di Pisa, volto a raccogliere ed analizzare i bollettini ufficiali dei due conflitti mondiali, le prime guerre nella storia dell'umanità documentate in maniera massiva.

Come spiegato dal coordinatore Alessandro Lenci, il progetto è volto a produrre un'analisi computazionale dei bollettini di guerra dei due conflitti mondiali, applicando le tecnologie allo stato dell'arte per quanto riguarda il Natural Language Processing e l'estrazione di informazione (Lenci et al., 2014).

[1] Alcuni esempi di risorse consultabili ed utilizzabili possono essere:

- http://www.europeana1914-1918.eu
- http://www.14-18.it/
- http://www.nationalarchives.gov.uk/first-world-war/

I risultati ottenuti dal progetto sono estremamente interessanti, come anche testimoniato dal sito stesso dove si possono esplorare in molteplici modi i dati raccolti, cercando parole, persone e luoghi dei conflitti. Tuttavia va sottolineato che il progetto è ancora in corso di realizzazione, e le sfide sono molteplici. Tra esse quella che suscita maggior interesse per il presente lavoro è senza dubbio la necessità di poter gestire testi con molto rumore, e soprattutto una grande variabilità, sia per quanto concerne le costruzioni grammaticali, che riguardo al lessico, in cui sono presenti ad esempio variazioni ortografiche per quanto riguarda le entità nominate (luoghi, persone, unità militari ecc.). La soluzione adottata dagli autori del progetto è stata quella di costruire ad hoc corpora di dominio su cui addestrare gli algoritmi di riconoscimento. La base di partenza è il classificatore ItaliaNLP NER (Dell'Orletta et al., 2014), basato su una Support Vector Machine che usa LIBSVM (Chang e Lin, 2001). Questo viene addestrato su I-CAB (Magnini et al., 2006), ma è utilizzato allo scopo anche un *gazetteer*. I gazetteers sono liste contenenti parole, che se compaiono all'interno del testo hanno un'alta probabilità di essere associate ad una entità. In particolare, rappresentano una relazione del tipo *è un* (Sharnagat, 2014). Quello utilizzato infatti contiene nomi di persona e di luogo riguardanti le guerre mondiali, creato manualmente con l'aiuto di un indice analitico presente per i documenti della Seconda Guerra Mondiale, a causa chiaramente dell'assenza di corrispondenza dei nomi contenuti nel corpus storico rispetto a I-CAB. Tale soluzione è sicuramente efficace e comprovata. Difatti è mostrato dai ricercatori come il tagger per le entità nominate risulti avere comunque un'accuratezza prossima, se non in linea, con gli strumenti che rappresentano lo stato dell'arte nel riconoscimento di entità nominate per l'italiano standard contemporaneo (Lenci et al., 2014). Nonostante ciò è innegabile come tale approccio sia estremamente dispendioso, soprattutto per quanto concerne l'effettivo dispendio di risorse umane per l'adattamento degli algoritmi al compito e la creazione appunto di appositi gazetteers.

Le problematiche sorte rispetto alla realizzazione di questo esperimento hanno quindi portato a riflettere sull'eventualità di utilizzare un diverso tipo di approccio e di strumenti per l'analisi dei testi. In particolare si è volto lo sguardo alle *Reti Neurali Artificiali*, una frontiera ormai già consolidata nel campo dell'informatica e delle scienze visive, e che sta negli ultimi anni aprendo nuovi orizzonti di ricerca proprio nel campo del Natural Language Processing, non solo per il corpus di bollettini della seconda guerra mondiale, ma più in generale per l'analisi di testi storici.

Va comunque chiarito come la presente ricerca non voglia assolutamente porsi come la trattazione e l'esposizione di una alternativa valida e comprovata al metodo standard per

l'analisi computazionale del testo, ma semplicemente come la discussione del possibile utilizzo, degli eventuali punti di forza e di debolezza di tale tecnologia, soprattutto in ambiti quali il contesto storico appena presentato, per i quali i metodi più largamente utilizzati necessitano, come già affermato, un notevole sforzo di adattamento al compito. Non si vuole inoltre scendere nei dettagli implementativi delle reti neurali artificiali, lasciando l'arduo compito ai testi citati in bibliografia, poiché oltre ad esulare dalle finalità del presente lavoro, si correrebbe il rischio di presentare spiegazioni parziali e poco chiare, data anche la elevata complessità della materia.

# 2. Artificial Neural Networks

Si può definire molto semplicemente una rete neurale artificiale come "*un sistema computazionale composto da un determinato numero di elementi semplici e altamente interconnessi, che processano informazione attraverso una risposta del loro stato dinamico a input esterni*" (Caudill, 1989). Sono dunque sistemi creati sul modello astratto della struttura neuronale presente nella corteccia cerebrale dei mammiferi, chiaramente in scala estremamente ridotta. In particolare, essi consistono di *input*, che vengono moltiplicati in base a dei *pesi*, e calcolati da una funzione matematica che determina l'*attivazione* del neurone. Un'altra funzione ne calcola l'*output* (Figura 1). La rete neurale artificiale combina questi neuroni artificiali per processare l'informazione (Gershenson, 2001). Chiaramente, in base al peso assegnato all'input, il calcolo effettuato dal neurone sarà differente. La struttura della rete è generalmente costruita su più livelli, come mostrato in Figura 2: un livello è assegnato alla ricezione degli input, che vengono processati e in seguito passati a uno o più livelli intermedi nascosti. Infine vengono ricevuti dal livello di output. Ogni livello cerca di apprendere un concetto dai livelli precedenti. Attraverso ogni successivo livello l'algoritmo di deep learning (tipicamente un algoritmo di *backpropagation*, proposto per la prima volta da Rumelhart et al (1988)) cerca di apprendere strati multipli di concetti di una difficoltà/astrazione crescente (Shanrnagat, 2014).

L'idea alla base di questo tipo di simulazione non è nuova. Il primo lavoro in merito è stato infatti condotto da McCulloch e Pitts risale agli inizi degli anni '40. Sono state in seguito

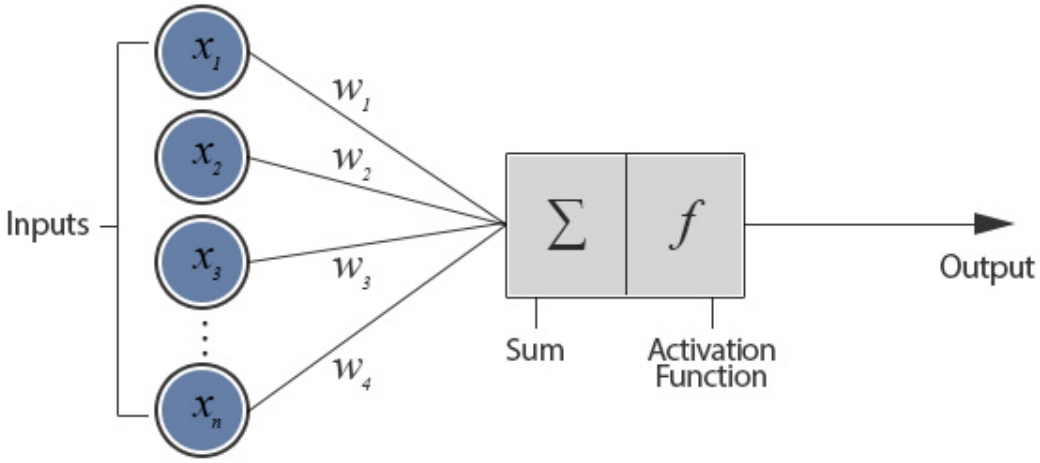


*Figura 1. Struttura di un neurone artificiale, fonte theprojectspot.com (2013)*

chiaramente proposte successive affinazioni, che tuttavia a causa dell'esosità in termini di risorse, hanno faticato a trovare applicazioni in contesti reali, essendo preferite ai modelli più semplici ed economici, come classificatori lineari o Support Vector Machines. Dagli inizi degli anni 2000 invece, con l'avvento del cosiddetto *deep learning* (LeCun et al., 1998), e con la disponibilità di calcolatori sempre più potenti, si è assistito ad una nuova crescita nell'interesse per la materia, anche grazie agli interessanti e incoraggianti risultati ottenuti.

In particolare negli ultimi anni un gran numero di ricerche si è focalizzato sull'applicazione delle reti neurali ai task più diffusi di Natural Language Processing, come ad *esempio POS-tagging*, *chunking*, *semantic role labelling* e *named entity recognition*.

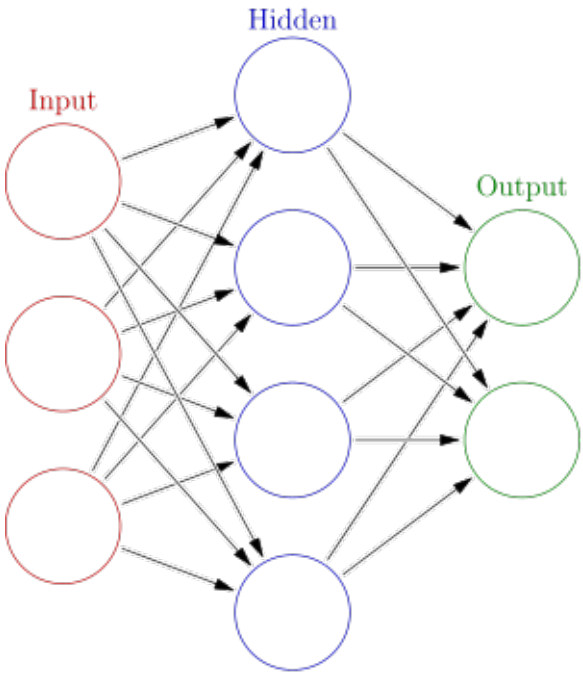


*Figura 2. Organizzazione a livelli della rete neurale artificiale, fonte Wikipedia*

Il focus principale da cui prendono le mosse questi lavori è la creazione di un'architettura neurale unificata e di un algoritmo di apprendimento che siano in grado di essere applicati a svariati task, cercando di evitare la costruzione di specifici sistemi e strumenti per ogni compito, limitando enormemente la necessità di conoscenza linguistica a priori, e utilizzando per l'addestramento larghe quantità di testo senza marcatura (Collobert et al., 2011b).

I risultati ottenuti attraverso l'uso di tali tecnologie sono senza dubbio estremamente interessanti e promettenti, mostrando infatti la capacità di essere all'altezza, in merito alle prestazioni di accuratezza, dei metodi più tradizionali che rappresentano lo stato dell'arte della materia.

## 2.1 Approcci empirici

È interessante quindi analizzare molto brevemente alcuni degli esperimenti effettuati con le reti neurali artificiali, sia per la loro efficacia nel dimostrare le potenzialità di un approccio del genere, sia per l'interesse da essi presentato per gli scopi del presente lavoro.

Zhang e LeCun (2015) propongono l'utilizzo del deep learning in un temporal convolutional network (LeCun et al., 1998) (ConvNets) per l'analisi del testo. Il loro approccio prevede l'assenza di conoscenza pregressa sia per quanto riguarda la struttura sintattica e semantica della proposizione sia in merito al lessico effettivamente presente nei testi analizzati, limitandosi infatti a lavorare con i caratteri in sequenza. È bene notare come questo possa significare la possibilità di applicazione, per un sistema di questo genere, indipendentemente dalla lingua dei testi che vengono considerati e analizzati.

I risultati dei test effettuati su vari compiti, quali ad esempio *sentiment analysis*, *news categorization* (in più di una lingua) e *topic classification*, hanno mostrato come le prestazioni del sistema proposto siano a tutti gli effetti generalmente all'altezza rispetto allo standard rappresentato dallo stato dell'arte per ognuno dei compiti, ponendo importanti basi per sviluppi futuri, soprattutto per quanto riguarda l'apprendimento non supervisionato di modelli linguistici partendo da dati non annotati (LeCun et al., 2015).

Un approccio del genere è stato tentato anche da Collobert et al. (2011b), tuttavia su task più tradizionali della linguistica computazionale, come il *Part-Of-Speech tagging* e, più interessante per i nostri scopi, il *riconoscimento di entità nominate*. Gli autori in questo caso hanno previsto vari approcci ai task proposti, mostrando come, sebbene siano necessari accorgimenti particolari rispetto all'utilizzo della rete e del sottostante algoritmo utilizzato così com'è, grazie a grandi quantità di dati non annotati, la loro rete neurale generica riesce comunque ad ottenere risultati vicini allo stato dell'arte, essendo in grado di individuare feature utili ai task proposti (Collobert et al, 2011b). Viene in seguito mostrato come attraverso l'utilizzo di tecniche più specifiche per ogni compito, e incorporando quindi nel loro sistema alcune procedure comuni in letteratura, i risultati migliorino significativamente. In particolare

in questa sede è importante sottolineare l'uso di *gazetteers* per il riconoscimento di entità nominate, il quale a detta degli autori permette di ottenere un chiaro miglioramento delle prestazioni, che anzi si mostrano leggermente superiori anche alla base di partenza dello stato dell'arte attuale. Il prodotto finale dello studio è un'architettura (SENNA), che raggiunge ottime prestazioni per ognuno dei task presi in considerazione, come mostrato in Tabella 1.

| TASK | BENCHMARK | SENNA |
|---|---|---|
| **PART OF SPEECH (POS)** | 97.24% | 97.29% |
| **CHUNKING (CHUNK)** | 94.29% | 94.32% |
| **NAMED ENTITY RECOGNITION (NER)** | 89.31% | 89.59% |
| **PARSE TREE LEVEL 0 (PT0)** | 91.94% | 92.25% |
| **SEMANTIC ROLE LABELING (SRL)** | 77.92% | 75.49% |

*Tabella 1. Prestazioni di SENNA e stato dell'arte a confronto*

## 2.2 Elaborazione lessicale in una prospettiva neuro-computazionale

Di particolare interesse per la presente ricerca è inoltre sicuramente il lavoro svolto da Marzi, Nahli e Pirrelli (2014), appartenenti al Comunication Physiology Lab del CNR di Pisa, sempre presentato nell'ambito del Seminario di Cultura Digitale.

Senza avere la pretesa, anche in questo caso, di presentare in maniera precisa e dettagliata il lavoro svolto, mostrato in Marzi, Nahli e Ferro (2014), si vuole invece evidenziare come, anche se la ricerca sia in particolare indirizzata allo studio di lingue arabe, possano da essa venir individuati alcuni punti cardine e tratte alcune considerazioni riguardo a prospettive molto interessanti per il contesto storico.

La ricerca presentata volge lo sguardo all'utilizzo di TSOM (Temporal Self Organizing Maps, mappe temporali auto-organizzanti), un particolare tipo di architettura a rete neurale artificiale. Grazie a queste mappe, si può cercare di simulare la competenza morfologica di un parlante, ed in particolare la capacità di distinguere parole note in contesti diversi, come ad esempio, per un parlante inglese l'abilità di riconoscere *book* nella sequenza *handbook*, e per un parlante arabo la radice di un verbo come *k-t-b* nella sua forma flessa *kataba* o *yaktubu* (Marzi, Nahli, Ferro, 2014).

La conoscenza lessicale, ovvero le parole, viene messa a disposizione della rete. Durante la fase di training, ogni nodo della rete sviluppa una sensibilità dedicata a una lettera

possibilmente specifica di una posizione, adattando incrementalmente i propri pesi sinaptici rispetto a pattern ricorrenti o strutture morfologiche.

Alla presentazione di un nuovo simbolo nel livello di input, tutti i nodi sono attivati simultaneamente attraverso le loro connessioni, con una funzione di sensibilità al simbolo corrente e familiarità con il suo contesto. La co-attivazione quindi degli stessi nodi (Best Matching Units, BMU) per differenti parole in input riflette l'estensione con cui la mappa percepisce relazioni morfologiche superficiali tra le parole immagazzinate (Marzi, Nahli, Ferro, 2014).

La co-attivazione di due stimoli può quindi, nella visione degli autori, essere considerata come la rappresentazione più semplice e basilare di ciò che nella percezione è la nozione di similarità, a differenza della prossimità topologica che risulta invece meno efficace.

Gli esperimenti sono stati effettuati sulla lingua araba (fattore, come si vedrà in seguito, comunque importante ai fini della nostra ricerca), ma appare chiaro come una tecnologia di questo tipo possa venire applicata indipendentemente dalla lingua utilizzata. Inoltre, sebbene lo studio sia principalmente volto a utilizzare la co-attivazione per individuare la similarità tra le varie forme flesse delle parole, si può ipotizzare la possibile applicazione anche in contesti dove la similarità e intesa in un contesto più ampio.

Tra le proprietà dimostrate da parte delle reti neurali artificiali in contesti d'uso reali, ai fini della presente ricerca risultano quindi essere di particolare interesse l'indipendenza da una lingua specifica per la comprensione del testo e la capacità di poter simulare in maniera più naturale rispetto agli approcci tradizionali la nozione di similarità.

Vedremo quindi i motivi per cui queste due proprietà possono potenzialmente giocare un ruolo importante se applicate a testi di matrice storica, in particolar modo per quanto riguarda il riconoscimento di entità nominate.

# 3. Prospettive di applicazione per documenti storici

Dopo aver brevemente presentato il funzionamento delle reti neurali artificiali e aver puntualizzato alcuni degli aspetti più interessanti come il fenomeno della co-attivazione e

l'indipendenza dalla lingua, si può andare ad analizzare come l'utilizzo di tali strumenti potrebbe risultare un'alternativa valida laddove gli strumenti di analisi linguistica nati e pensati per essere applicati su testi moderni hanno maggiori difficoltà e maggior necessità di essere manualmente adattati, non con poco sforzo da parte degli studiosi.

Verrà utilizzato come esempio il già citato progetto Memorie di Guerra, di cui verranno analizzate alcune fasi, cercando di capire dove l'implementazione di una rete neurale artificiale possa portare effettivi benefici. Tuttavia, si ritiene che le opinioni espresse possano mostrarsi valide anche in ambiti che non riguardano specificamente l'analisi dei bollettini delle due guerre mondiali.

## 3.1 Problematiche riscontrate

Innanzitutto è necessario chiarire quali siano tali punti problematici, sia nel contesto di Memorie di Guerra, ma più in generale per i documenti testuali di matrice storica. Come già affermato le difficoltà maggiori riscontrate dagli autori sono state causate dalla lingua utilizzata per i bollettini. È chiarito infatti come essi contengano strutture lessicali e sintattiche che caratterizzano l'italiano del secolo scorso (In particolare per quanto riguarda la Prima Guerra Mondiale, quando l'italiano considerato oggi standard era ancora in via di formazione), oltre a lessico specifico di dominio. Sono presenti strutture ellittiche e omissioni causate dallo stile telegrafico. Infine si riscontra un gran numero di costrutti sintattici e lessicali arcaici che potrebbero ostacolare l'annotazione linguistica (Lenci et al., 2014). Termini quali *riparto* in luogo di *reparto* o *schiatori* invece di *sciatori* rischiano di creare difficoltà a strumenti addestrati sull'italiano standard, sia per quanto riguarda analisi linguistiche di base, sia soprattutto in una prospettiva di riconoscimento di entità nominate.

Inoltre, anche la georeferenziazione viene resa più complessa da alcuni fattori, tra i quali soprattutto la presenza di entità nominate non presenti sulle carte odierne, sia a causa di toponimi particolari che indicano luoghi poco frequentati e quindi mal rappresentati (zone nel deserto in Etiopia, luoghi montani ecc.), sia per la variazione dei toponimi stessi. In questo caso può essere discriminante sia il fattore diacronico (nomi caduti in disuso o modificati), sia l'utilizzo di varianti dello stesso nome (viene portato l'esempio di *ValFurva*, *Valle Furva* e *Valfurva*).

Ci si è a questo punto chiesti se questo tipo di problematiche in particolare, ma forse in senso più generale l'intero campo del trattamento e analisi di testi linguistici non conformi allo standard delle lingue odierne, potesse essere affrontato attraverso tecniche alternative e

maggiormente adattabili, senza grande sforzo da parte dei ricercatori, alla risoluzione delle problematiche sorte.

## 3.2 Metodi per il Named Entity Recognition

Prima di considerare tuttavia ipotesi alternative, sembra opportuno effettuare anche una breve panoramica volta a esporre, rimandando comunque per descrizioni più accurate e dettagliate ai testi in bibliografia, alcune delle tecniche diffuse in letteratura per effettuare il riconoscimento di entità nominate all'interno di testi.

Lo scopo del Named Entity Recognition è chiaramente quello di individuare automaticamente attraverso l'uso di algoritmi una parola o una frase che fa riferimento a una particolare entità nel testo (Sharnagat, 2014).

L'interesse per il riconoscimento delle entità nominate non è certamente nuovo, difatti i primi studi in merito risalgono agli inizi degli anni novanta. Mentre i sistemi più primitivi facevano uso di algoritmi che applicavano regole preparate a mano, i sistemi moderni fanno sempre più spesso uso di tecniche di *machine learning* (Nadeau e Sekine, 2007). I progressi di cui si è stati testimoni, soprattutto negli ultimi anni, sono sicuramente incoraggianti, ma è chiaro come il problema del riconoscimento sia ancora oggi un problema tutt'altro che risolto, e un campo di studi estremamente attivo per le scienze linguistico-computazionali.

Gli approcci a questo task si possono generalmente suddividere in tre categorie: supervisionato, semi-supervisionato, e non supervisionato.

Gli approcci di tipo supervisionato sono rappresentati da una classe di algoritmi che apprende il modello tramite cui effettuare il riconoscimento in base a un riferimento costituito da un corpus annotato di addestramento. Tipicamente gli algoritmi imparano a disambiguare in base a feature estratte dal campione in addestramento o cercando di individuare i parametri della distribuzione per cui la somiglianza con il campione di partenza è massima (Sharnagat, 2014). Alcuni esempi di tecniche supervisionate possono essere Hidden Markov Model, Decision Tree models, e Support Vector Machines. Questa ultima tecnica è stata quella utilizzata nel progetto Memorie di Guerra. È chiaro come questo tipo di approccio sia forse il più efficiente, ma anche come richieda un gran numero di risorse, soprattutto per quanto riguarda il corpus di addestramento, che se non costruito precedentemente generalmente necessita di essere creato manualmente dagli studiosi. Tale eventualità si verifica, come per il caso di Memorie di Guerra, quando si vuole analizzare un testo che si discosta, soprattutto per quanto riguarda il lessico, dalla norma della lingua.

Gli approcci semi-supervisionati sono molto simili, ma fanno utilizzo in fase di apprendimento sia di corpus annotati che non annotati per creare le loro ipotesi. Tipicamente gli algoritmi iniziano con un piccolo corpus annotato e procedono a creare le ipotesi considerando grandi quantità di testo non annotato (Sharnagat, 2014). Un esempio possono essre gli algoritmi di bootstrapping, come AdaBoost (Carreras et al., 2002).

Gli approcci non supervisionati cercano invece di risolvere il problema, molto importante soprattutto per alcune lingue, della mancanza di grandi quantità di testo annotato da cui estrarre informazione riguardo le feature utili al riconoscimento, utilizzando principalmente algoritmi di clustering. Alcuni di essi, come ad esempio KNOWITALL (Etzioni et al., 2005) e il sistema di Munro e Manning (2012) cercano rispettivamente essere indipendenti dal dominio a cui vengono applicati e dalla lingua. Come infatti già evidenziato, il problema del riconoscimento presenta tre fattori fondamentali, ovvero lingua, dominio, e tipo di entità, su cui i ricercatori devono focalizzare la loro attenzione.

Infine, come già precedentemente mostrato, la frontiera presentata dal cosiddetto deep learning e dalle reti neurali risulta interessante anche per l'approccio al problema del riconoscimento di entità nominate.

## 3.3 L'ipotesi della rete neurale artificiale

Alla luce anche di quanto evidenziato all'interno del capitolo 2, si è cercato di ragionare su come la prospettiva di utilizzo di tecniche basate su reti neurali artificiali potesse portare ad un effettivo miglioramento delle prestazioni e soprattutto ad una minore necessità di elaborazione manuale dei corpora da parte di personale specializzato.

Come già affermato, risultano essere estremamente interessanti proprio a questo scopo due delle proprietà evidenziate dagli studi precedenti: la potenziale capacità di adattamento a qualunque lingua, senza necessità di una conoscenza a priori particolare per ognuna di esse, e la possibilità di individuare meglio e in maniera più naturale, più simile a quanto avviene attraverso la percezione in un essere umano, la similarità tra due stimoli.

È proprio partendo da questi presupposti che si è cercato di immaginare una soluzione più semplice ed adattabile.

Per quanto riguarda il problema della lingua utilizzata dai testi storici, essa risulta differente da quella utilizzata al giorno d'oggi, particolarmente per lingue come l'italiano, che all'alba

del XX secolo ancora era in larga parte in via di formazione, anche a causa della recente unificazione, avvenuta appena 50 anni prima della Prima Guerra Mondiale (Lenci et al., 2014). L'utilizzo di una struttura basata su rete neurale artificiale potrebbe quindi portare effettivi benefici. Come più volte evidenziato infatti, uno dei vantaggi più immediati di un approccio che non fa uso di dizionari e l'applicabilità immediata a altri tipi di linguaggi naturali umani (Zhang, LeCun, 2015). Si potrebbe quindi ipotizzare che grazie a una struttura di questo tipo, indipendente da particolari conoscenze a priori specifiche per ogni lingua, si possa innanzitutto ridurre la difficoltà iniziale di elaborazione, causata invece dalla necessità dei sistemi più diffusi di essere addestrati sulla stessa lingua in analisi. Sia essa una lingua completamente differente, o semplicemente una variazione arcaica di una lingua odierna, lo strumento utilizzato necessita di pesanti modifiche ed adattamenti per essere efficace. Potenzialmente quindi si potrebbe ottenere un sistema più affidabile già in partenza rispetto ai tool tradizionali per l'elaborazione del linguaggio naturale, per i quali è invece ampiamente riconosciuto un calo nelle prestazioni quando utilizzati su corpora che differiscono dalla tipologia di testi sui quali sono stati addestrati (Lenci et al., 2014). Nel caso invece di una rete neurale artificiale si potrebbe dunque semplicemente limitare la fase di addestramento ai testi storici in analisi, a patto che essi siano una quantità sufficientemente grande da permettere alla rete di effettuare delle predizioni accurate. La capacità quindi della rete neurale di apprendere le feature necessarie all'analisi senza avere una conoscenza pregressa di parole, strutture sintattiche e semantiche della lingua presa in considerazione, qualunque essa sia[2], potrebbe dunque svolgere un ruolo importante anche nelle fasi preliminari del processo di analisi linguistica, fornendo uno strumento più adatto ed adattabile al compito, soprattutto quando i corpora presentano differenze anche significative per quanto riguarda le strutture sintattiche e morfo-sintattiche.

Per quanto riguarda invece l'individuazione e il riconoscimento di entità nominate, i meccanismi diventano più complessi, e di conseguenza risulta più complicato anche ipotizzare la costruzione di un sistema in grado di effettuare questo tipo di elaborazione. Tuttavia anche in questo caso la letteratura offre spunti molto interessanti. Ad esempio gli esperimenti condotti da Petasis et al. (2000), volti a comparare in particolare il comportamento di un *decision tree* basato sull'algoritmo C4.5 e di una rete neurale nel riconoscimento di entità nominate mostrano evidenza che, anche con meno dati come input, la rete neurale ha risultati migliori, sia rispetto

[2] Le evidenze mostrate da Zhang e LeCun (2015) suggeriscono fortemente questa ipotesi. Tuttavia è necessario chiarire ancora una volta come si tratti solo di ipotesi che necessitano senza ombra di dubbio di ulteriori verifiche empiriche.

al classificatore basato sull'albero decisionale sia rispetto al sistema costruito ad hoc, garantendo migliori performance sul task e soprattutto riducendo significativamente lo sforzo necessario per adattare il sistema di riconoscimento di entità nominate a un particolare dominio (Petasis et al., 2000).

Il successo di questo ed altri esperimenti precedentemente menzionati, come in particolare Collobert et al. (2011b), mostrano come i sistemi di riconoscimento basati su reti neurali possano essere molto promettenti.

La capacità inoltre di simulare più efficacemente, tramite co-attivazione di gruppi di neuroni, la nozione di similarità, come suggerito da Marzi, Nahli, e Ferro (2014) ed evidenziato nel capitolo 2.2, potrebbe risultare importante, se non cruciale proprio laddove le entità nominate presenti nei testi siano al giorno d'oggi conosciute con nomi differenti o variazioni lessicali rispetto a quelli presenti nei testi. Si vorrebbe infatti suggerire l'implementazione all'interno del processo di riconoscimento una TSOM. L'idea proposta è quella di utilizzare una rete neurale per il riconoscimento di entità nominate, che faccia uso di un gazetteer. Tuttavia quest'ultimo, invece di essere necessariamente creato semi-manualmente dal corpus storico di dominio che si vuole analizzare, potrebbe in particolar modo per i luoghi, essere costruito partendo molto più semplicemente dalle denominazioni odierne. In combinazione a questo, la TSOM potrebbe venir addestrata sullo stesso elenco di nomi, utilizzando dunque il meccanismo di co-attivazione per cercare di individuare entità nominate simili presenti nell'elenco fornito al sistema. In questo modo, almeno teoricamente, si potrebbe riuscire a far fronte alla variabilità presente nei testi per la denominazione dei luoghi, e ai mutamenti che le stesse subiscono in una prospettiva storica.

Chiaramente, è necessario precisare come la presente idea mostri comunque dei limiti evidenti. Si basa infatti sull'assunzione che la denominazione dei luoghi, come la lingua stessa, sia soggetta a un mutamento progressivo pertanto è verosimile che in casi particolari, in cui i mutamenti sono troppo repentini o drastici per poterne efficacemente tenere traccia basandosi solamente sulla similarità attivata fra le due denominazioni, il sistema possa non essere altrettanto efficace, a meno di particolari accorgimenti. Inoltre, è chiaro come questa proposta, se da un lato potrebbe essere efficace per quanto riguarda le entità che rappresentano i luoghi, non lo è altrettanto chiaramente per altre entità, come ad esempio le persone, che potrebbero non essere adeguatamente rappresentate nella lingua odierna.

# 4. Conclusioni

Le ipotesi sviluppate dalla presente ricerca, senza avere la presunzione volersi affermare come la via corretta da seguire, possono comunque risultare molto interessanti, o quantomeno possono tentare di aprire un dibattito sulla possibilità di affrontare una tematica attuale in maniera originale e, si sostiene verosimilmente efficace alla luce degli studi considerati e presenti in letteratura riguardo le reti neurali artificiali.

Nondimeno appare chiaro come si ponga assolutamente necessario un raffronto con la pratica, prima di poter con un margine sufficiente di sicurezza confermare o smentire ciò che è stato in queste pagine presentato. È chiaro anche, tuttavia, che l'impiego di risorse atte a condurre ricerche di questo genere risulta necessariamente massiccio.

È comunque auspicabile che in futuro l'argomento venga tenuto in considerazione, soprattutto alla luce degli avanzamenti incredibili nel campo delle scienze neuro-computazionali e neuro-linguistiche, che proprio negli ultimi anni stanno mostrando una prospettiva nuova, fresca e potenzialmente più efficace, perché più vicina a quello che può essere il modo di ragionare ed interpretare il mondo proprio dell'essere umano, per affrontare in maniera più semplice ed automatica alcuni problemi di natura linguistico-computazionale che ancora oggi risultano non avere una soluzione ottimale.

Rimane tuttavia da notare come, soprattutto per quanto riguarda la prospettiva storica, di maggiore interesse per questa ricerca, sia necessario fare un enorme balzo in avanti nella digitalizzazione dei documenti. Se infatti negli ultimi anni si è assistito ad un progressivo interesse per la possibilità di avere testi storici digitali, sospinto oltre che dalla volontà di creare biblioteche digitali sempre più ampie e facilmente fruibili e dalla necessità di adattare la didattica storica a un contesto digitale come è quello delle nuove generazione, anche proprio dalle ricerche e dagli incredibili avanzamento nel campo dell'analisi linguistico-computazionale. È quindi assolutamente necessario educare alla volontà di rendere digitale tutta quella conoscenza del mondo prima di noi che ancora giace su carta, ma che ha tutte le potenzialità per fornirci nuove conoscenze e nuove e forse inattese prospettive sulla nostra storia più recente.

# Bibliografia


Boschetti, F., Cimino, A., Dell'Orletta, F., Lebani, G.E., Passaro, L., Picchi, P., Venturi, G., Montemagni, S., e Alessandro Lenci. *Computational Analysis of Historical Documents: An Application to Italian War Bulletins in World War I and II*, in Proceedings of the LREC 2014 Workshop on "Language resources and technologies for processing and linking historical documents and archives – Deploying Linked Open Data in Cultural Heritage" (LRT4HDA 2014), Reykjavik, 2014.

Burger, J. T. *A Basic Introduction To Neural Networks,* URL pages.cs.wisc.edu/~bolo/shipyard/neural/local.html , Giugno 2015

Carreras, X., Marquez, L and Lluıs Padro. *Named entity extraction using adaboost*. In proceedings of the 6th conference on Natural language learning - Volume 20, COLING-02, pp. 1–4, Stroudsburg, PA, USA, 2002. URL http://dx.doi.org/10.3115/1118853.1118857.

Caudill, Maureen. *Neural Network Primer: Part I*, AI Expert, Volume 2 Issue 12, pp. 46-52, 1989.

Chang, C-C., e Lin, C-J. *LIBSVM: a library for Support Vector Machines*. 2001 Software disponibile presso http://www.csie.ntu.edu.tw/~cjlin/libsvm

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K, e Pavel Kuska. *Natural language processing (almost) from scratch*. J. Mach. Learn. Res., 12, pp. 2493–2537, 2011b.

Dell'Orletta, F., Venturi, G., Cimino, A., e Simonetta Montemagni. *T2K$^2$: a System for Automatically Extracting and Organizing Knowledge from Texts.* Proceedings of the 9th International Conference on Language Resources and Evaluation *(LREC'14)*, Reykjavik (Iceland), 2014.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A., Shaked, T., Soderland, S., Weld, D.S. e Alexander Yates. *Unsupervised named-entity extraction from the web: an experimental study*. Artif. Intell., 165, pp. 91–134, 2005, URL http://dx.doi.org/10.1016/j.artint.2005.03.001

Gershenson, Carlos. *Artificial Neural Networks for Beginners*, Formal Computational Skills Teaching Package, COGS, University of Sussex, 2001.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. *Gradient-based learning applied to document recognition*. Proceedings of the IEEE, 86(11), pp. 2278–2324, 1998.

LeCun, Yan, e Xiang Zhang. *Text Understanding from scratch*. URL http://arxiv.org/abs/1502.01710, 2015

Magnini, B., Pianta, E., Girardi, C., Negri, M., Romano, L., Speranza, M., Bartalesi Lenzi, V., Sprugnoli, R. *I-CAB: the Italian Content Annotation Bank*. Proceedings of the 5th International Conference on Language Resources and Evaluation *(LREC'06)*, 2006

Marzi C., Nahli O., e Ferro M., *Word Processing for Arabic Language: A reappraisal of morphology induction through adaptive memory self-organisation strategies.*
In: IEEE - CiST14 - Third IEEE International Colloquium in Information Science and Technology (CIST) (Tetuan (Marocco), 20-22 Ottobre 2014). Proceedings, vol. Catalog Number: CFP1467R-ART pp. 241 – 247, **2014**.

Memorie di Guerra, <http://www.memoriediguerra.it/>, Giugno 2015

Munro, R. e Christopher D. Manning. *Accurate unsupervised joint named-entity extraction from unaligned parallel text*. In Proceedings of the 4th Named Entity Workshop, NEWS '12, pp. 21–29, Stroudsburg, PA, USA, 2012. URL http://dl.acm.org/ citation.cfm?id=2392777.2392781.

Nadeau, D. e Satoshi Sekine, *A survey of named entity recognition and classification*. Linguisticae Investigationes 30, pp. 3–26, 2007

Petasis, G., Petridis, S., Paliouras, G., Karkaletsis, V., Perantonis, S.J., Spyropoulos, C.D. *Symbolic and Neural Learning for Named-Entity Recognition*. In: Proceedings of European

Best Practice Workshops and Symposium on Computational Intelligence and Learning (COIL 2000), Chios, Greece, pp. 58–66, 2000

Rumelhart, D.E., Hinton, G.E. e Ronald J. Williams. *Neurocomputing: foundations of research*. Capitolo Learning representations by back-propagating errors, pp. 696–699. MIT Press, Cambridge, MA, USA, 1988. URL http://dl.acm.org/citation.cfm?id= 65669.104451.

Sharnagat, R., *Named Entity Recognition: A Literature Survey*, Indian Institute of Technology Bombay, 2014

Wikipedia, voce *Artificial Neural Network*, URL https://en.wikipedia.org/wiki/Artificial_neural_network , Giugno 2015