UNIVERSITÀ DEGLI STUDI DI PISA

Corso di Laurea in Informatica Umanistica

Tesi di Laurea

Riconoscimento automatico di entità nominate nel dominio della Pubblica Amministrazione: il caso delle organizzazioni

Relatori:

prof. Alessandro Lenci

prof. Felice Dell'Orletta

Candidato:

Roswita Candusso

matricola: 482938

Anno accademico 2015-2016

Ringraziamenti

Desidero ringraziare la dott.ssa Lucia Passaro, che con la sua collaborazione e la sua guida, ha reso possibile la buona riuscita di questo progetto di tesi. I miei amici e colleghi, vecchi e nuovi, ognuno con la sua peculiarità, che mi hanno accompagnata fino a questo punto; in particolare Andrea e Michele, che grazie alla loro esperienza ed amicizia hanno saputo consigliarmi nei momenti di difficoltà. Vorrei inoltre ringraziare il mio ragazzo, per l'infinita pazienza dimostrata. La mia famiglia per tutto l'amore che mi ha donato: mia madre, per i suoi sacrifici e il continuo supporto, e mia sorella, che mi auguro possa un giorno sentirsi come me oggi.

Indice

Elenco delle figure

Capitolo 1

Introduzione

Il principio di trasparenza, in parallelo al diritto a conoscere, applicato alle Pubbliche Amministrazioni ha conosciuto un'evoluzione significativa nel corso degli ultimi anni in Italia. Evoluzione alimentata da un interesse crescente nell'ambito dell'*open governement*[1]. Questo principio si pone infatti come chiave per garantire l'apertura del patrimonio informativo pubblico, che permette un controllo costante dell'attività da parte dei cittadini, promuovendo al tempo stesso la responsabilità degli amministratori pubblici. Il cosiddetto "Decreto Trasparenza"[2] cita all'articolo 1, comma 1:

[1]Letteralmente: governo aperto

[2]Decreto Legislativo 14 Marzo 2013, n. 33

> La trasparenza è intesa come accessibilità totale delle informazioni concernenti l'organizzazione e l'attività delle pubbliche amministrazioni, allo scopo di favorire forme diffuse di controllo sul perseguimento delle funzioni istituzionali e sull'utilizzo delle risorse pubbliche.

Il principio di trasparenza trova un forte alleato nel *web*, capace di rendere fruibile un'informazione a un numero indefinito di soggetti. Il "Decreto Trasparenza" ha disposto il diritto alla conoscibilità di documenti, informazioni e dati oggetto di pubblicazione obbligatoria: ha previsto una specifica sezione del sito web istituzionale, denominata "Amministrazione Trasparente", in cui tali contenuti devono essere presentati, descrivendone dettagliatamente organizzazione e struttura, dedicando un'attenzione particolare alla qualità delle informazioni. Questa totale accessibilità delle informazioni si forma sul modello offerto dal *Freedom of Information Act*[3] statunitense, finalizzato ad assicurare l'accessibilità a qualsiasi documento o dato in possesso delle Pubbliche Amministrazioni.

È in questo panorama che si inserisce il progetto SEMPLICE, a servizio della Pubblica Amministrazione nella comunicazione, analisi

[3] "atto per la libertà di informazione", è una legge sulla libertà di informazione, emanata negli Stati Uniti il 4 luglio 1966 durante il mandato del presidente Lyndon B. Johnson

delle performance e anche di valutazione del gradimento delle attività degli enti locali. Il progetto risulta quindi in accordo con le attuali necessità di un ente locale in tema di semplificazione, razionalizzazione, efficienza nella gestione dell'informazione della Pubblica Amministrazione. Tenendo presente che gli enti locali, in primis i Comuni, sono chiamati a garantire gli stessi servizi di un ente nazionale facendo conto però su minori risorse, SEMPLICE mira a fornire un'infrastruttura informatica, in linea con i principi dell'Agenda Digitale Italiana, che sia uno strumento di trasparenza verso il cittadino e un supporto alla programmazione.

Alcuni degli obiettivi del progetto sono: fornire una percezione più nitida del territorio amministrato e una conoscenza più rapida delle problematiche e dei bisogni della popolazione. La realtà del territorio viene descritta per mezzo di indicatori quantitativi georefenziati, elaborati per orientare le scelte della *governance* e rappresentati tramite strumenti di *Business Intelligence*. I metodi statistici adottati permettono di verificare concretamente le scelte e prevederne l'impatto futuro.Tra i vari scenari misurabili attraverso gli indicatori di SEMPLICE ci sono performance e spesa, funzionali a dare supporto all'organo politico o alla dirigenza amministrativa nelle scelte relative al personale e al regime economico generale.
Il progetto SEMPLICE è stato realizzato congiuntamente da *01S srl*,

Seacom srl, *Bnova Srl* e il *Dipartimento di Filologia Letteratura e Linguistica* dell'Università di Pisa.

> Dal punto di vista scientifico l'apporto dell'Università sarà inerente lo studio e definizione di sistemi di *ontology learning text mining* e classificazione dell'informazione non strutturata.
> (Semplice PA, Home)

Questi sono *tasks* propri della Linguistica Computazionale, in particolare possiamo classificarli in quel aspetto della disciplina che è l'elaborazione del linguaggio naturale. Attraverso questi processi informatico-linguistici si è in grado di compiere delle estrazioni di informazione strutturata partendo da testo scritto in linguaggio naturale, non strutturato. Il compito dell'Università di Pisa è stato proprio quello di fornire gli strumenti per poter applicare questo tipo di analisi. Questa tesi tratta dei procedimenti che sono stati eseguiti, in particolare nel caso delle organizzazioni. È stato creato ed annotato un nuovo *corpus*, per migliorare le *performances* degli algorimi linguistici all'interno del progetto SEMPLICE, in seguito è stata messa una lente di ingrandimento sulle organizzazioni: la loro classificazione è un ulteriore livello di analisi utile a fini statistici. L'obiettivo è quello di poter riconoscere e distinguere le aziende e gli enti, poterli dividere tra aziende ed enti pubblici o privati.

Capitolo 2

Il progetto SEMPLICE

SEMPLICE (acronimo di "*SEMantic instruments for PubLIc administrators and CitizEns*") è un progetto nato nel 2012, finanziato dalla Regione Toscana[1] basato sull'idea imprenditoriale presentata da un gruppo di aziende ed enti di ricerca capeggiato da *01s s.r.l.*, azienda attiva nella ricerca, sviluppo e innovazione tecnologica. Partner di questo progetto è il Dipartimento di Filologia, Letteratura e Linguistica (FiLeLi) dell'Università di Pisa, che ha collaborato alla realizzazione degli strumenti di analisi semantica dei documenti. L'idea che sta alla base del progetto è essenzialmente la creazione di:

> "Una piattaforma italiana per valorizzare gli *open data* della Pubblica Amministrazione trasformandoli in strumenti utili al

[1] A valere sul Bando Unico POR CREO 2012

cambiamento."
(SEMPLICE PA, SEMPLICE applicato alla Pubblica Amministrazione)

Figura 2.1. Logo del progetto SEMPLICE

Il progetto propone la sperimentazione di una piattaforma di servizi informatici per la gestione e l'organizzazione ragionata dell'informazione nella Pubblica Amministrazione.
Al termine di un progetto finanziato della durata di due anni e mezzo, e terminato nel 2015 nasce un prototipo attualmente in fase di industrializzazione e gestito dalla start-up *Eti3 s.r.l.*, nata da *01s* e deputata specificatamente a gestire le attività legate al progetto SEMPLICE, con il supporto e la collaborazione del Dipartimento FiLeLi dell'Università di Pisa.

2.1 Stato dell'arte

Lo strumento essenziale alla realizzazione di questo progetto è il *Named Entity Recognition*[2], d'ora in poi NER. Il NER può essere definito come un'applicazione di *Natural Language Processing*[3], d'ora in poi NLP, in particolare di *Information Extraction*[4]. Le modalità di utilizzo di questo strumento nell'ambito di NLP sono molteplici, come la traduzione automatica o i sistemi di *question answering*.

2.1.1 Natural Language Processing

Il *Natural Language Processing*, è un campo di ricerca che mira a creare una relazione tra i computer e il linguaggio umano, con l'obiettivo di rendere quest'ultimo interpretabile dal calcolatore. Le difficoltà in questo campo sono soprattutto interpretative e sono dovute all'ambiguità e variabilità del linguaggio. Una frase (o una parola) è ambigua quando le si possono associare più interpretazioni. La complessità del linguaggio umano è tale che si è soliti classificarlo in diversi livelli di astrazione conosciuti come fono-, morfo-, logico/sintattico, semantico e pragmatico.

[2]Riconoscimento di Entità Nominate

[3]Elaborazione del Linguaggio Naturale

[4]Estrazione di informazione strutturata da testo non strutturato

Per risolvere questi problemi NLP procede per livelli:

1. Tokenizzazione

   La tokenizzazione è un processo che consente di dividere un testo in porzioni più piccole dette "token". I *tokens* sono l'unità di base del testo digitale. La nozione di token include quella di "parola", ma è al tempo stesso assai più semplice di quest'ultima.

2. PoS *tagging*

   Il PoS tagging consiste nell'assegnazione di una categoria grammaticale (*Part of Speech*[5]) ad ognuno dei tokens.

3. Parsing

   Il termine *parsing* in linguistica computazionale descrive il processo, eseguito da un *parser*[6], di analisi formale di divione in costituenti di una frase. Il risultato è un *parse tree*[7] che evidenzia le relazioni sintattiche, o dipendenze, tra i costituenti.

4. Lemmatizzazione

   È il processo che, attraverso un algoritmo, permette il riconoscimento del lemma di un *token*, ed eventualmente il raggruppamento di tutti i *tokens* aventi lo stesso lemma. Il lemma è

[5]Parte del discorso

[6]Il parser è un componente sofware che esegue questo tipo di analisi

[7]Albero sintattico

definibile come la forma di citazione di una parola, ossia quella parola che per convenzione è scelta per rappresentare tutte le forme di una flessione.

2.1.2 Named Entity Recognition

Il primo articolo di ricerca sul NER è stato presentato alla "*Seventh IEEE Conference on Artificial Intelligence Applications*" da Lisa F. Rau (1991) in cui viene descritto un metodo per il riconoscimento e l'estrazione dei nomi di aziende, basato su euristiche e regole scritte manualmente. Dal 1996, dopo il primo importante *task* in MUC-6[8], la ricerca in questo settore è stata sempre molto attiva. L'obiettivo di un NER è quello di identificare e classificare in un testo i nomi propri in categorie semantiche definite a priori. Esistono tre categorie definite *Named Entities*[9] universalmente riconosciute: nomi di persona, di luogo e di organizzazioni.

Il compito di un NER può quindi essere diviso in due *sub-tasks*: individuazione delle entità e la loro successiva categorizzazione. Per svolgere questo compito sono stati proposti due approcci: *Rule Based*, ovvero basato su un insieme di regole, e *Machine Learning Based*,

[8]Message Understanding Conference, R. Grishman & Sundheim, 1996

[9]Entità Nominate

basato invece sull'apprendimento automatico[10]. I sistemi che sfruttano insiemi di regole danno buoni risultati, ma richiedono un elevato tempo di elaborazione da parte di linguisti esperti. L'apprendimento automatico, grazie all'utilizzo di tecniche che utilizzano raccolte (*corpora*) di documenti opportunamente annotati per addestrare computazionalmente il classificatore permette di spostare il tempo di sviluppo dalla definizione manuale di regole alla compilazione e annotazione dei documenti del corpus stesso.

2.1.3 Named Entity

La *Named Entity* può essere definita come un sintagma nominale costituito da un nome proprio, come ad esempio un nome di persona o di luogo.

es. "Mario Rossi", "Pisa"

Secondo le specifiche definite da LDC[11] (2005) ci sono sette classi semantiche di *Named Entities* sulle quali i NER lavorano:

- Nomi di persona (Person, PER)

[10] L'apprendimento automatico rappresenta una delle aree fondamentali nel campo dell'intelligenza artificiale e si occupa della realizzazione di sistemi e algoritmi che si basano su osservazioni come dati per la rappresentazione di nuovi contenuti informativi

[11] Linguistic Data Consortium

- Nomi di organizzazioni (Organization, ORG)
- Entità Geo-Politiche (Geo-Political Entity, GPE)
- Nomi di luogo (Location, LOC)
- Edifici o strutture (Facility)
- Strumenti atti allo spostamento (Vehicle)
- Strumenti fisici (Weapon)

Le *Named Entities* da identificare dipendono dal dominio sul quale il NER verrà utilizzato. Nell'eventualità che durante l'addestramento del NER si voglia inserire o modificare un'entità questo processo richiede un'ulteriore annotazione di tutti i documenti già precedentemente annotati.

2.1.4 Features

Per *features* si intendono le caratteristiche di una parola utili ai fini del suo riconoscimento attraverso algoritmi di apprendimento automatico, possono essere costituite da un valore booleano, numerico o nominale. Alcuni esempi di *features* che si possono utilizzare in un NER sono:

- un attributo booleano[12] impostato a *true* se la parola è scritta con la lettera maiuscola, *false* altrimenti.

- un attributo numerico corrispondente alla lunghezza in caratteri della parola.

- un attributo nominale corrispondente alla parola minuscola, o al lemma.

2.1.5 Stanford NER

Lo Stanford NER[13] (conosciuto anche come *CRFClassifier*) è un'implementazione Java di un *Named Entity Recognizer* ed è disponibile per il download sotto licenza *GNU General Public License*[14]. Il software è stato sviluppato da Jenny Finkel presso la Stanford University e le funzioni di estrazione sono ad opera di Jenny Finkel, Dan Klein e Christopher Manning.

Il modello sul quale lo Stanford NER si basa è il *Conditional Random Field* o CRF. I modelli CRF sfruttano la proprietà markoviana

[12]Un attributo booleano è un attributo che può assumere solo due valori, 0 o 1, vero o falso

[13]Per approfondimenti tecnici, le informazioni sono accessibili alla pagina http://nlp.stanford.edu/software/CRF-NER.shtml

[14] https://www.gnu.org/licenses/gpl-3.0.en.html

secondo la quale una decisione da prendere su uno stato in una particolare posizione di una sequenza può dipendere solo da una piccola finestra di altri stati intorno ad esso. Al fine di facilitare il calcolo delle probabilità condizionate, viene usato il *Gibbs Sampling*[15]. La struttura dello Stanford NER permette di addestrare il classificatore per scopi specifici, utilizzando corpus o corpora appositamente annotati.

2.2 NER in SEMPLICE

Essendo il progetto legato al dominio della Pubblica Amministrazione, è stato necessario creare un corpus apposito. Per quanto riguarda le *Named Entities* le tipologie standard sono state modificate definendo quelle che sono necessarie per gli obiettivi del progetto SEMPLICE.

2.2.1 Corpus di training

Inizialmente lo Stanford NER è stato addestrato usando come training corpus I-CAB[16], creato nell'ambito del progetto "*Ontotext*" adottando gli schemi di annotazione sviluppati per "*ACE Entity Detection and*

[15]Campionamento di Gibbs, è un algoritmo comunemente usato come uno strumento per eseguire inferenza statistica, specialmente inferenza bayesiana

[16]I-CAB: the Italian Content Annotation Bank PAPER

TimeExpressions Recognition and Normalization tasks for English". I-CAB è composto di 525 articoli del quotidiano locale "L'Adige" distribuito nella Provincia di Trento. I-CAB si divide in due sezioni, una di *training* e una di *test*, che contengono rispettivamente 335 e 190 documenti.

Successivamente il NER è stato addestrato utilizzando un corpus composto da 461 documenti, a loro volta composti da 724623 tokens. Il corpus è il frutto di una selezione effettuata tra le migliaia di documenti scaricati dagli Albi Pretori[17] dei Comuni italiani attraverso degli strumenti di data crawling. La tipologia di questi documenti è molto variegata, ma principalmente riconducibile a delibere, determinazioni, ordinanze e decreti. La selezione dei documenti è stata casuale, ma rappresentativa: generalmente per ogni Comune sono stati scelti 3-4 documenti.

Ogni documento del corpus è stato annotato seguendo il modello CoNLL[18]. I file di questo tipo contengono dati strutturati come in una tabella, presentano al loro interno un numero predefinito di colonne, in questo caso sette. Ogni colonna rappresenta un diverso livello di annotazione sul token in analisi. I file contengono frasi separate da

[17] L'Albo Pretorio è un luogo virtuale dove ogni Comune pubblica una serie di atti soggetti a pubblicazione obbligatoria. Concretamente si tratta di una sezione specifica del sito del Comune.

[18] Conference on Computational Natural Language Learning

una linea vuota. Una frase consiste in uno o più tokens, ogni token è posto su una nuova linea. Un token è costituto da sette campi descritti nella tabella in figura 2.2. I diversi campi sono separati da un singolo carattere di tabulazione. I dati sono codificati in UTF-8[19] (Unicode). Gli spazi bianchi non sono ammessi all'interno delle righe.

ID	Contatore di token, inizia da 1 per ogni nuova frase	***48***
TOKEN	Il token analizzato nella riga	***legge***
LEMMA	Il lemma del token	***legge***
CPOSTAG	*Coarse-grained tag*, che rappresenta la parte del discorso	***S***
POSTAG	*Fine-grained tag*, descrive la parte del discorso in modo più specifico, se non disponibile è identico al CPOSTAG	***S***
FEATS	*Morphed tag* che aggiunge informazioni morfologiche separate da barra verticale '|' , se non disponibile si presenta come carattere underscore '_'	***num=s|gen=f***
NER TAG	Rappresenta il tag dell'entità annotata	***I-LAW***

Figura 2.2.
[Tabella descrittiva del modello CoNLL]

Coarse-grained tag, fine-grained tag e *morphed tag* sono *features* date in input tramite *Tanl POS Tagset*[20]

[19]UTF-8 è una codifica dei caratteri Unicode in sequenze di lunghezza variabile di byte

[20]È un tagset basato sul tagset ILC/PAROLE ed è conforme agli standard internazionali EAGLES, per maggiori informazioni consultare la sezione Siti consultati

In ultima posizione sulla riga troviamo il tag dell'entità annotata, che è così rappresentato:

B- : BEGIN + tipo entità	rappresenta il primo token dell'entità.
I- : INTERNAL + tipo entità	sono i token facenti parte dell'entità successivi al primo.
O: OUTSET	è un token che non rappresenta nessuna entità.

Figura 2.3. Struttura NER TAG nei documenti del corpus

```
48  legge   legge   S   S   num=s|gen=f I-LAW
```

Figura 2.4. Esempio di riga tratto da un documento tokenizzato del corpus

2.2.2 Adattamento al dominio

Per l'adattamento al dominio della Pubblica Amministrazione sono state utilizzate le seguenti Entità Nominate:

- **PER**: persone

  Vengono annotati il nome e il cognome di una persona. Il prefisso professionale è normalmente escluso, può accadere però che sia inserito tra il nome e il cognome, in quel caso è stato annotato insieme agli altri tokens.

Dott. [Mario Rossi]*PER*

[Mario dott. Rossi]*PER*

```
11  MONNET   MONNET   S    S    num=s|gen=f B-PER
12  DANIELE  DANIELE  S    SP   _    I-PER
```

Figura 2.5. Esempio di entità PER tratto da un documento tokenizzato del corpus

- **ORG**: organizzazioni

  Le organizzazioni rappresentano un mix molto eterogeneo, infatti ogni ente (pubblico, privato, ecclesiastico) è stato classificato come ORG. Un'azienda privata (ente privato):

  ditta [Mario Rossi spa]*ORG*

  Un Ministero (ente pubblico):

  [Ministero dell'Interno]*ORG*

  Successivamente, essendo questo insieme troppo generico e variegato, si è sentita l'esigenza di specificare dei nuovi tipi, o sottotipi, di entità[21].

[21] Si veda il capitolo 3, in particolare la sezione 3.1.2

```
38  Comunità    Comunità    S   SP   _    B-ORG
39  Montana Montana S   SP   _    I-ORG
40  del di  E   EA  num=s|gen=m I-ORG
41  Pinerolese  Pinerolese  S   SP   _    I-ORG
```

Figura 2.6. Esempio di entità ORG tratto da un documento tokenizzato del corpus

- **LOC**: luoghi

  Vengono identificate così le entità geopolitiche. Nomi di città, indirizzi, Stati. Quando si individua l'espressione "Comune di..." essa va annotata come luogo e non come organizzazione amministrativa, allo stesso modo rappresentano entità di luogo espressioni come "Provincia di Milano" o "Città Metropolitana di Torino".

  Es. [Comune di Pisa]LOC

  Un'espressione contenente due diverse entità, come ad esempio: "Città di Pisa Comune di Montescudaio" è stata annotata nel seguente modo:

  [Città di Pisa]LOC

  [Comune di Montescudaio]LOC

ma nel caso si tratti di un indirizzo come: "via Mario Rossi, Comune di Montescudaio, Pisa" questo è annotato interamente, comprensivo di via, numero, cap come un'unica entità:

Es. [via Mario Rossi, Comune di Montescudaio, Città di Pisa]LOC

```
17  Torre   Torre   S   SP  _   B-LOC
18  Pellice Pellice S   SP  _   I-LOC
```

Figura 2.7. Esempio di entità LOC tratto da un documento tokenizzato del corpus

- **ACT**: atti

  Le entità ACT rappresentano i riferimenti a delibere, atti, decreti emanati dal Comune. Non sono stati annotati, ad esempio, atti regionali o delibere di autorità varie.

  Es. [Delibera comunale n° 3 del 11-11-11]ACT

  [Delibera regionale n° 3 del 11-11-11]O

  Per questa entità, estremamente importante considerato il dominio di appartenenza, sono stati definiti dei sotto-tag:

_T: *type*	definisce il tipo, potremmo ad esempio trovare: "Delibera di Giunta Comunale", "Determinazione del responsabile"..
_N: *number*	definisce il numero dell'atto, è quasi sempre specificato singolarmente
_D: *data*	definisce la data di emanazione dell'atto
_U: *unparsed*	definisce una costruzione imprecisa (*unparsed*) che andrà chiarificata in fase di normalizzazione. Ad esempio può succedere di avere il numero nella stessa stringa della data.
_X	definisce all'interno dell'atto tutto quello che non è T, N, D o U. Ad esempio le preposizioni o 'n°'

Figura 2.8. Struttura del tipo di entità ACT

Vengono annotati gli atti che sono provvisti di numero e data, altrimenti non sono dei riferimenti utili. Nell'esempio: "vista la precedente Delibera comunale", "Delibera comunale" non sarà annotato.

```
1   DETERMINAZIONE  determinazione  S   S   num=s|gen=f B-ACT_T
2   DEL di  E   EA  num=s|gen=m I-ACT_T
3   RESPONSABILE    responsabile    S   S   num=s|gen=n I-ACT_T
4   DEL di  E   EA  num=s|gen=m I-ACT_T
5   SERVIZIO    servizio    S   S   num=s|gen=m I-ACT_T
6   N.  N.  S   SP  _   I-ACT_X
7   17  17  N   N   _   I-ACT_N
8   DEL DEL S   SP  _   I-ACT_X
9   24-02-2015  24-02-2015  N   N   _   I-ACT_D
```

Figura 2.9. Esempio di entità ACT tratto da un documento tokenizzato del corpus

- **LAW**: leggi

  Rappresentano tutti i riferimenti che presentano rango normativo, quindi leggi dello stato, come leggi, decreti legislativi, testi unici ecc. Il riferimento legislativo è stato annotato interamente, comprensivo di articolo, comma, lettera. In alcuni casi la fonte normativa sarà espressa con il richiamo a un numero e a una data:

  "Dlsg 267/2000"

  in altri casi invece la fonte sarà indicata con il solo nominativo:

  "Testo unico degli enti locali"

  Esempi annotati:

  dato l'[Art. 25 del TUEL]$_{LAW}$

  visto il [TUEL]$_{LAW}$

```
28  art.3   art.3   S   S   num=p|gen=m B-LAW
29  e   e   C   CC  _   I-LAW
30  17  17  N   N   _   I-LAW
31  del di  E   EA  num=s|gen=m I-LAW
32  D.Lgs   D.Lgs   N   N   _   I-LAW
33  3.2.93  3.2.93  N   N   _   I-LAW
34  n.29    n.29    S   SA  _   I-LAW
```

Figura 2.10. Esempio di entità LAW tratto da un documento tokenizzato del corpus

- **ORG_PA**: partizioni comunali

  L'entità ORG_PA è stata aggiunta alle altre entità in un secondo momento. Tramite questa sottocategorizzazione facciamo riferimento all'insieme di ORG che sono le partizioni organizzative in cui è articolato il Comune. Esse sono molto eterogenee, la denominazione delle partizioni cambia potenzialmente da un Comune a un altro, infatti la loro individuazione è particolarmente delicata. Generalmente sono state individuate come ORG_PA le espressioni che iniziano con "Settore", "Direzione", "Ufficio" ed altri.

  [Servizi Sociali]*ORG_PA*

  [Ufficio Elettorale]*ORG_PA*

```
30  Servizio    Servizio    S   SP  _   B-ORG_PA
31  di  di  E   E   _   I-ORG_PA
32  Polizia Polizia S   SP  _   I-ORG_PA
33  Municipale  Municipale  S   SP  _   I-ORG_PA
```

Figura 2.11. Esempio di entità ORG_PA tratto da un documento tokenizzato del corpus

2.2.3 Features in SEMPLICE

I tipi di *features* che sono state usate nel progetto SEMPLICE sono:

- *Features* a livello di parola:

  Queste descrivono tratti ortografici come ad esempio la prima lettera maiuscola o la punteggiatura per gli acronimi. Tra le *features* di questo gruppo troviamo anche l'inclusione di sotto-pattern. Per esempio, "s.r.l." o "s.p.a." tra le parole che identificano una entità di tipo ORG sono molto utili alla determinazione della classe. Si possono individuare tratti di inizio o fine parola (es. suffissi come "-ista" o "-ore", sono suffissi con i quali terminano molte parole che descrivono professioni umane, come "musicista" o "professore"). Inoltre la parola precedente o successiva all'entità può aiutare a definire la stessa, ad esempio in questa frase:

  "Il presidente [Sergio Mattarella]*PER*..."

  la parola "presidente" aiuta a determinare che le parole seguenti appartengono alla classe *PER*. In SEMPLICE è stata presa in considerazione una finestra di 5 parole (2 precedenti, e 2 successive alla parola target).

- *Features* linguistiche:

  Il NER è stato addestrato fornendo in input caratteristiche linguistiche che in particolare sono: la posizione della parola all'interno della frase con un attributo numerico, il lemma con un

attributo nominale e sempre con un attributo nominale la *PoS* della parola.

- I *gazetteers*:
  I *gazetteers* sono liste di entità note: nei sistemi a regole, questi possono essere usati in maniera deterministica, mentre in molti sistemi statistici vengono usati soltanto come "indizio", dal momento che spesso gli stessi nomi presenti nelle liste non possono essere classificati prescindendo dal contesto, un nome di persona infatti può essere anche il nome di una via. In SEMPLICE vengono utilizzate liste di nomi di persona, di organizzazioni, di luoghi e di entità geopolitiche.

2.2.4 Annotazione

L'annotazione è stata eseguita da due annotatori. I documenti che sono stati annotati sono i 461 documenti appartenenti alla sezione di *training* del corpus, questi ultimi erano già parzialmente annotati, elaborati dal NER implementato in precedenza. Successivamente sono stati annotati anche gli ultimi 25 documenti riservati alla sezione di *test*.

L'annotazione consiste nello scorrere progressivamente il documento controllando che ogni token abbia il giusto tag annotato. Non sono

stati riscontrati omissioni o errori ricorrenti in gran misura . Un errore incontrato è stato, ad esempio. il seguente:

"Deliberazione di [Giunta Comunale]ORG n. 10 del 10-10-10"

al posto di:

"[Deliberazione di Giunta Comunale n. 10 del 10-10-10]ACT"

In casi come questi, non resta che procedere alla correzione modificando mediante un editor di testo piano[22] il documento analizzato.

La classificazione di un token come entità o meno non è un processo immediato dato che, soprattutto in fase iniziale ci sono stati dei dubbi riguardanti l'annotazione di determinati tokens. Per risolvere questi dubbi è stato adottato un approccio collaborativo: è stato creato un documento caricato su Google Docs[23] modificabile da quattro persone, i due annotatori, la dott.ssa Lucia Passaro e Anna Gabbolini, esperte rispettivamente del NER e del dominio della Pubblica Amministrazione. Date queste molteplici conoscenze è stato possibile compiere delle valutazioni competenti sotto ogni punto di vista ed è stato più facile discriminare nel migliore dei modi.

[22] È stato utilizzato Notepad++

[23] è un programma gratuito e basato su Web di elaborazione testi, fogli elettronici, presentazioni e sondaggio, tutto parte di una suite per ufficio offerta da Google come parte del servizio Google Drive. La suite consente agli utenti di creare e modificare documenti online e di collaborare con altri utenti in tempo reale.

2.2.5 Analisi dei risultati

Per poter analizzare le performance è stata eseguita la *Cross-Validation*[24] sulle sezioni di *training* e di *test* del corpus. I valori calcolati sono:

- P: precision, è il quoziente della divisione tra i valori positivi corretti e quelli identificati come appartenenti alla classe (*true positive* e *false positive*). Risponde alla domanda: quanti tra i valori selezionati sono effettivamente corretti?

- R: recall, invece è il quoziente della divisione tra i valori positivi corretti e tutti quelli realmente appartenenti alla classe (*true positive* e *false negative*). Risponde alla domanda: quanti tra i valori effettivamente corretti sono stati selezionati?

- F1: F1 measure, è la media armonica tra i due precedenti valori

[24]In italiano: convalida incrociata, è una tecnica statistica utilizzabile in presenza di una buona numerosità del campione osservato o training set

Entity	P	R	F1
ACT	0.792842	0.869712	0.826468
LAW	0.82675	0.83815	0.83236
LOC	0.70917	0.74939	0.72746
ORG	0.70853	0.68194	0.69407
PER	0.83366	0.86593	0.84895
ORG_PA	0.60898	0.77885	0.6819
Micro AVG	**0.76274**	**0.8166**	**0.78834**
MacroAVG	**0.7467**	**0.7973**	**0.7685**

Figura 2.12. Cross-Validation sulla sezione di *training*

La tabella in figura 2.12 fornisce i risultati relativi alla sezione di *training*, dettagliati per tipo di entità.

	P	R	F1
ACT	0.99202	0.84632	0.9069
LAW	0.9524	0.88	0.9148
LOC	0.8593	0.7095	0.7773
ORG	0.8649	0.7328	0.7934
PER	0.9157	0.76	0.8306
ORG_PA	0.9211	0.8642	0.8917
Micro AVG	**0.9388**	**0.8354**	**0.8841**
MacroAVG	**0.9176**	**0.7988**	**0.8525**

Figura 2.13. Cross-Validation sulla sezione di *test*

La tabella in figura 2.13 riguarda invece la sezione di *test* del corpus, di 25 documenti.

Facendo un confronto tra i risultati ottenuti dopo l'addestramento successivo all'annotazione dei documenti dell'albo pretorio e i risultati ottenuti dopo l'addestramento col primo corpus (I-CAB) nel 2011 possiamo notare che, per le *Named Entities* che sono rimaste invariate (LOC, ORG e PER) c'è stato un miglioramento significativo nel riconoscimento delle entità.

Entity	P	R	F1
LOC	0,8	0,3304	0,4672
ORG	0,64115	0,5447	0,58855
PER	0,85815	0,84125	0,8491
TOTALS	0,7696	0,59765	0,656688

Figura 2.14. Valori riferiti al corpus I-CAB

Capitolo 3

Il caso delle organizzazioni

3.1 Differenziazione delle organizzazioni

La trasparenza è uno degli elementi più importanti con il quale il progetto SEMPLICE deve confrontarsi per soddisfare le esigenze delle Pubbliche Amministrazioni. Nel caso delle organizzazioni è importante poter identificare, ad esempio, le aziende, in modo tale da poter quantificare i servizi resi al comune da una certa azienda, o ancora gli acquisti, le vendite, le convenzioni o gli appalti vinti. Proprio per queste ragioni, il fatto di poter creare uno schema a partire dagli atti comunali, è sicuramente di grande aiuto per chi vuole svolgere un'indagine del tutto limpida.

Inizialmente, in netto contrasto con un'impostazione *trasparente*,

il tipo di entità ORG racchiudeva al suo interno un insieme estremamente variegato di enti, istituzioni e aziende. Questo ha creato da subito il problema di differenziare dalle ORG, tutte quelle organizzazioni che sono invece partizioni comunali[1], risolto poi introducendo la nuova entità ORG_PA. In seguito è stato posto l'obiettivo di differenziare maggiormente questo macro insieme di organizzazioni, in primis rendendo possibile il riconoscimento e la distinzione delle aziende stesse, e di specificare dei sottotipi che comprendessero l'intero insieme dei vari enti all'interno delle ORG. Dati questi nuovi *tasks* si è deciso di proseguire nel progetto inserendo nuove funzioni all'interno dei processi post annotazione.

3.1.1 Pubblico e privato

Inizialmente erano stati creati tre nuovi sottotipi: un insieme racchiudeva gli enti pubblici, uno le aziende e l'ultimo le entità rimanenti. Questa prima schematizzazione non teneva però di conto della dicotomia tra pubblico e privato all'interno dell'insieme di aziende. Questa distinzione è appunto uno dei problemi più importanti da risolvere, sia per quanto riguarda gli enti che per quanto riguarda le aziende. Le aziende infatti, a seconda del loro soggetto giuridico[2], possono essere

[1] Servizi, Uffici, Direzione ed altri

[2] Il soggetto giuridico dell'azienda è la persona fisica o giuridica che assume i diritti e gli obblighi derivanti dalle operazioni aziendali

classificate in aziende private, che hanno un soggetto giuridico privato ed hanno come obiettivo il conseguimento di profitto, e aziende pubbliche che al contrario di quelle private hanno un soggetto giuridico pubblico e come obiettivo il raggiungimento di un obiettivo di interesse pubblico.

3.1.2 Quattro nuovi sottotipi

- **ORG_EPU, Enti Pubblici**

  Un ente pubblico, nell'ordinamento giuridico italiano, è un ente costituito o riconosciuto da norme di legge[3], attraverso il quale la Pubblica Amministrazione svolge la sua funzione amministrativa per il perseguimento di un interesse pubblico.

  In questa tipologia rientrano *Named Entities* come ad esempio:

  Ministero dell'Interno

  Giunta Comunale

  Tribunale

  ed altri.

- **ORG_EPR, Enti Privati**

  Con l'espressione ente privato si intendono organizzazioni con

[3] Legge 20 Marzo 1975, n° 70

persone giuridiche governate dalle norme di diritto privato, come ad esempio onlus, associazioni, partiti politici, sindacati, agenzie. Le aziende non sono state considerate appartenenti a questo sottotipo poiché sono stati riservati loro due ulteriori sottotipi appositi. Possono appartenere a questo sottotipo ad esempio queste organizzazioni:

Partito Democratico
Associazione Nazionale Ufficiali di Stato Civile e d'Anagrafe
Autorità per la vigilanza sui contratti pubblici

ed altri.

- **ORG_APR, Aziende Private**
  Un'azienda privata è un'azienda il cui soggetto economico, che detiene il potere di decidere gli indirizzi strategici, è un istituto di diritto privato.

  Alcuni esempi di questo tipo sono:

  Scarato Mauro srl
  Banca Carisbo
  Totalerg Italia
  Azienda agricola Villanetti

ed altri.

- **ORG_APU, Aziende pubbliche**

  Un'azienda pubblica è un'azienda il cui soggetto economico, che detiene il potere di decidere gli indirizzi strategici, è direttamente o indirettamente un istituto di diritto pubblico, ossia lo stato o altri enti pubblici.

  Questo è stato l'ultimo sottotipo creato. Questa distinzione è stata operata per avere la possibilità di distinguere un'azienda pubblica, come possono essere le Poste Italiane o il Mercato Elettronico della Pubblica Amministrazione (MEPA), dall'altro insieme di aziende private.

  Alcuni esempi, oltre a quelli già citati:

  Azienda Sanitaria Locale

  Ferrovie dello Stato

  Sogepu

3.2 Procedura

Sono stati utilizzati due script scritti in Python, il primo attraverso una funzione "normalizzazione" si occupa della classificazione delle organizzazioni nei quattro nuovi sottotipi, il secondo, quando eseguito, restituisce i file dai quali sono tratte le entità, in questo caso i file del corpus di *training*, con i nuovi sottotipi di organizzazioni correttamente normalizzati.

3.2.1 Python

Python è un linguaggio di programmazione dinamico orientato agli oggetti utilizzabile per molti tipi di sviluppo software. Offre un forte supporto all'integrazione con altri linguaggi e programmi, è fornito di una estesa libreria standard ed è il linguaggio di programmazione utilizzato per le funzioni implementate per il NER. Python, nella versione 2.7.3, è disponibile al download sotto licenza *Open-Source* approvata dalla OSI[4].

3.2.2 La funzione "normalizzazione"

Come primo passo sono stati creati un *gazetteer* e una lista di parole spia che seguono la struttura rappresentata in figura 3.1:

[4]Open System Interconnection

```
ORG_EPU i.n.p.s.
ORG_APR TotalErg
ORG_APR Fininvest
```

Figura 3.1. Esempio tratto dal gazetteer

In prima posizione sulla riga si trova il nuovo sottotipo, in seconda posizione separato da tabulazione si trova l'entità che gli appartiene. La differenza tra il *gazetteer* e le parole spia è la seguente: mentre il primo contiene una lista di intere entità confrontabili, come ad esempio: "Ministero dell'Interno", le parole spia sono delle singole parole che possono, e non, essere incluse all'interno dell'entità analizzata, ad es. "s.p.a.". Lo script viene eseguito su una lista di entità ORG estratte dal corpus di *training*, inserite in output in un file di testo piano in questo modo:

```
001026_2015_119.rtf.conll       188     pubbliche amministrazioni       ORG
001026_2015_119.rtf.conll       275     ANAC    ORG
001026_2015_119.rtf.conll       305     Dipartimento della Funzione Pubblica    ORG
```

Figura 3.2. Esempio tratto dal file contenente le entità ORG estratte dal corpus

ogni riga rappresenta un'entità con un riferimento univoco al documento di provenienza dato dal nome del documento in prima posizione, immediatamente seguito dal numero della riga del primo token dell'entità. Prendendo come punto di partenza questo file è stata creata la funzione "normalizzazione", visibile in figura 3.3.

```
def normalizzazione(i, s, t):
    s_norm = s.upper()
    t_norm = t
    gazetteer = list(open(input_dir_1+'gazetteer.txt'))
    spia = list(open(input_dir_1+'spia.txt'))
    for line in gazetteer:
        uLine = unicode(line.strip(),'utf8')
        gaz = uLine.split('\t')
        e = gaz[1].upper()
        if e in s_norm:
            t_norm = gaz[0]
            break
    if t_norm == 'ORG':
        for line in spia:
            uLine = unicode(line.strip(),'utf8')
            spia = uLine.split('\t')
            e = spia[1].upper()
            if e in s_norm:
                t_norm = spia[0]
                break
    s_norm = s_norm.title()
    return (i, s_norm, t_norm)
```

Figura 3.3. Funzione tratta dallo script in python

La funzione riceve in input un indice (i), l'entità (s) e il tipo di entità (t) e restituisce in output lo stesso indice (i), l'entità normalizzata (s_norm) e il tipo normalizzato (t_norm).

input → (indice, entità, tipo)

output → (indice, entità_normalizzata, tipo_normalizzato)

Figura 3.4. Input e output

Il riconoscimento dei nuovi sottotipi si basa su cicli e confronti. In prima istanza le entità vengono confrontate con quelle presenti nel *gazetteer*, si confronta l'entità ORG estratta dal corpus di *training* con la stringa presente nel *gazetteer* ed eventualmente si assegna all'entità analizzata il nuovo sottotipo di ORG associato alla stringa. Al momento dell'assegnamento l'iterazione si interrompe. Successivamente, per le entità non presenti nel *gazetteer*, alle quali quindi non è stato assegnato alcun sottotipo, si ripete lo stesso procedimento con la lista delle parole spia.

L'intero script restituisce in output un file di testo piano che è così costituito:

```
001026_2015_119.rtf.conll   326 ANAC    Anac    ORG_EPR
001026_2015_119.rtf.conll   33  Giunta Comunale Giunta Comunale ORG_EPU
001026_2015_339.pdf.conll   82  SODEXO ITALIA S.P.A.    Sodexo Italia S.P.A.    ORG_APR
001076_2016_26.pdf.conll    23  MEPA    Mepa    ORG_APU
```

Figura 3.5. Esempio tratto dal file output dello script della funzione Normalizzazione

In ogni riga sono presenti in ordine: nome del file dal quale proviene l'entità, un indice che corrisponde al numero della riga del primo token

dell'entità estratta, l'entità così come è scritta nel file appartenente al corpus originale, l'entità normalizzata, e il nuovo sottotipo.

3.2.3 Normalizzazione del corpus

In un secondo script avviene la modifica dei file appartenenti al corpus. Questo secondo script utilizza in input i valori creati con la funzione 'normalizzazione', quindi entità e tipo normalizzati. Le entità ORG, all'interno del documento tokenizzato, si presentano infine nel formato in figura 3.6:

```
5   GIUNTA  giunta  S   S   num=s|gen=f B-ORG_EPU   Giunta Comunale
6   COMUNALE    comunale    A   A   num=s|gen=n I-ORG_EPU   Giunta Comunale
```

Figura 3.6. Esempio tratto da un documento del corpus tokenizzato dopo la normalizzazione

ovvero in sostituzione al tipo ORG c'è il tipo normalizzato ed è stata aggiunta una nuova colonna che rappresenta l'entità normalizzata, questa è ripetuta per tutti i tokens dell'entità.

3.2.4 Organizzazioni non riconosciute

Dopo l'esecuzione dello script sono rimaste 500 entità ORG per cui non è stato possibile individuare una classificazione all'interno dei quattro sottotipi. Sono state trasferite in output in un file di testo, per poter fare dei controlli ed eventuali miglioramenti all'interno del *gazetteer*.

Si è notato che alcune di queste sono errori di annotazione come ad esempio:

Ministro dell'Interno

"Ministro" al posto di "Ministero"

Sutera giovanna

che dovrebbe appartenere al tipo PER

Servizi sociali

che dovrebbe essere un'entità di tipo ORG_PA

La maggior parte di queste entità escluse sono, invece, aziende private 'non riconoscibili' in quanto non presentano nessuna delle particolarità che le contraddistinguono come tali nello schema, come ad esempio:

Ivo Ferrini

Autotrasporti ed escavazioni

Questo tipo di fenomeno legato alle ORG è da ritenersi fisiologico al metodo di classificazione delle stesse, in quanto il riconoscimento dell'entità avviene in base alla presenza di parole spia come "spa", "snc" e altre sigle societarie simili.

3.3 Risultati

Contando le occorrenze di ogni sottotipo alla fine del processo di normalizzazione sono stati ottenuti i risultati in figura 3.7 e 3.8.

ORG_EPU	1173
ORG_EPR	744
ORG_APU	310
ORG_APR	868
ORG "escluse"	500

Figura 3.7. Numero di occorenze di ogni sottotipo di organizzazione

Figura 3.8. Grafico raffiguarante il numero di occorenze di ogni sottotipo di organizzazione

Facendo una prima analisi, visibile nel grafico a torta in figura 3.9, si nota una maggioranza di occorrenze di enti pubblici, il 37.9%, come poteva essere previsto, dato il dominio di appartenenza di questi documenti. A seguire le aziende private, gli enti privati e le aziende pubbliche.

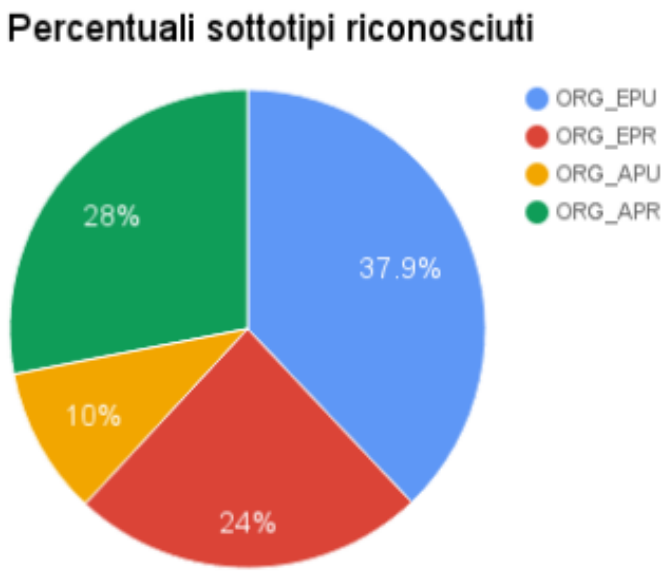

Figura 3.9. Grafico raffiguarante le percentuali di ogni sottotipo di organizzazione

Per quanto riguarda le organizzazioni incluse ed escluse dalla classificazione lo script riconosce e classifica all'interno del corpus di *training* l'86.1% delle entità annotate di tipo ORG. Risultati visibili nel grafico

in figura 3.10.

Figura 3.10. Grafico raffiguarante le percentuali delle organizzazioni classificate nei vari sottotipi

3.3.1 Problemi incontrati

Avendo una conoscenza elementare in ambito giuridico, sono stati riscontrati degli ostacoli che hanno provato dei rallentamenti notevoli, come ad esempio il non saper classificare un'entità nel giusto sottotipo e la conseguente ricerca di informazioni su quest'ultima. Della stessa

natura è stato anche l'errore di riconoscimento delle entità che ora sono ORG_EPR, o enti privati, come ad esempio:

Associazione Nazionale Pubbliche Assistenze (ANPAS)

inizialmente inseriti nel sottotipo ORG_EPU, enti pubblici. Il problema è stato risolto con la creazione del sottotipo adeguato (ORG_EPR) che rendesse rilevante la dicotomia pubblico-privato.

Il problema tutt'ora più significativo è il mancato riconoscimento delle molte aziende private, nel tentativo di porvi rimedio sono stati inseriti nel *gazetteer* una serie di nomi di aziende, provenienti da liste trovate sul web delle maggiori aziende italiane, in modo tale da ampliare il numero delle parole chiave e quindi aumentare la percentuale di riconoscimento di questo insieme. Tuttavia, considerato il fatto che le aziende sono state annotate solo utilizzando il loro nome, qualora questo non fosse presente nel *gazetteer* o non comprendesse la sigla societaria, il riconoscimento non è avvenuto. Un ulteriore ostacolo è costituito proprio dal nome dell'azienda, che, nel caso in cui si tratti di una piccola azienda locale, spesso coincide con il nome del proprietario e, di conseguenza, rende pressoché impossibile l'identificazione di un *pattern* comune.

3.3.2 Sviluppi futuri

Nel futuro il primo obiettivo potrebbe essere quello di trovare un metodo più efficiente per il riconoscimento di tutte quelle aziende private che ad ora sono rimaste escluse, ad esempio avendo una lista delle aziende locali si potrebbe ampliare il *gazetteer* a seconda della Pubblica Amministrazione della quale verranno analizzati i documenti. Inoltre, nonostante le *performances* siano già accettabili, potrebbe essere fatta un'ulteriore ottimizzazione del codice.

Capitolo 4

Conclusioni

Grazie a strumenti linguistici e computazionali è stato possibile operare un tipo di analisi volto all'estrazione di informazione strutturata, da testo scritto in linguaggio naturale e quindi non strutturato. L'annotazione del corpus di *training*, come si è visto nel cap. 2.2.5 "Analisi dei risultati", ha migliorato le prestazioni del NER nel riconoscimento di tutte le entità nominate. La subclassificazione dell'insieme delle organizzazioni ha dato risultati positivi ed è uno strumento indispensabile per svolgere analisi e valutazioni *trasparenti*.

Nell'ambito del progetto SEMPLICE, con questo tipo di informazioni, si intende aiutare le Pubbliche Amministrazioni sia dal punto di vista computazionale nello svolgere analisi sul buon funzionamento dell'ente locale, sia dal punto di vista del rapporto col cittadino,

rendendo disponibili in modo chiaro ed intiutivo notizie che altrimenti sarebbero più difficili da decifrare. Questo rinnovato interesse alla trasparenza è un atteggiamento positivo e stimolante, sia per il cittadino che si informa, sia per la Pubblica Amministrazione che è così maggiormente responsabile del suo operato.

Bibliografia

[1] Landi, Giulio. 1965. Ente (premessa), voce dell'*Enciclopedia del diritto*, Vol. XIV. Milano. Giuffrè. Venetia, 1612.

[2] CoNLL. 2003. *Proceedings of the seventh conference on Natual language learning at HLT-NAACL 2003*, Vol IV. pp. 142-147, 188-191

[3] Monti, Paolo. 15 Febbraio 2006. *Il diritto... E il rovescio*. Bologna. Zanichelli. p. 50

[4] 14 Marzo 2013. *Il Decreto Trasparenza (Decreto legislativo, 14 Marzo 2013, n.33)*, a cura di Ernesto Belisario e Guido Scorza.

[5] Lenci, Alessandro. Montemagni, Simonetta. Pirrelli, Vito. 2005. *Testo e computer. Elementi di linguistica computazionale*. Carocci.

[6] Buitelar, Paul. Cimiano, Philipp. Magnini, Bernardo. 2005. *Ontology learning from text: methods, evaluation and application*. IOS press.

[7] Finkel. Grenager. Manning. 2005. *Incorporating Non-local Information Into Information Extraction Systems by Gibbs Sampling.*

Stanford. Stanford University.

[8] Casella. George. Agosto 1992. *The American Statitician*. Vol 46. Issue 3. pp. 167-174.

[9] Montemagni, Simonetta. 2013. *Studi Italiani di Linguistica Teorica e Applicata (SILTA)*. Numero 1. pp. 145-172.

[10] Schreibman. Siemens. Unsworth. 2008. *The Digital Humanities and Humanities Computing: An Introduction*

Siti Web consultati

[1] Wikipedia. voce:
Azienda Pubblica. https://it.wikipedia.org/wiki/Azienda_pubblica

[2] Wikipedia. voce:
Azienda Privata. https://it.wikipedia.org/wiki/Azienda_privata

[3] Wikipedia. voce:
Ontology Learning. https://en.wikipedia.org/wiki/Ontology_learning

[4] Wikipedia. voce:
Cross-validation. https://it.wikipedia.org/wiki/Convalida_incrociata

[5] Wikipedia. voce: *Campionamento di Gibbs*
https://it.wikipedia.org/wiki/Campionamento_di_Gibbs

[6] Stanford NER. *Software, CRF-NER*.
http://nlp.stanford.edu/software/CRF-NER.shtml

[7] Semplice PA. *Semplicepa, Semplicepa*.
http://www.semplicepa.it/semplicepa

[8] 01s Community. *Home*. http://www.01s.it

[9] Python. *Home*. http://www.python.it

[10] Linguistic Data Cconsortium. *Home*. https://www.ldc.upenn.edu

[11] Youtube. *SemplicePA*.
https://www.youtube.com/watch?v=GvAIsnzjZSU

[12] Medialab. *Tanl POS Tagset*
http://medialab.di.unipi.it/wiki/Tanl_POS_Tagset