

UNIVERSITÀ DEGLI STUDI DI TORINO

DIPARTIMENTO DI INFORMATICA

Corso di Laurea Magistrale in Informatica



Tesi di Laurea Magistrale

**Estrazione di conoscenze per una
storiografia digitale**

Relatore:

Prof. Vincenzo Lombardo

Candidato:

Giuseppe Biondi

Anno Accademico 2020/2021

Abstract (Italiano)

Historygraphia (HSG) è un progetto in ambito storico, nato dall'obiettivo di proporre una piattaforma web per l'interconnessione e la diffusione della conoscenza storica, con applicazioni alla ricerca, la formazione, i servizi per la cittadinanza attiva. Il sistema è distribuito su tre livelli di rappresentazione: narrazioni su specifici argomenti, fonti storiche e conoscenza semantica sottostante. L'obiettivo di questa tesi è l'automazione del processo di estrazione della conoscenza contenuta nelle storie, per la costruzione di grafo di conoscenza (knowledge graph), a supporto dell'interconnessione delle narrazioni. Questo progetto descrive tematiche e sfide tipiche del processo di rappresentazione digitale della conoscenza e propone uno studio in ambito Information Extraction (IE), un ramo del Natural Language Processing (NLP), per la combinazione di diversi tool esistenti col fine di costruire un grafo di conoscenza partendo da saggi narrativi, compilati da esperti nelle discipline storiche.

Abstract (English)

Historygraphia (HSG) is a project in the historical field, which is born from the aim of promoting a web platform for the interconnection and dissemination of historical knowledge, with applications to research, training, services for active citizenship. The system is distributed on three levels of representation: narration on specific topics, historical sources and underlying semantic knowledge. The goal of this thesis is the automation of the process of extracting the knowledge contained from stories, for the construction of a knowledge graph to support the interconnection of narrations. This work describes typical issues and challenges about the digital representation process of knowledge and it proposes a study in the field of Information Extraction (IE), a branch of Natural Language Processing (NLP), for the combination of several existing tools in order to build a knowledge graph starting from narrative essays, compiled by experts in the historical disciplines. e

Indice

1	Introduzione	1
2	L'elaborazione automatica dei documenti storici	6
2.1	Ciclo di vita dell'informazione storica	6
2.2	Classificazione dei documenti storici	9
2.3	Problemi aperti	12
2.3.1	Problemi legati alle fonti storiche	13
2.3.2	Problemi legati alle relazioni tra le fonti	15
2.3.3	Problemi legati all'analisi	16
2.4	Conclusioni di capitolo	19
3	Information extraction (IE)	20
3.1	Wrapping	21
3.2	Estrazione di entità	23
3.2.1	Task	24
3.2.2	Tecniche	30
3.3	Estrazione di relazioni	32

INDICE

3.4	Conclusioni di capitolo	35
4	HiStoryGraphia	36
5	Il modulo Storytelling2Knowledge	43
5.1	Architettura	44
5.2	Wrapper	47
5.3	Estrazione di entità	50
5.3.1	Tools	51
5.3.2	Euristiche per la sovrapposizione	57
5.4	Estrazione di relazioni	58
5.5	Conclusioni di capitolo	60
6	Risultati	62
6.1	Conclusioni di capitolo	76
7	Conclusioni e sviluppi futuri	78
	Bibliography	85

DICHIARAZIONE DI ORIGINALITÀ

"Dichiaro di essere responsabile del contenuto dell'elaborato che presento al fine del conseguimento del titolo, di non avere plagiato in tutto o in parte il lavoro prodotto da altri e di aver citato le fonti originali in modo congruente alle normative vigenti in materia di plagio e di diritto d'autore. Sono inoltre consapevole che nel caso la mia dichiarazione risultasse mendace, potrei incorrere nelle sanzioni previste dalla legge e la mia ammissione alla prova finale potrebbe essere negata."

Capitolo 1

Introduzione

History and computing (in italiano, storia e computer), più recentemente rinominato come *scienza dell'informazione storica* da Boonstra et al. [1], è un campo di ricerca che studia l'utilizzo delle tecnologie informatiche applicate alla ricerca storica.

Il rapporto tra computer e ricerca storica esiste quasi dalla nascita dei computer stessi e nel corso degli anni ha creato una forte connessione: le continue evoluzioni delle tecnologie informatiche, come ad esempio l'introduzione dei database, hanno influenzato il modo di fare ricerca. I database hanno fornito uno strumento per memorizzare l'informazione e recuperarla in maniera più semplice ed efficace [2].

Tale collaborazione ha portato ad una crescente diffusione nel processo di digitalizzazione del materiale storico. Tuttavia, sono emerse delle difficoltà legate alla diffusione e al confronto di tali contenuti.

Quaresma afferma che [3]:

«*I documenti storici hanno un'enorme quantità potenziale di informazioni, che non è facilmente accessibile ai ricercatori o ai cittadini.*»

Questo avviene poiché la semplice digitalizzazione e condivisione dei documenti non permette di risalire all'informazione contenuta in essi. È necessario, infatti, distinguere i concetti di informazione e documento, in quanto, citando Seth Den-

CAPITOLO 1. INTRODUZIONE

bo [4], "gli studiosi cercano informazioni piuttosto che semplici documenti".

Digitalizzare una fonte e renderla accessibile a chiunque non garantisce che gli interessati riescano a raggiungerla facilmente. Lo studioso necessita di un supporto che lo indirizzi ai documenti che contengono l'informazione a cui è interessato.

Il *Semantic Web*, termine coniato da Tim Berners-Lee, rappresenta un'estensione del web classico che permette di descrivere le informazioni contenute nei documenti in un formato elaborabile in modo automatico tramite l'aggiunta di metadati.

I metadati sono definiti come «*informazioni, comprensibili dalla macchina, relative a una risorsa Web o a qualche altra cosa*» [5, 6], essi permettono di associare i dati a delle risorse descrittive, permettendo una cooperazione migliore tra uomo e computer.

L'aggiunta di metadati al materiale storico permette di rappresentarne il contenuto semantico, in altre parole descrive, in un linguaggio comprensibile dai computer, le informazioni contenute al suo interno. Inoltre, permette l'utilizzo di strumenti di ricerca per risalire al documento in base all'informazione desiderata e di creare relazioni tra i dati per migliorarne l'esplorazione.

Questo nuovo tipo di rappresentazione dell'informazione è detta *grafo di conoscenza*. Partendo da un documento possiamo estrarne gli elementi principali, che costituiranno i nodi del grafo, e unirli in relazioni tra di loro, tramite archi associati ad un significato semantico che ne identifica il legame.

Ad esempio, dato il documento storico in figura 1.1, un estratto del contenuto è rappresentato dal grafo di conoscenza in figura 1.2 in cui l'informazione descritta tratta la commissione, da parte della Confraternita di Santa Croce, del ciclo di 12 teleri a Lorenzo Gastaldi.

CAPITOLO 1. INTRODUZIONE

Article: Entracque – Confraternita di Santa Croce

Author: Gelsomina Spione

diocesi di Mondovì

La chiesa della confraternita viene edificata nel 1538, segno dell'affermazione di un notabilato che controlla la vita economica e sociale della comunità e sviluppatiso tra la metà del Quattrocento e la fine del Cinquecento, a seguito dell'incremento dei traffici commerciali. La rete economica favorisce anche la circolazione di artisti. I confratelli di Santa Croce commissionano (1658 - 1660) un ciclo di dodici tele a Lorenzo Gastaldi. La scelta del pittore, che in quel momento si trova a Monaco, e realizza molte opere per le valli del Nizzardo e il basso

...

Si tratta di dodici teleri raffiguranti episodi della vita di Cristo e della Vergine. Una scelta che risponde ai dogmi affermati dalla Chiesa cattolica della Controriforma, l'Immacolata Concezione della Vergine, la Trinità, il riscatto del genere umano attraverso il Sacrificio di Cristo. A questo si aggiungono i riferimenti alle pratiche liturgiche svolte dai confratelli durante il Giovedì Santo, rievocate nell'Ultima Cena e soprattutto nella Lavanda dei piedi. Il modello decorativo di riferimento è quello della Confraternita di Santa Croce a Cuneo, dove nel 1626 sono allestite le tele con i Miracoli della Vera Croce dipinti nel 1626 da Giulio e Giovanni Battista Bruno (che a loro volta rimandano ai modelli decorativi di primo Seicento delle Casacce (oratori), genovesi.

...

Figura 1.1: Estratto narrazione: Entracque – Confraternita di Santa Croce

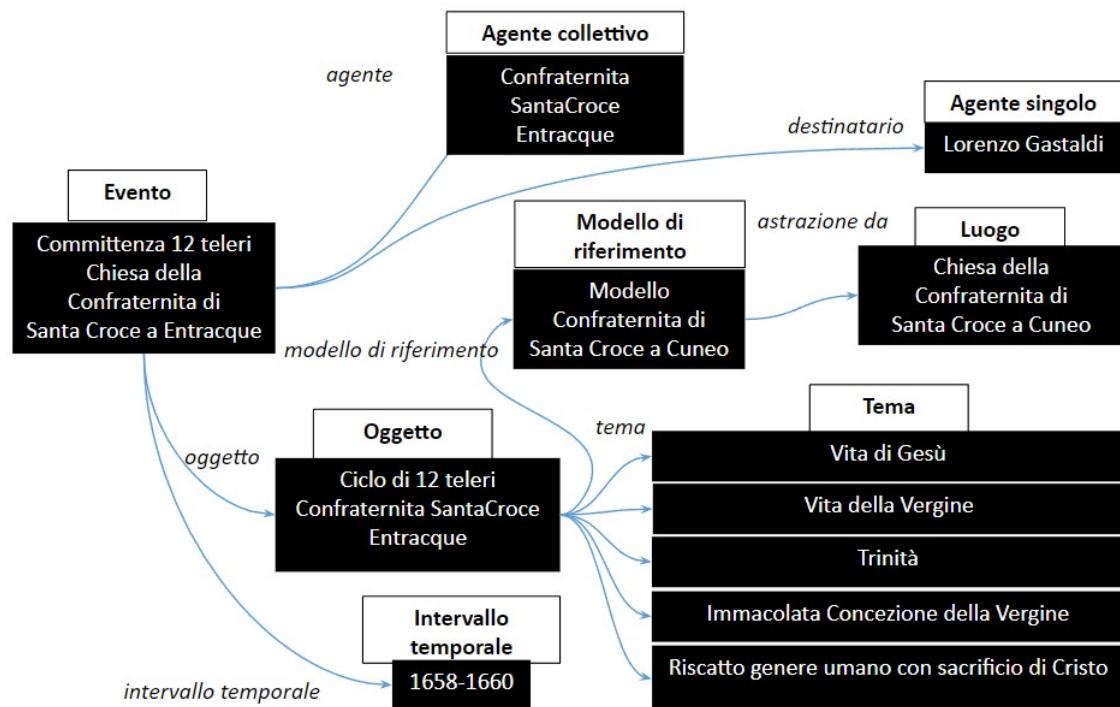


Figura 1.2: Parte del grafo di conoscenza relativo a 1.1

CAPITOLO 1. INTRODUZIONE

La rappresentazione di dati storici tramite i grafi di conoscenza, non solo permette di migliorarne l'accessibilità, grazie a sistemi di ricerca basati sul contenuto semantico, ma permette anche di confrontare il contenuto di diversi documenti e trovare nuove connessioni tra le informazioni presenti (ad esempio, i pittori che hanno lavorato per gli stessi clienti).

In questo contesto nasce il progetto HiStoryGraphia, il cui obiettivo è di costruire una piattaforma web finalizzata all'interconnessione e diffusione della conoscenza storica. Questa risorsa prevede due tipi di utenti: i narratori, che si occupano di inserire nuove narrazioni, e i visitatori, che ricercano informazioni e connessioni.

I saggi, o narrazioni, contengono l'informazione da introdurre nel sistema. Per rappresentare questa informazione in un linguaggio comprensibile dalla macchina, è necessario rappresentarne il contenuto semantico tramite l'aggiunta di metadati.

Attualmente questo lavoro avviene attraverso un processo manuale da parte degli autori, utilizzando un'interfaccia web. L'obiettivo di questa tesi è la costruzione di un sistema di supporto a questa fase, che suggerisca i metadati necessari, estraendoli automaticamente dai saggi storici e generando un grafo di conoscenza.

In particolare, la tesi vuole: individuare alcuni aspetti legati alla rappresentazione digitale delle informazioni da documenti storici, identificare una serie di strumenti e risorse open source per la lingua italiana, presentare un'architettura modulare per l'interconnessione delle conoscenze che sia aggiornabile con l'aggiunta di nuove soluzioni e/o euristiche con l'evoluzione della piattaforma.

La tesi è organizzata come segue. Nei capitoli 2 e 3 si presenterà l'attuale stato dell'arte riguardo, rispettivamente, l'elaborazione di informazione storica, con particolare attenzione alle difficoltà legate al dominio, e l'estrazione di informazione. Nel capitolo 4 sarà presentato brevemente il progetto HiStoryGraphia, la sua struttura ed esempi di narrazioni. Nel capitolo 5 si presenterà l'idea di architettura proposta, i componenti e come questi interagiscono e si introduranno le

CAPITOLO 1. INTRODUZIONE

risorse utilizzate, o costruite, in relazione al contenuto del capitolo 3. Si termina con due capitoli: uno dedicato ai risultati prodotti e l'ultimo, il settimo, conterrà una conclusione con possibili idee per il futuro.

Capitolo 2

L'elaborazione automatica dei documenti storici

Esistono molte sfide e problemi aperti nel dominio storico. Queste difficoltà riguardano principalmente problemi testuali, di collegamento, strutturazione, interpretazione e visualizzazione [2, 1].

La finalità del capitolo è di presentare un contesto legato all'elaborazione automatica dei documenti storici riportando il contenuto principale di un survey condotto da Meroño et al. [2]. Il tema riguarda il ciclo di vita dell'informazione storica e alcuni tipi di classificazioni dei documenti storici, rispettivamente, trattati nella prima e nella seconda sezione. Essi saranno utili per introdurre e illustrare, nella terza sezione, i problemi specifici affrontati, contestualizzandoli all'interno del ciclo di vita e accompagnandoli con relativi esempi.

2.1 Ciclo di vita dell'informazione storica

Nel libro titolato "Past, present and future of historical information science", pubblicato da Boonstra et al. nel 2004 [1], viene affrontato il tema del progresso tecnologico in relazione al contenuto storico. In particolare, sono state evidenziate una

CAPITOLO 2. L'ELABORAZIONE AUTOMATICA DEI DOCUMENTI STORICI

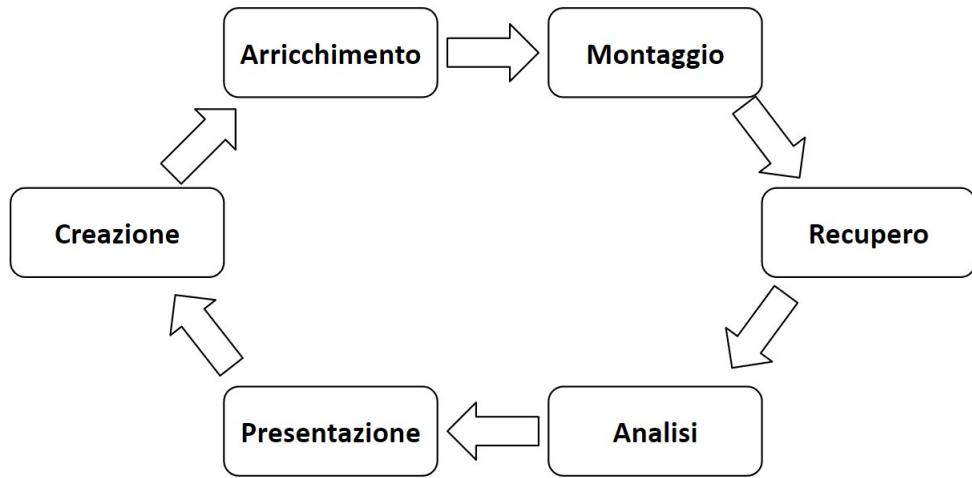


Figura 2.1: Il ciclo di vita dell'informazione storica [2]

serie di fasi che caratterizzano la vita dell'informazione storica, dalla creazione all'utilizzo, nel ciclo di vita mostrato in Figura 2.1.

Le fasi, non tutte necessariamente presenti, descrivono il flusso di lavoro dell'informazione storica:

- **Creazione:**

Comprende la creazione fisica dei dati digitali, incluso il design della struttura e del progetto.

Esempi di attività legate a questa fase potrebbero essere la pianificazione dell'inserimento dei dati, la digitalizzazione (i.e., OCR¹) oppure considerare il software database più appropriato.

- **Arricchimento:**

In seguito alla creazione avviene la fase di arricchimento dell'informazione storica, ottenuta tramite l'utilizzo di metadati descrittivi, preferibilmente utilizzando standard come ad esempio Dublin Core [1].

Questa fase comprende anche il collegamento di diversi individui riferiti alla stessa entità reale. Per esempio, la nazione degli *Stati Uniti* può essere

¹Riconoscimento ottico dei caratteri(Optical Character Recognition)

CAPITOLO 2. L'ELABORAZIONE AUTOMATICA DEI DOCUMENTI STORICI

riconosciuta anche come "*USA*" e "*US*" [7]. Un secondo esempio è papa Francesco che può anche essere riferito come Jorge Mario Bergoglio, due nomi associati alla stessa entità reale.

- **Montaggio:**

Consiste nella fase di codifica del testo, inserendo mark-up tag o introducendo i dati in un database. Questa fase include anche una sottofase di *enhancement* in cui i dati sono trasformati tramite processi algoritmici preliminari per la successiva fase di analisi.

Il montaggio estende l'originale annotazione dei dati con informazioni di base, riferimenti a bibliografie e collegamenti a pagine relazionate.

- **Recupero:**

In questa fase l'informazione è pronta per essere selezionata, cercata e utilizzata. Il recupero è basato su meccanismi di ricerca quali ad esempio le query SQL per i tradizionali database.

In alcuni progetti questa avviene successivamente alla pubblicazione dei dati e dunque questa fase può essere spostata in un punto successivo del ciclo di vita.

- **Analisi:**

L'analisi dell'informazione può assumere diversi significati nella ricerca storica. Varia dal confronto qualitativo e dalla valutazione di risultati delle query all'analisi statistica su set di dati.

- **Presentazione:**

L'ultima fase consiste nella presentazione, in varie forme, dell'informazione storica. Anche se è rappresentata come fase finale, spesso capita che questa avvenga a valle delle fasi precedenti.

La presentazione può assumere varie forme: edizioni di testi elettronici, database online o visualizzazioni dei risultati di ricerca all'interno di un progetto.

CAPITOLO 2. L'ELABORAZIONE AUTOMATICA DEI DOCUMENTI STORICI

Il ciclo descrive concettualmente il percorso dell'informazione storica.

Il progetto di tesi sarà concentrato in due di queste fasi: l'obiettivo è di estrarre il contenuto semantico dei saggi storici tramite l'utilizzo di metadati, processo legato alla fase di *arricchimento*, e successivamente suggerirlo per l'aggiornamento della base di conoscenza, nella fase di *montaggio*. Tale ciclo, inoltre, permette di contestualizzare i problemi, i quali saranno presentati più avanti, che possono intercorrere per via del dominio.

2.2 Classificazione dei documenti storici

Un secondo aspetto interessante del dominio storico sono i molteplici possibili criteri di classificazione per i documenti storici. Si introducono, di seguito, i principali proposti da Meroño et al. [2], con attenzioni particolari per l'ultimo che sarà richiamato durante la fase di estrazione della conoscenza, si vedrà che è possibile applicare diverse tecniche in base al tipo di documento da trattare.

La prima classificazione presentata è la divisione delle fonti in **primarie** e **secondarie**.

Per fonti primarie si considerano i materiali creati al tempo dello studio, ovvero ricerche che riportano i dati esattamente come sono, privi di interpretazione. Ad esempio, gli articoli scientifici, che riportano i risultati di esperimenti, i documenti governativi e documenti legali.

Le fonti secondarie sono derivate dalle primarie. Si tratta di documenti scritti da storici riguardo il passato. Queste descrivono, interpretano, analizzano e valutano le fonti primarie per uno specifico obiettivo o per un particolare pubblico [8]. Ne sono un esempio encyclopedie, articoli riassuntivi e bibliografie.

Un esempio storico, tratto da un articolo dell'università James Cook [8], propone come fonte primaria il diario dell'esploratore e come fonte secondaria derivata il libro sull'esplorazione.

Le narrazioni introdotte nella piattaforma HiStoryGraphia sono tipicamente fonti

CAPITOLO 2. L'ELABORAZIONE AUTOMATICA DEI DOCUMENTI STORICI

secondarie, mentre le fonti sono fonti primarie o secondarie a seconda dei casi.

Una seconda classificazione è legata alla finalità: quando si lavora con documenti storici è importante decidere nelle prime fasi se è necessario modificare un documento secondo un approccio orientato alla *fonte* o all'*utilizzo*.

I documenti orientati alla fonte (*source oriented*) posticipano ogni tipo di operazione sui dati, ad esempio imporre degli standard o classificare, e quindi lasciano ampia libertà alle diverse interpretazioni. I documenti orientati all'obiettivo (*goal oriented*) si focalizzano nel ristrutturare i dati per un fine, diminuendo così le difficoltà legate all'interpretazione e alla gestione delle inconsistenze.

Infine, l'ultima classificazione citata è quella legata alla struttura del file ed è quella che più caratterizza la scelta di approccio per la tecnica di costruzione del semantic web, come vedremo nella parte di *information extraction*.

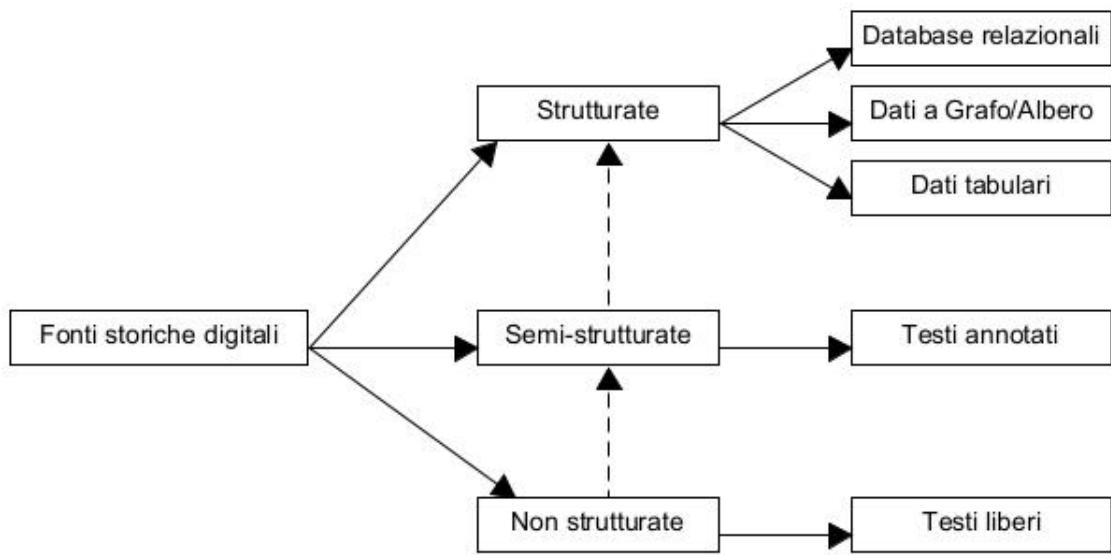


Figura 2.2: Classificazione dei documenti storici in base al livello di struttura [2]

Esistono tre classi legate alla struttura del documento, come rappresentato in figura 2.2. Le linee tratteggiate che le uniscono rappresentano la direzione che avviene normalmente nel workflow di estrazione della conoscenza:

CAPITOLO 2. L'ELABORAZIONE AUTOMATICA DEI DOCUMENTI STORICI

- **Strutturate:**

Questi documenti hanno un modello astratto ben definito che ne rende semplice la comunicazione.

Esempi di questa classe sono i file rdf dataset, file xml e cartelle di lavoro di fogli di calcolo (*spreadsheet workbooks*). Questo genere di file viene normalmente gestito usando database relazionali, graphtree/alberi o tabular data

- **Semi-strutturate:**

Si tratta di una rappresentazione intermedia, generalmente in questa classificazione troviamo dei documenti che contengono annotazioni di alcune parti di testo tramite markup language come xml. In questa categoria un esempio sono gli *annotated corpora*.

- **Non strutturate:**

Si tratta di documenti privi di alcun tipo di data model, per esempio i *raw corpora*.

In figura 2.3 è riportato un esempio della classificazione tratto dall'articolo di Cardoso [9]. L'esempio riporta il confronto di uno stesso contenuto presentato in documenti di diversa struttura. Nel primo caso, il documento a sinistra, si tratta di una fonte non strutturata, priva di alcuna forma di modello del contenuto. Il documento centrale riporta le informazioni rilevanti riguardo gli studenti. Si tratta di una fonte semi strutturata, in formato XML, dove i dati sono racchiusi da descrittori tra parentesi angolari che ne descrivono il tipo di informazione. Infine, nella figura a destra, si presenta la stessa informazione ma in formato strutturato, come potrebbe essere una tabella di un database relazionale. Le righe della tabella rappresentano gli studenti e le colonne i campi con cui vengono descritti.

CAPITOLO 2. L'ELABORAZIONE AUTOMATICA DEI DOCUMENTI STORICI

Unstructured data	Semi-structured data	Structured data																								
<pre>The university has 5600 students. John's ID is number 1, he is 18 years old and already holds a B.Sc. degree. David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.</pre>	<pre><University> <Student ID="1"> <Name>John</Name> <Age>18</Age> <Degree>B.Sc.</Degree> </Student> <Student ID="2"> <Name>David</Name> <Age>31</Age> <Degree>Ph.D. </Degree> </Student> ... </University></pre>	<table border="1"><thead><tr><th>ID</th><th>Name</th><th>Age</th><th>Degree</th></tr></thead><tbody><tr><td>1</td><td>John</td><td>18</td><td>B.Sc.</td></tr><tr><td>2</td><td>David</td><td>31</td><td>Ph.D.</td></tr><tr><td>3</td><td>Robert</td><td>51</td><td>Ph.D.</td></tr><tr><td>4</td><td>Rick</td><td>26</td><td>M.Sc.</td></tr><tr><td>5</td><td>Michael</td><td>19</td><td>B.Sc.</td></tr></tbody></table>	ID	Name	Age	Degree	1	John	18	B.Sc.	2	David	31	Ph.D.	3	Robert	51	Ph.D.	4	Rick	26	M.Sc.	5	Michael	19	B.Sc.
ID	Name	Age	Degree																							
1	John	18	B.Sc.																							
2	David	31	Ph.D.																							
3	Robert	51	Ph.D.																							
4	Rick	26	M.Sc.																							
5	Michael	19	B.Sc.																							

Figura 2.3: Esempio classificazione in base alla struttura [9]

Nel progetto HiStoryGraphia, si possono avere più tipi di documenti, a seconda della sensibilità degli autori, fino a mappe interattive e documenti ipertestuali. Le classificazioni presentate, unite al ciclo di vita della sezione precedente, ci permettono di creare un contesto per il progetto.

Si vedrà che i saggi storici prodotti dagli autori che desiderano immettere nuova conoscenza saranno tipicamente fonti secondarie e spesso avranno connessioni con altre fonti di riferimento da cui è tratta l'informazione.

Inoltre, la classificazione basata sulla struttura ci permette di distinguere i documenti per trattarli diversamente durante l'estrazione del contenuto, tramite procedure specifiche in base al formato del testo semi-strutturato o libero.

2.3 Problemi aperti

L'ultima sezione del capitolo è dedicata alla trattazione di problemi legati all'elaborazione delle fonti storiche.

Basandosi sul lavoro di A. Meroño-Peñuela et al. [2], si presentano di seguito tre tipologie di problemi, contestualizzandoli all'interno delle fasi del ciclo di vita proposto da Boonstra [1] e accompagnandoli da esempi esplicativi.

2.3.1 Problemi legati alle fonti storiche

La prima categoria coinvolge la prima fase del ciclo di vita in Figura 2.1: la fase di creazione.

Questi problemi riguardano il processo di trasformazione in digitale, indipendentemente dal fatto che il processo sia eseguito manualmente o tramite sistema automatico, tipo OCR². Una prima difficoltà potrebbe essere un problema di lettura di alcuni caratteri che sono rovinati o impossibili da leggere, si tratta di un'informazione persa del documento originale.

Inoltre, anche supponendo che la fase di digitalizzazione avvenga senza difficoltà, servono ulteriori informazioni per contestualizzare l'informazione digitalizzata: chi l'ha scritta, con quale finalità, a chi era indirizzata ecc., che, come vedremo, servono per gestire il successivo problema dell'interpretazione.

Di particolare interesse in questa categoria è la necessità di dover scegliere come codificare l'informazione digitalizzata. Come accennato in precedenza, durante il ciclo di vita, nella fase di creazione avviene anche la scelta del software per memorizzare l'informazione digitalizzata. Se si usasse uno strumento scorretto si potrebbe causare una perdita di informazione.

Per chiarire meglio si riporta l'esempio pubblicato da Richard Brath [10] nel 2020 dove spiega che "il contenuto semantico è aggiunto al testo quando le parole sono combinate in sequenza". L'esempio riportato dall'articolo riguarda le poesie di Jack and Jill e Humpty Dumpty:

«*Jack and Jill went up the hill To fetch a pail of water. Jack fell down and broke his crown, And Jill came tumbling after.*» «*Humpty Dumpty sat on a wall Humpty had a great fall all the king's horses and all the king's men Couldn't put Humpty together again.*»

²Sistemi di riconoscimento ottico dei caratteri

CAPITOLO 2. L'ELABORAZIONE AUTOMATICA DEI DOCUMENTI STORICI

In particolare, le frasi:

- Jack and Jill went up the hill.
- Humpty Dumpty sat on a wall.

Contengono informazione sul fatto che Jack, Jill e Humpty Dumpty si trovano ad una certa altezza. Se si utilizzasse un sistema di conversione considerando solo le singole parole e non il loro ordine, come ad esempio delle *word cloud*, perderemmo l'informazione semantica legata alla sequenza delle parole, ottenendo per esempio il risultato in Figura 2.4.



Figura 2.4: Word cloud per "Jack and Jill went up the hill" e "Humpty Dumpty sat on wall" generata tramite: <https://www.wordclouds.com>

Questa conversione genererebbe una perdita di informazione legata all'altezza dei soggetti e dunque impedirebbe di rispondere a semplici domande come "Dumpty è in alto?", "Possono Jack e Jill cadere?" in quanto si tratta di una informazione semantica andata persa.

È quindi importante, in base alle finalità del documento, cercare di preservarne quanta più informazione possibile, ad esempio, memorizzandone anche il formato originale.

2.3.2 Problemi legati alle relazioni tra le fonti

Nella ricerca storica normalmente si attingono le informazioni da più fonti, ne consegue la necessità di dover collegare tra di loro i concetti che si riferiscono allo stesso elemento della realtà.

Si tratta di un problema della fase di arricchimento in Figura 2.1.

Per esempio, Meroño et al. [2] descrivono una situazione in cui bisogna collegare, o unire, due registri diversi: partendo dunque da due insiemi di concetti, bisogna unificare i riferimenti alle stesse entità del mondo reale. Supponendo di avere nel primo registro il nome “*Lars Erikson*” e nel secondo il nome “*Lars Eriksson*”, si tratta della stessa persona o di soggetti diversi? Non essendo un match perfetto diventa difficile determinare se unificare o separare le due entità. Inoltre, può capitare che due persone differenti siano riferite con lo stesso nome.

Infatti, il dominio storico presenta un particolare problema relazionato al precedente, ovvero che le entità sono delimitate da un contesto spaziale e temporale: i confini di un territorio variano nel corso degli anni e due persone con lo stesso nome potrebbero esser vissute in zone e periodi completamente differenti.

Si pensi al termine “Roosevelt”, che può riferirsi a diverse figure famose negli anni: Theodore Roosevelt (1858 – 1919), ventiseiesimo presidente degli Stati Uniti e Franklin D. Roosevelt (1882 - 1945), trentaduesimo presidente degli Stati Uniti.

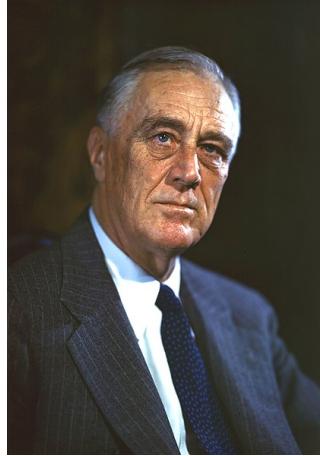
Data la semplice frase “*Roosevelt è stato eletto governatore di New York nel 1928*” è necessario contestualizzare la parola “Roosevelt” per determinare a quale delle due entità si riferisce. In questo caso si può determinare il riferimento a Franklin in quanto l’evento è occorso nel 1928 e Theodore morì nel 1919.

Ne consegue che la rappresentazione del contesto ha un ruolo particolarmente importante in questo dominio perché può fornire delle informazioni preziose legate all’identificazione dei concetti corretti.

CAPITOLO 2. L'ELABORAZIONE AUTOMATICA DEI DOCUMENTI STORICI



Fotografo: Leon A. Perskie
Digitalizzazione: FDR Presidential Library & Museum



(a) F.D.Roosevelt(1882-1945)



(b) T.Roosevelt (1858–1919)

Fotografo: Pach Bros
Digitalizzazione: Divisione Stampe e Fotografie della Biblioteca del Congresso

Figura 2.5:

- (a) Franklin Delano Roosevelt, 32° presidente degli Stati Uniti [11]
- (b) Theodore Roosevelt, 26° presidente degli Stati Uniti [12]

2.3.3 Problemi legati all'analisi

L'ultima categoria introdotta tratta problemi legati alla fase di analisi del ciclo di vita, in Figura 2.1, o più in generale quando i dati vengono utilizzati.

Nella ricerca storica il significato dei dati non può esistere senza interpretazione [1]. Differenti interpretazioni possono esistere in riferimento allo stesso dato. Infatti, Wertsch e Poleman [5] sostengono che nella storia è possibile che due esperti storici, lavorando con le stesse fonti e utilizzando gli stessi metodi, possano raggiungere differenti ma legittime e sensate interpretazioni dello stesso evento passato.

Per approfondire questo problema, si riporta parte del contenuto pubblicato da Kolikant et al. nel 2019 [13]. Lo studio nasce con l'obiettivo di introdurre gli studenti alla multi-prospettiva e alla natura interpretativa della storia.

L'esperimento riguarda un diverso punto di vista sviluppato tra due etnie in conflitto, condotto nello stato di Israele dove gli Ebrei costituiscono la maggioranza

CAPITOLO 2. L'ELABORAZIONE AUTOMATICA DEI DOCUMENTI STORICI

della popolazione e gli Arabi, che sono circa il 20% della popolazione totale della città, sono la minoranza più grande.

L'esperimento consiste nel coinvolgere studenti arabi ed ebrei, i quali sono soliti frequentare scuole separate, in un'attività collaborativa riguardo un evento rilevante per il passato travagliato comune dei due gruppi: la pubblicazione britannica del Libro Bianco di Churchill nel 1922.

I libri bianchi della Palestina consistono in una serie di leggi pubblicate dalle autorità Britanniche. In particolare, il primo libro bianco del 1922 fu pubblicato con lo scopo di ridurre le tensioni della zona in quegli anni. Il contenuto di maggiore rilevanza era il diritto del popolo ebraico di ritornare nelle terre ancestrali, all'epoca popolate dagli Arabi [14].

L'esperimento consisteva nel sottoporre gli studenti alla lettura di narrazioni storiche contenenti almeno due diverse prospettive dell'evento.

L'attività fu divisa in due fasi: nella prima si crearono coppie etnicamente omogenee, formando gruppi di Arabi e gruppi di Ebrei, i gruppi vennero dunque sottoposti alla lettura delle narrazioni e successivamente dovettero rispondere alle seguenti domande:

“Cosa promisero gli Inglesi agli Arabi e agli Ebrei nel libro bianco del 1922? Come entrambe le fazioni risposero a tale documento e perché? Per quale motivo il libro bianco è significativo per gli Inglesi?”

Nella seconda fase vennero creati gruppi misti, ognuno composto da due coppie etnicamente omogenee, gli studenti dovettero leggere reciprocamente le risposte date e discuterne.

I risultati dimostrarono che nonostante entrambi i gruppi partissero dalle stesse fonti, nella prima fase vennero posti in rilievo i pensieri allineati con la propria cultura e di rifiutare o ignorare quelli opposti. Per esempio, le coppie Arabe diedero particolarmente attenzioni a frasi che li caratterizzavano come vittime: “Gli Arabi furono ingannati”, “La terra gli fu sottratta” ecc. Al contrario tra le coppie

CAPITOLO 2. L'ELABORAZIONE AUTOMATICA DEI DOCUMENTI STORICI

Ebrei prevalse il pensiero secondo cui gli Arabi si rifiutavano di condividere.

Tuttavia, in seguito all'unione dei gruppi nella seconda fase, si ottenne un allineamento dei due pensieri che si trovarono d'accordo con la frase:

"La popolazione Araba non fu d'accordo con il libro bianco perché fu percepito come l'inizio del processo in cui le terre gli venivano sottratte"

Questo esperimento dimostra come l'interpretazione sia un fattore intrinseco dell'informazione storica. Di conseguenza, è fondamentale per le rappresentazioni semantiche considerare il contesto e la struttura dei sorgenti. In caso contrario si potrebbero ottenere delle contraddizioni o perdite di informazioni. Ad esempio, nel caso precedente sarebbe complicato determinare se la terra fosse stata condivisa o sottratta partendo esclusivamente dalle fonti storiche usate nell'esperimento.

Nel progetto HiStoryGraphia, le narrazioni rappresentano punti di vista specifici degli autori; l'interconnessione basata sulla rappresentazione semantica permette di mettere a confronto interpretazioni differenti di argomenti correlati.

In conclusione, sono state presentate tre categorie di problemi che intercorrono durante l'elaborazione automatica dei testi storici. Questi, sono accomunati dall'importanza legata alla rappresentazione del contesto e della preservazione della fonte originale da cui è estratto il contenuto semantico.

Tra questi, sono particolarmente interessanti i problemi legati alle relazioni tra le fonti, i secondi trattati precedentemente. Vedremo, nei capitoli successivi, come queste difficoltà, riscontrate in diversi lavori simili, abbiano portato all'introduzione nei sistemi di una tecnica denominata *gazetteer* e quali benefici sono legati alla sua applicazione.

2.4 Conclusioni di capitolo

In questo capitolo si è trattata una contestualizzazione dell'elaborazione automatica di testi in dominio storico. La prima sezione ha presentato il lavoro di Boonstra et al. [1] legato al trattamento dell'informazione storica, in particolare il ciclo di vita che la coinvolge. In seguito, sono state introdotte una serie di classificazioni rilevanti per il settore, di cui l'ultima, la classificazione legata alla struttura del documento, sarà ripresa in seguito per differenziare i documenti e il tipo di lavoro necessario per estrarne il contenuto.

Si sono analizzate tre categorie di problemi legati a questo dominio. La prima categoria espone le difficoltà che avvengono in seguito ad una povera rappresentazione del materiale. La seconda categoria presenta le problematiche legate alla disambiguazione dei concetti storici e come questi siano fortemente legati al contesto. Infine, la terza categoria, presenta il problema dell'interpretazione dell'informazione storica e come questa possa causare una perdita di informazione se non si rappresenta il contesto da cui è tratta.

Capitolo 3

Information extraction (IE)

L'idea di fornire ai computer l'abilità di elaborare il linguaggio umano esiste dalla nascita dei computer stessi [15]. Il campo di ricerca dedito alla realizzazione di sistemi in grado di comprendere il linguaggio umano è denominato *Natural Language Processing (NLP)*. Esso viene definito da Singh [16] come:

«*L'uso di metodi computazionali per elaborare la forma scritta o parlata di testi liberi che agiscono come forma di comunicazione comunemente usata tra umani*»

Si tratta dunque di un settore molto vasto, che a sua volta si ramifica in diverse specializzazioni. Tra queste, l'estrazione di informazione (IE – Information extraction) è un ramo in cui certi tipi di informazioni devono essere riconosciute ed estratte dal testo [17]. Si tratta di tecniche per riconoscere delle informazioni rilevanti al fine di presentarle in forma strutturata [18] e quindi facilmente utilizzabili da una macchina.

La combinazione del successo legato all'estrazione automatica di informazioni con l'emersione del più recente Semantic Web ha portato alla nascita di sotto aree, tra cui l'estrazione di informazioni basate su Ontologie (OBIE - Ontology Based Information Extraction). Per dare una definizione formale, citiamo Wilamura [17] che la definisce come:

« *Un sistema che elabora un testo in linguaggio naturale, libero o semi-strutturato, at-*

traverso meccanismi guidati da ontologie per estrarre determinati tipi di informazioni e presentarli utilizzando ontologie »

Nel contesto informatico, il termine ontologia assume il significato di una rappresentazione concettuale. Secondo Guarino [19], un'ontologia è:

Un prodotto ingegneristico, costituito da specifici vocabolari usati per descrivere una certa realtà.

L'idea generale del suo utilizzo è di rappresentare e organizzare tutto ciò che esiste nel mondo all'interno di una gerarchia di categorie [20].

Il seguente capitolo presenterà una serie di task per l'estrazione dell'informazione rilevanti ai fini del progetto.

La prima sezione tratta un task per gestire documenti semi strutturati, in quanto, come vedremo nel capitolo successivo, gli articoli da analizzare avranno una struttura di base che potremo sfruttare tramite l'applicazione di un Wrapper.

Seguono due sezioni dedicate all'estrazione dei contenuti da testi liberi: la prima riguarda il riconoscimento delle entità presenti in un testo mentre la seconda ricerca come queste entità si relazionano tra di loro.

La finalità della tesi è di presentare un'architettura che integri moduli già esistenti. L'obiettivo di queste sezioni è di presentare i task, il loro funzionamento e i concetti chiave necessari per il capitolo sull'implementazione del modulo di estrazione della conoscenza.

3.1 Wrapping

Le metodologie linguistiche utilizzate per i testi liberi possono essere difficili da applicare o inefficaci su testi altamente strutturati come le pagine web prodotte da database [21].

Questo genere di pagine sono normalmente generate tramite script automatici che vengono riempiti con i dati del database e dunque hanno una struttura simile



(a) Pagina web

```
<HTML><TITLE>Some Country Codes</TITLE><BODY>
<B>Some Country Codes</B><P>
<B>Congo </B> <I>242 </I><BR>
<B>Egypt </B> <I>20 </I><BR>
<B>Belize </B> <I>501 </I><BR>
<B>Spain </B> <I>34 </I><BR>
<HR><B>End</B></BODY></HTML>
```

(b) Codice della pagina

Figura 3.1: Pagina web contenente i prefissi telefonici di alcuni paesi [22]

tra di loro.

Per esempio, immaginando di avere a disposizione una pagina web, come in figura 3.1(a), contenente i prefissi telefonici di ogni paese, si potrebbe sfruttare la struttura, mostrata in figura 3.1(b), per estrarre i paesi e i loro rispettivi prefissi; ottenendo l'informazione strutturata in figura 3.2(a). Questo è permesso dalla ripetizione dei tag all'interno della costruzione della pagina stessa che ci permette di realizzare una procedura come in figura 3.2(b).

Questo task assume il nome di *wrapping* e viene definito come:

«*Una procedura per estrarre un particolare contenuto di una risorsa* » [22]

Consiste nell'individuare un insieme di regole che consentono l'estrazione sistematica di specifici record di dati dalle pagine [23]. In altre parole, si tratta di una procedura applicabile quando si opera con documenti provvisti di una

$\{\langle \text{Congo}, 242 \rangle, \langle \text{Egypt}, 20 \rangle, \langle \text{Belize}, 501 \rangle, \langle \text{Spain}, 34 \rangle\}$

(a) Risultato estratto dalla procedura 3.2(b)

```

ExtractCCs(page  $P$ )
    skip past first occurrence of <P> in  $P$ 
    while the next occurrence of <B> is before the next occurrence of <HR> in  $P$ 
        for each  $\langle \ell_k, r_k \rangle \in \{(\langle B \rangle, \langle /B \rangle), (\langle I \rangle, \langle /I \rangle)\}$ 
            extract from  $P$  the value of the next instance of the  $k^{\text{th}}$  attribute
            between the next occurrence of  $\ell_k$  and the subsequent occurrence of  $r_k$ 
    return all extracted tuples

```

(b) Procedura per generare il risultato (a)

Figura 3.2: Risultato e procedura per ottenere l'informazione contenuta nella figura 3.1

[22]

certa struttura tale che possa permettere di costruire delle regole per estrarne il contenuto senza analizzare la semantica del testo.

Il wrapping è una soluzione applicabile ogni qual volta si operi con documenti semi strutturati. Nel progetto HiStoryGraphia, i saggi creati dagli utenti, durante l'inserimento di nuova conoscenza, saranno dotati di una struttura e quindi, sarà possibile applicare questa metodologia per realizzare una estrazione di una parte del contenuto informativo da essi.

3.2 Estrazione di entità

Esistono molteplici nomenclature per i task legati al trattamento di entità. Per entità, o con nome Named Entity (NE), si intende una sequenza di parole che identificano una qualche entità del mondo reale, ad esempio “Barack Obama” [24].

Il riconoscimento e la classificazione di tali entità è uno dei task più frequenti nell'information extraction, nonché la base di partenza per task successivi quali l'estrazione di relazioni o eventi.

Il processo tradizionale di riconoscimento delle entità prevede il susseguirsi di due task distinti: il primo, il riconoscimento e la classificazione di entità all'interno di un testo, prende il nome di Named Entity Recognition (NER), mentre il secondo, l'assegnazione di entità ambigue ad un riferimento univoco viene detto, Named entity Disambiguation (NED).

Tuttavia, l'attuale disponibilità di ampie basi di conoscenza accessibili pubblicamente ha consentito di fondere i due task in uno solo, denominato Entity Linking (EL). Esso mira a collegare un parte del testo alla voce corrispondente in una risorsa come DBpedia o Wikipedia [25].

Di seguito, si affronteranno singolarmente i due task di NER e NED, per spiegarne il funzionamento con esempi. Infine, si vedrà una soluzione particolarmente diffusa nei domini specifici, tipo gli umanitari, detta Gazetteer.

3.2.1 Task

Named entity recognition

Il Named Entity recognition and classification (NERC) [26], recentemente riferito semplicemente come NER, si occupa di identificare entità generiche, tipo persone, località e organizzazioni, o specifiche di un dato dominio, ad esempio malattie, farmaci, sostanze chimiche, proteine [27].

In figura 3.3 sono presentati due esempi, per la lingua inglese e italiana, in seguito al riconoscimento e classificazione di entità al loro interno. La frase italiana "*Mario, che abitava a torino, si unì a Medici Senza Frontiere*" contiene al suo interno tre entità: una di tipo PER, Persona, associata a Mario, una di tipo LOC, Località, associata a Torino e l'ultima di tipo ORG, Organizzazione, riferita a Medici Senza Frontiere.

[ORG U.N.] official [PER Ekeus] heads for [LOC Baghdad]

(a) Esempio ENG

The image shows the English sentence "Mario , che abitava a Torino , si unì a Medici Senza Frontiere" with three entities annotated: "Mario" is labeled PER (Person), "Torino" is labeled LOC (Location), and "Medici Senza Frontiere" is labeled ORG (Organization). The labels are placed above the corresponding words in the sentence.

(b) Esempio ITA

Figura 3.3: Esempi NER. Classi: PER (Persona), LOC (Località), ORG (Organizzazione)

(a) Esempio ufficiale CoNLL-2003 [28]

(b) Esempio in lingua italiana tramite la risorsa Stanza [29]

Le entità definite come generiche derivano da dataset costruiti al fine di addestrare e testare risorse in tema NER. Ad esempio, un famoso dataset inglese è il CoNLL¹-2002/2003, pubblicato nel 2002.

CoNLL è una conferenza annuale organizzata da SIGNLL² che tratta argomenti in ambito NLP. Nel 2002 pubblicarono un dataset contenente le seguenti classi:

- **PER:** Persone
- **ORG:** Organizzazioni
- **LOC:** Località
- **MISC:** Misto, tutto ciò che non rientra nelle altre classi

Tuttavia, si trattava di una risorsa per le lingue inglese e tedesco. Per la lingua italiana nacque EVALITA, campagna di valutazione periodica del NLP e degli strumenti vocali.

L'edizione del 2007, denominata EVALITA 2007, contenne un task dedicato al NER e fu reso disponibile un corpus annotato in lingua italiana I-CAB³.

L'introduzione dei dataset e dei tag utilizzati sarà ripresa nella descrizione dei

¹Conference on Computational Natural Language Learning

²Special Interest Group on Natural Language Learning

³Italian Content Annotation Bank - <https://ontotext.fbk.eu/icab.html>

tool integrati nella piattaforma HiStoryGraphia in quanto si tratta di risorse addestrate su dataset simili e dunque utilizzano gli stessi tag per classificare le entità.

Esistono diversi formati per rappresentare le entità riconosciute in un testo.

Tra questi ne citiamo due, i formati **BIO** e **BIOES/BILOU**, in quanto sono quelli utilizzati dalle risorse che verranno integrate. La rappresentazione prevede una struttura simile al formato CoNLL, usato nei dataset, in cui le frasi sono rappresentate in una forma tabellare in cui nelle righe si hanno le parole della frase e nelle colonne i vari descrittori di tale parola.

I formati *BIO* e *BIOES* possono esser utilizzati come ulteriore colonna descrittiva nel formato CoNLL o da soli, come negli esempi in figura 3.4, dove la singola colonna descrittiva ne indica il tag NER associato.

Alex	B-PER	Alex	S-PER
sta	O	sta	O
andando	O	andando	O
a	O	con	O
Los	B-LOC	Marty	B-PER
Angeles	I-LOC	A.	I-PER
in	O	Rick	E-PER
California	B-LOC	a	O
		Los	B-LOC
		Angeles	E-LOC

(a) Esempio BIO

(b) Esempio BIOES

Figura 3.4: Esempi rappresentazioni in formati BIO e BIOES [30]

L'esempio, in figura 3.4, associa ad ogni parola della frase l'eventuale tag NER se è un'entità nominale, ad esempio Alex è classificato come persona. Oltre al tag sono presenti delle lettere che aggiungono informazioni per riconoscere quando le entità sono composte da più parole, utilizzando il seguente schema di lettere:

- **B**: L'inizio di una entità
- **I**: Una parola contenuta all'interno di un'entità
- **O**: Una parola che non appartiene ad un'entità
- **E\L**: L'ultima parola che compone un'entità
- **S\U**: Un'entità composta da una singola parola

La differenza tra i due formati è la presenza di due lettere in più per maggior precisione di rappresentazione. Per esempio, nella prima frase la parola Alex è associata alla lettera B, dunque bisogna controllare la parola successiva per determinare se è l'inizio di una entità di più parole o se si tratta di una entità a singola parola. Al contrario nella seconda frase, l'utilizzo della lettera S, la classifica subito come entità a singola parola.

Le informazioni introdotte riguardo il task di NER, saranno riprese nella descrizione dei componenti di HiStoryGraphia, dove saranno impiegate alcune risorse per l'estrazione delle entità che utilizzeranno proprio i tag sopra descritti e i risultati saranno descritti secondo i formati BIO/BIOES. Questo task sarà utile per identificare le entità presenti nei testi liberi e associarle a classi generali di riferimento, in particolare per le situazioni in cui sarà necessario definire nuove entità nella base di conoscenza e quindi associarle ad una classe concettuale.

Named entity disambiguation

Il task di Named entity Linking o Disambiguation consiste nel connettere le entità estratte dal testo a dei riferimenti univoci all'interno di basi di conoscenza, tipo Wikipedia.

Per esempio, in figura 3.5, la frase inglese *Floyd rivoluzionarono il rock con The Wall* contiene diverse entità ambigue: Floyd rappresenta la band Pink Floyd ma in altri contesti potrebbe rappresentare un nome di persona, analogamente il rock in inglese significa anche roccia e Wall può far pensare al famoso muro di Berlino

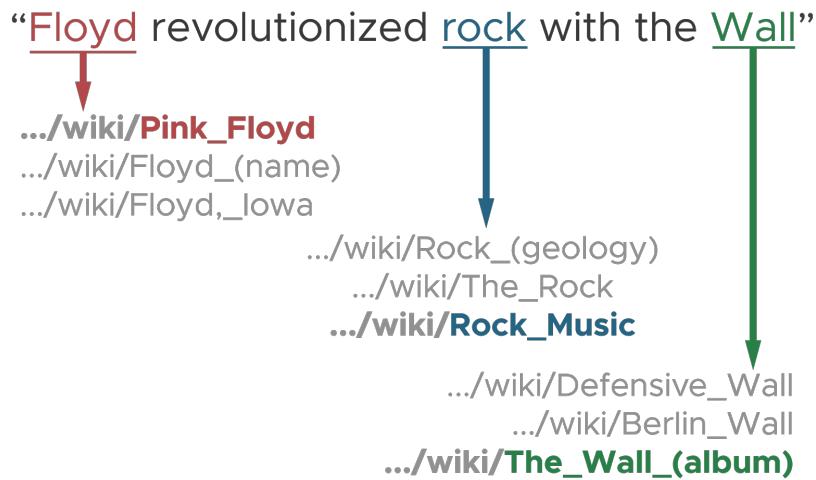


Figura 3.5: Esempio NED [31]

ma nel contesto della frase rappresenta il disco, The Wall, della band rock inglese. La finalità del task è quindi di riconoscere i concetti nascosti dietro alle entità e differenziarli nelle situazioni ambigue.

Esistono una vasta varietà di soluzioni al problema, Martinez et al. [32] raggruppano queste tecniche in cinque categorie di macro livello in base al tipo di caratteristiche che prendono in considerazione.

In particolare, citiamo la prima di queste, la menzione, in quanto sarà utilizzata nella fase di implementazione:

- **basata sulla menzione**

Il termine menzione si riferisce alla sequenza di parole riconosciute come entità all'interno del testo. Ad ogni menzione estratta corrispondono uno o più candidati all'interno della base di conoscenza, cioè gli identificatori univoci per diversi concetti assegnabili a tale porzione di testo.

Ad esempio, nella figura 3.5, "Floyd" è la menzione e "Pink_Floyd", "Floyd_(name)" e "Floyd,_Iowa" sono i candidati, cioè gli identificatori in Wikipedia.

La categoria "*basata sulla menzione*" raggruppa tutte le tecniche di disambiguazione che si basano sulle caratteristiche della menzione: il testo, la

presenza di lettere maiuscole iniziali, la presenza di altre menzioni che si sovrappongono nel testo, la presenza nello stesso testo di forme abbreviate di menzioni, ecc..

Ad esempio, data la menzione «*Bryan L. Cranston*» estratta dal testo e i candidati «*Bryan L. Reuss*» e «*Bryan Cranston*» contenuti nella base di conoscenza. Una tecnica di disambiguazione di questa categoria potrebbe effettuare una misura sulla distanza di edit, o distanza di Levenshtein, che corrisponde al minimo numero di modifiche necessarie per ottenere una sequenza di caratteri partendo da un'altra sequenza. Per esempio, date le parole *bar* e *biro*, la distanza di Levenshtein equivale a 2. Infatti, servono due modifiche per ottenere la parola *biro*: sostituire la 'a' con la 'i' e inserire una 'o' finale. Nell'esempio precedente di «*Bryan L. Cranston*», otterremmo come risultati, rispettivamente, 7 e 3.

Un secondo esempio relativo a questa categoria è la presenza di menzioni abbreviate. Si immagini la presenza di "*Jimmy Wales*" nell'introduzione di un articolo e successivamente, all'interno del testo, l'entità "*Wales*". È possibile che si tratti di un riferimento allo stesso concetto.

Tuttavia, questo specifico caso di disambiguazione può rivelarsi poco efficace se preso singolarmente. Si prenda il concetto di *Enzo Ferrari*, il fondatore del brand *Ferrari*. È evidente che in questo caso si tratta di concetti diversi che facilmente potrebbero comparire all'interno dello stesso testo. Spesso questo genere di caratteristiche si utilizzano in combinazione con altre per evitare questo tipo di errori.

Un modulo della piattaforma HiStoryGraphia, denominato Storytelling2Knowledge, incorpora al suo interno diverse risorse, alcune delle quali utilizzano dei propri sistemi di disambiguazione. Tuttavia, il combinare risultati da diverse fonti può portare a situazioni in cui esistano dei risultati in contraddizione e, quindi, è necessario decidere quale accettare. A tal scopo, sarà presentata una strategia di scelta, o euristica, legata a questa categoria, che quindi sfrutterà

una caratteristica delle menzioni trovate nel testo per decretare il candidato migliore.

3.2.2 Tecniche

Gazetteer

La tecnica più diffusa per i task legati al riconoscimento e alla disambiguazione delle entità sono i Gazetteer. Questa tecnica si rivela particolarmente importante in domini come quelli umanistici, dove può capitare di dover estrarre entità non presenti nelle basi di conoscenza comuni.

Questo problema viene affrontato anche da Rovera et al. [25] per estrarre entità da bollettini di guerra, in quanto risorse come Wikipedia non contenevano, ad esempio, i nomi dei soldati. Inoltre, spesso tali soldati venivano nominati usando soprannomi o alias, rendendo più complicato il riconoscimento e la disambiguazione.

L'idea di base del gazetteer è di avere una lista di termini, che possono essere associati a delle classi o a dei riferimenti in basi di conoscenza, e di cercare tali termini all'interno del testo: se una sequenza di parole soddisfa i criteri di confronto allora viene segnata come menzione e associata alla classe o al riferimento rappresentato nel gazetteer.

In figura 3.6 è presentato un esempio tratto dal lavoro di Busemann [33]. Il gazetteer contiene diverse righe, una per ogni candidato. Ognuna è rappresentata da quattro informazioni: la menzione da trovare all'interno del testo, il tipo della classe associata, la lingua di riferimento e l'identificativo nella base di conoscenza.

Ogni volta che compare una sequenza di parole che corrisponde ad un elemento della prima colonna viene riconosciuta l'entità associata a quella riga e rappresentata dall'identificativo della base di conoscenza dell'ultima colonna.

CAPITOLO 3. INFORMATION EXTRACTION (IE)

```
cote d'ivoire | GTYPE:gaz_country | LANG:french | CONCEPT:c_ivory_coast
côte d'ivoire | GTYPE:gaz_country | LANG:french | CONCEPT:c_ivory_coast
elfenbeinkueste | GTYPE:gaz_country | LANG:german | CONCEPT:c_ivory_coast
elfenbeinkuste | GTYPE:gaz_country | LANG:german | CONCEPT:c_ivory_coast
elfenbeinküste | GTYPE:gaz_country | LANG:german | CONCEPT:c_ivory_coast
ivory coast | GTYPE:gaz_country | LANG:english | CONCEPT:c_ivory_coast
```

Figura 3.6: Voci di un gazetteer per riconoscere un concetto in diverse lingue [33]

Esistono molteplici costruzioni per i gazetteer, in base alle necessità del problema: una semplice implementazione potrebbe essere una lista di termini da associare ad una classe, permettendo così di riconoscere entità specifiche di un dato dominio, come potrebbe capitare nel settore sanitario in cui si hanno termini specifici e standard comuni [34], ad esempio UMLS⁴, oppure sistemi più complessi come quello descritto in precedenza in figura 3.6 che associa per ogni termine un identificativo all'interno di una base di conoscenza.

Quando si adoperano gazetteer di grosse dimensioni, aumentano le possibilità che più menzioni, e quindi più candidati, vengano attivate per lo stesso insieme di parole; può esser necessaria una fase di disambiguazione.

A tal fine, per alcune implementazioni, può essere necessario aggiungere un campo descrittivo ai termini del gazetteer contenente una serie di parole legate al contesto.

Il gazetteer è una soluzione efficace per casi in cui si lavora con entità difficilmente riconoscibili, ad esempio per via della loro assenza su basi di conoscenza pubbliche, o che in un determinato dominio assumono un preciso significato, semplificando quindi il processo di disambiguazione.

Il progetto HiStoryGraphia raccoglie contenuti storici locali e quindi difficilmente presenti all'interno di grosse basi di conoscenza pubbliche. In questo contesto, l'inserimento di gazetteer, che contiene principalmente entità di dominio, può fornire una soluzione importante nel riconoscere le entità presenti nei saggi creati dagli utenti.

⁴Unified Medical Language System

I task di riconoscimento e disambiguazione delle entità nominali, insieme alla tecnica del gazetteer, saranno ripresi nella descrizione del modulo Storytelling2Knowledge, nel quale saranno integrate differenti risorse per combinarne i risultati. In particolare, si utilizzeranno risorse open source per l'estrazione di entità generiche e si integrerà la funzionalità del gazetteer per eventuali classi più specifiche.

3.3 Estrazione di relazioni

L'estrazione di relazioni focalizza l'attenzione nella ricerca delle connessioni tra le entità all'interno di un documento testuale. Di solito, questo processo avviene in seguito ad una precedente fase di analisi finalizzata al riconoscimento, ed eventualmente alla disambiguazione, delle entità presenti.

Le relazioni sono generalmente rappresentate da un termine centrale, generalmente un verbo, che rappresenta l'informazione semantica della relazione, e due o più argomenti, cioè concetti coinvolti in un legame rappresentato dal verbo stesso.

Si distinguono principalmente due tipi di relazioni, in base al numero di argomenti che coinvolgono:

– **Binarie:**

Si definiscono relazioni binarie se coinvolgono solo due elementi. Un classico esempio sono le triple soggetto - verbo - oggetto, in cui il verbo indica il legame tra le due entità.

Ad esempio, data la frase "*Barack Obama è nato alle Hawaii*", si otterrebbe la relazione "*è nato*" che lega i concetti "*Barack Obama*" e "*Hawaii*".

– **N-arie:**

Si definiscono relazioni n-arie se coinvolgono tre o più elementi, o argomenti. Si usano per situazioni in cui il tipo di verbo descrive situazioni, o eventi,

in cui sono coinvolti più partecipanti.

Ad esempio, data la frase "*Barack Obama ha sposato Michelle Obama a Chicago*", ci sono tre entità coinvolte dal verbo "*ha sposato*": "*Barack Obama*", "*Michelle Obama*" e il luogo in cui è avvenuto il fatto, "*Chicago*".

Un approccio emerso nell'ultimo decennio, legato al riconoscimento di relazioni, è Open IE (Open information extraction). Questo task, partendo da testi in linguaggio naturale, estrae una o più proposizioni rappresentate da un verbo e i suoi argomenti [36], spesso rappresentate sotto forma di tuple.

Per esempio, data la frase in inglese "*she born in a small town*" ("lei è nata in una piccola città"), tratta dal sito di Stanford NLP Group [35], la relazione riconosciuta riguarda gli argomenti "*she*" (lei) e "*town*" (città), legati dal verbo "*born in*" (nascerre in). L'estrazione di tale relazione crea un legame di "essere nata in" tra lei e la città, rappresentata dalla tupla "*(she, born in, town)*" (lei, nata in, città).

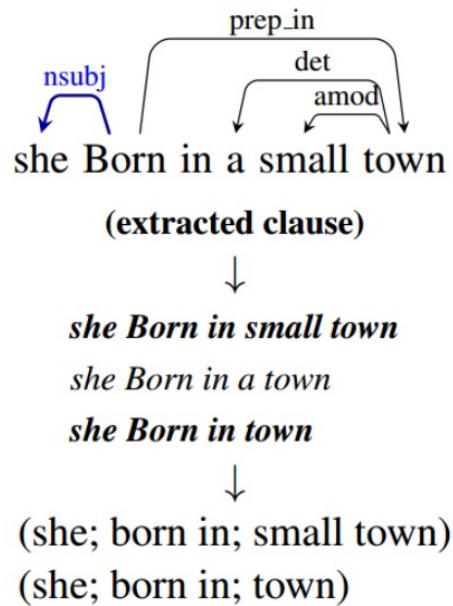


Figura 3.7: Esempio Open IE [35]

Questo genere di approcci sfruttano la struttura della frase, in particolare le dipendenze tra le parole che la compongono: data una frase è possibile estrarne una rappresentazione strutturale delle parole e di come queste sono in relazione al suo interno.

CAPITOLO 3. INFORMATION EXTRACTION (IE)

In particolare, la rappresentazione a dipendenze mette in relazione coppie di parole in un rapporto di superiore e inferiore, generando quindi un legame di dipendenza tra le due parole. Si tratta di relazioni binarie e direzionate, che possono essere associate a tag, che rappresentano il tipo di legame che intercorre tra le due parole. Ad esempio, data la frase "*Mario Biondi ama la pizza*", una rappresentazione della sua struttura a dipendenza è data dalla figura 3.8.

I tag, o tipi, attribuibili alle relazioni di dipendenza variano in base alla lingua, l'esempio in figura 3.8 pone il verbo "ama" come superiore di entrambi "Mario" e "pizza" e li pone in relazione, rispettivamente, di soggetto, rappresentato dal tag *nsubj*, e oggetto, rappresentato dal tag *obj*, della parola "ama".

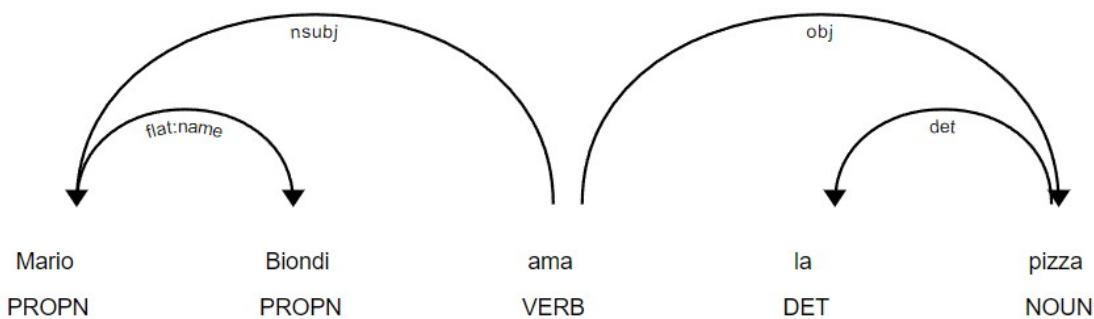


Figura 3.8: Esempio albero a dipendenze per la frase "Mario Biondi ama la pizza"

Questo genere di rappresentazione permette di poter costruire regole specifiche per la lingua ed estrarre gli argomenti legati al verbo in base al tipo di dipendenza che li relaziona.

Grazie a questo meccanismo OpenIE, a differenza dei tradizionali metodi IE, non si limita ad un ristretto insieme di relazioni conosciute in anticipo, ma al contrario estrae tutte quelle presenti in un testo. In questo modo facilita le applicazioni indipendentemente dal dominio [37].

Per via della loro rappresentazione del risultato, i sistemi di questo tipo si applicano meglio ad estrarre relazioni binarie, dove il verbo indica il tipo di legame che intercorre tra i suoi argomenti come nel caso precedente di *ha-sposato* (*Barack Obama, Michelle Obama*).

L'attività di estrazione delle relazioni è un tema problematico in relazione con la lingua italiana, a causa della scarsità di risorse open source accessibili. Essendo il progetto orientato all'integrazione di moduli già esistenti, per soppiare a questa mancanza, sarà costruito un semplice script ispirato da uno basato sulla lingua inglese e riadattato a quella italiana. Questo si baserà sulle idee dei sistemi di Open IE, ovvero l'utilizzo delle relazioni a dipendenza tra le parole della frase e la costruzione di semplici regole per riconoscere i tag principali, come il soggetto e l'oggetto.

3.4 Conclusioni di capitolo

In questo capitolo è stata presentata una panoramica generale, riguardo l'estrazione di informazione da documenti testuali, dei concetti che saranno ripresi nella successiva parte di progettazione e implementazione del modulo Storytelling2Knowledge. La prima sezione ha trattato il task del wrapping, che consiste nel costruire una procedura per estrarre informazioni da documenti dotati di una certa struttura e che per via della stessa possono dare scarsi risultati con le metodologie utilizzate per i testi liberi. Seguono due sezioni per l'estrazione di conoscenza da testi non strutturati. Nella seconda sono stati descritti due task legati all'estrazione di entità, rispettivamente, il riconoscimento e la disambiguazione. Questi task permettono di estrarre, classificare e associare ad identificatori univoci, presenti in basi di conoscenza, le entità presenti in un testo libero. In seguito, è stata vista una tecnica particolarmente interessante per l'estrazione di entità in domini specifici, il gazetteer, che consiste nel raccogliere una lista di termini rilevanti da cercare all'interno dei testi. Infine, l'ultima sezione del capitolo, riguardava l'estrazione delle relazioni. Si è vista la struttura a dipendenze tra le parole e come questa sia utile al fine di costruire regole per il task Open IE che permette di estrarre relazioni rappresentate da un verbo e i suoi argomenti, indipendentemente dal dominio.

Capitolo 4

HiStoryGraphia

Questo capitolo presenta una panoramica del progetto HiStoryGraphia, la missione da cui nasce la ricerca in tesi al suo interno.

Esiste una mole di conoscenze storiche nascoste nei territori locali: elementi artistici, come le committenze di dipinti, eventi storici, quali le conquiste o le migrazioni, informazioni su divisioni politico-amministrative, come le ripartizioni di territori in comuni o in diocesi, ecc. Tutti i fenomeni e gli accadimenti sono interconnessi, ma, nelle analisi, le discipline specialistiche spesso hanno difficoltà a individuare le connessioni e a presentare in modo sinergico i fatti [38].

HiStoryGraphia(HSG) è un progetto di studio in ambito storico nato presso l'Università degli studi di Torino con l'obiettivo di costruire una piattaforma per l'interconnessione e la diffusione della conoscenza storica.

Si tratta di un sistema, tutt'ora in fase di sviluppo, che vuole raccogliere, unire e connettere le conoscenze territoriali, spesso poco diffuse e a volte difficilmente reperibili su risorse famose quali Wikipedia, garantendo la preservanza nel tempo e allo stesso tempo migliorandone l'accessibilità e l'analisi.

Per realizzare questa missione, il sistema prevede una partecipazione attiva del-

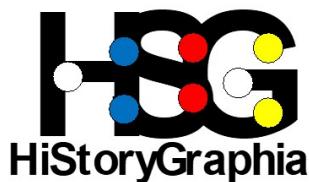


Figura 4.1: Logo HSG

CAPITOLO 4. HISTORYGRAPHIA

l'utenza, che contribuisce inserendo nuova conoscenza. Questo approccio è coerente con le principali tendenze internazionali in materia di patrimonio culturale, a partire dalla Convenzione di Faro, che dichiara "chiunque da solo o collettivamente ha diritto di contribuire all'arricchimento dell'eredità culturale ed è dunque necessario che i cittadini partecipino al processo di identificazione, studio, interpretazione, protezione, conservazione e presentazione dell'eredità culturale" [38].

Si pongono, quindi, al centro del progetto gli utenti, che possono partecipare attivamente al contenuto culturale del sistema, registrandosi e introducendo le conoscenze che desiderano condividere.

L'introduzione di nuove conoscenze avviene tramite la stesura di saggi storici, o narrazioni, parzialmente strutturate, vedi figura 4.2.

Tramite specifici divisorii, rappresentati da parentesi quadre, si differenziano i vari frammenti che compongono gli articoli, ognuno rappresenta un diverso contenuto informativo, il cui tipo è descritto dal divisore stesso. Inoltre, le parti di testo interne al frammento possono essere associate ad una o più fonti di riferimento, anch'esse identificate da parentesi quadre.

Questo formato rispecchia la rappresentazione interna del sistema. Essa si articola in tre livelli di descrizione, vedi figura 4.3: il livello delle **narrazioni**, il triangolo rosso, contiene i saggi precedentemente descritti, creati dai contributori del sito. Il livello della **conoscenza**, il triangolo giallo, contiene il contenuto semantico di rilevanza delle narrazioni. Infine, il livello delle **fonti**, il triangolo grigio, rappresenta il materiale di riferimento associato alle parti di testo delle narrazioni.

Più nello specifico, il sistema si basa su un'ontologia suddivisa nei tre macro livelli sopra presentati. In figura 4.4, è riportata la struttura dettagliata:

- Il livello della **conoscenza** divide il contenuto semantico in sette classi di alto livello che descrivono fatti ed eventi nelle narrazioni.

CAPITOLO 4. HISTORYGRAPHIA

[Articolo]: San Bernardo Castelletto Stura

[Localizzazione]: Castelletto Stura

[Diocesi]: Diocesi di Asti; dal 1430 Diocesi di Mondovì e in seguito alla riorganizzazione post-napoleonica (1817) passa sotto la Diocesi di Cuneo [fonte: L. Berra, *Riordinamento delle Diocesi di Mondovì, Saluzzo, Alba e Fossano ed erezione della Diocesi di Cuneo nel 1817*, in "Bollettino della Società per gli Studi Storici, Archeologici e Artistici della Provincia di Cuneo", 36, 1955, p. 51]

[Dipendenze]: Dal 1430 il territorio di Castelletto è assorbito in quello di più vasta pertinenza della "villanova" di Cuneo [fonte: R. Comba, *Due resoconti inediti della castellania di Cuneo (1388-1409)*, in "Bollettino della Società per gli Studi Storici, Archeologici e Artistici della Provincia di Cuneo", 67, 1972, pp. 32-33]. Nel 1619 Castelletto è infeudato ad Amedeo Ponte di Scarnafigi; nel 1661 passa a Francesco Bartolomeo Sandri Trottì, marchese di Montanera, che nel 1668 lo vende a Giovanni Battista Lamberti, famiglia a cui apparterrà fino al periodo della conquista francese [fonte: M. Ristorto, *Castelletto Stura. Storia civile e religiosa*, Cuneo 1977, pp. 56-73; G. Comino, *Castelletto Stura*, 1998 <https://www.archiviocasalis.it/localized-install/content/castelletto-stura>].

[Cronologia]: sulla base della lettura dei dati di stile, la cronologia di esecuzione cade intorno al 1480 [fonte: E. Brezzi Rossetti, *Giovanni Mazzucco*, in *La Pittura in Italia. Il Quattrocento*, Milano 1987, pp. 708-709]

Figura 4.2: Estratto narrazione. In giallo i divisorii che identificano i diversi frammenti di testo, in grigio le fonti associate ad ogni parte di testo

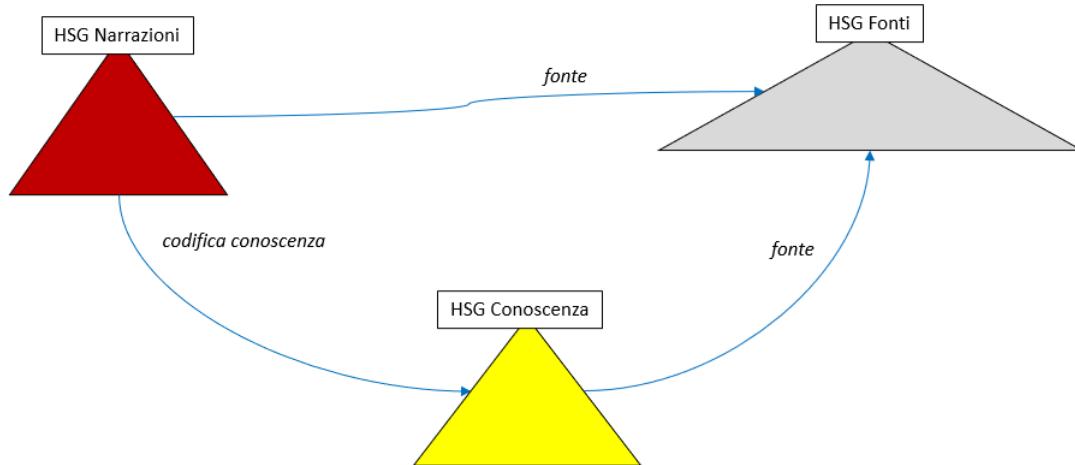


Figura 4.3: Rappresentazione concettuale HSG

CAPITOLO 4. HISTORYGRAPHIA

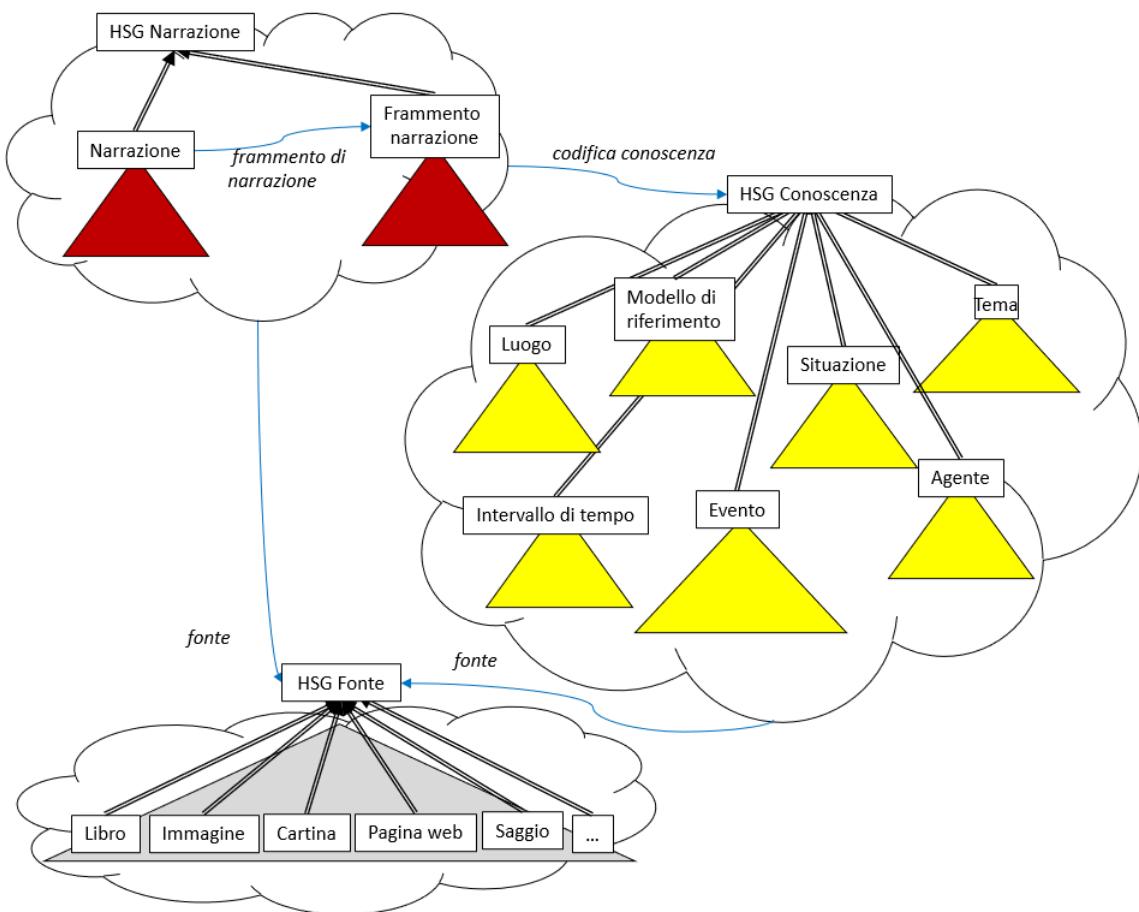


Figura 4.4: Rappresentazione dettagliata dei concetti di macro livello in HSG

- Il livello delle **fonti** contiene le fonti di riferimento che possono essere, ad esempio, libri, pagine web, immagini, cartine, saggi ecc., permettendo all'utente di poter risalire e verificare personalmente il contenuto dei paragrafi.
- Il livello delle **narrazioni** rappresenta separatamente i frammenti che compongono i vari articoli, ognuno assegnato ad una classe diversa in base al tipo di contenuto.

CAPITOLO 4. HISTORYGRAPHIA

... [Committenza]

La chiesa della confraternita viene edificata nel 1538, segno dell'affermazione di un notabilato che controlla la vita economica e sociale della comunità e sviluppatisi tra la metà del Quattrocento e la fine del Cinquecento, a seguito dell'incremento dei traffici commerciali. La rete economica favorisce anche la circolazione di artisti.

I confratelli di Santa Croce commissionano (1658 - 1660) un ciclo di dodici tele a Lorenzo Gastaldi. La scelta del pittore, che in quel momento si trova a Monaco, e realizza molte opere per le valli del Nizzardo e il basso cuneese, si spiega con interessi economici della comunità e alle frequentazioni liguri-nizzarde della popolazione locale. Durante il Seicento gli entracquesi vendono ovini e bovini ai macelli di Grasse, Nizza e Genova. Sovente alcuni abitanti si recavano per affari a Tenda, raggiunta attraverso i colli della Finestra e del Sabbione. A partire dagli anni Settanta del Seicento è potenziata l'attività mercantile, a cui corrisponde il tentativo di rilanciare l'antica strada del colle delle Finestre, che in seguito al potenziamento della strada del colle di Tenda aveva perso la sua importanza. I rapporti di Entracque con la valle della Vesubie, attraverso il colle delle Finestre, sulla strada di Nizza, determinarono almeno un altro episodio di committenza: il "mastro da bosco" (falegname) Giacomo Rosso di Lantosca realizza nel 1684 un armadio per la parrocchiale di Sant'Antonino.[F. ARNEODO, D. DEIDDA, L. VOLPE, *Attività pastorizia ed evoluzione degli equilibri socio-economici a Entracque (secoli XV-XVIII)*, in *Entracque : una comunità alpina tra Medioevo ed Età moderna*, Atti della giornata di studio, Entracque, 13 Aprile 1997, a cura di R. Comba, M. Cordero, Cuneo 1997, pp. 107-143]. [FONTE: B. Palmero, Entracque 2008, <https://www.archiviocasalis.it/localized-install/biblio/cuneo/entracque>].

[Iconografia]

Si tratta di dodici teleri raffiguranti episodi della vita di Cristo e della Vergine. Una scelta che risponde ai dogmi affermati dalla Chiesa cattolica della Controriforma, l'Immacolata Concezione della Vergine, la Trinità, il riscatto del genere umano attraverso il Sacrificio di Cristo. A questo si aggiungono i riferimenti alle pratiche liturgiche svolte dai confratelli durante il Giovedì Santo, rievocate nell'Ultima Cena e soprattutto nella Lavanda dei piedi.

Il modello decorativo di riferimento è quello della Confraternita di Santa Croce a Cuneo, dove nel 1626 sono allestite le tele con i Miracoli della Vera Croce dipinti nel 1626 da Giulio e Giovanni Battista Bruno (che a loro volta rimandano ai modelli decorativi di primo Seicento delle Casacce (oratori), genovesi.

...

Figura 4.5: Esempio articolo: Confraternita di Santa Croce a Entracque

Per esempio, preso un estratto dell'articolo della "*Confraternita di Santa Croce a Entracque*", riportato in figura 4.5, una parte della rappresentazione concettuale nell'ontologia di HiStoryGraphia è riportata in figura 4.6: in rosso, il livello di narrazione, è rappresentata la narrazione e i collegamenti con i vari frammenti che la compongono; in grigio, il livello delle fonti, è presente la fonte relativa alla sezione della committenza; mentre in giallo, il livello della conoscenza, vi è la rappresentazione delle informazioni contenute nei frammenti, relative agli eventi di "*committenza del ciclo di 12 teleri a Lorenzo Gastaldi*" e "*erezione della Chiesa della Confraternita di Santa Croce a Entracque*".

CAPITOLO 4. HISTORYGRAPHIA

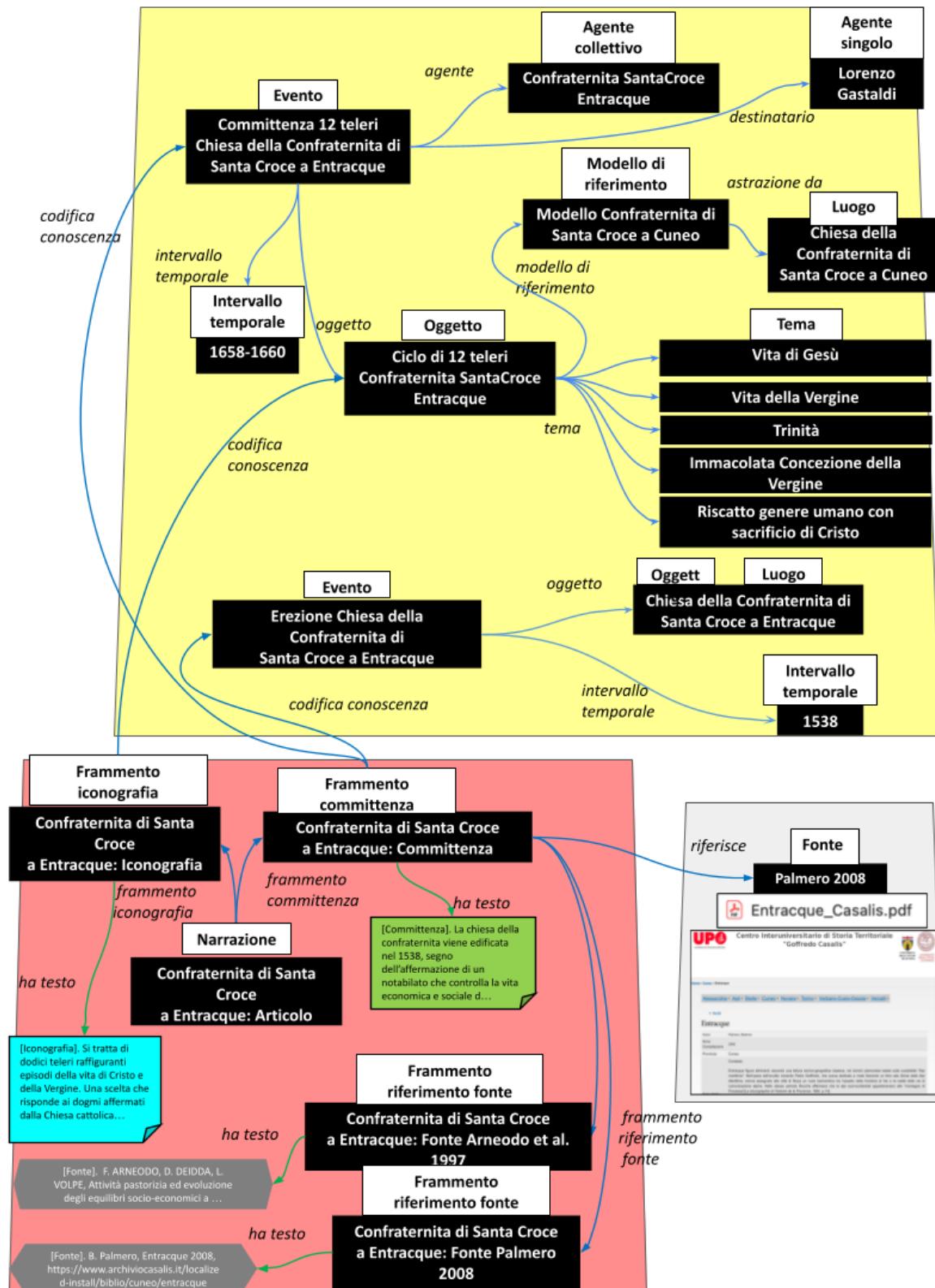


Figura 4.6: Esempio rappresentazione concettuale dell'articolo in figura 4.5

CAPITOLO 4. HISTORYGRAPHIA

Questo tipo di memorizzazione garantisce la non perdita di informazioni accentuata nei problemi presentati nel capitolo 2, in quanto si mantiene il file originale e si rappresenta il contesto e la fonte da cui è tratto il contenuto semantico del livello di conoscenza.

La ricerca di questa tesi si posiziona lungo la connessione tra i livelli di narrazione e conoscenza.

Attualmente il processo di codifica della conoscenza contenuta nei saggi avviene tramite un lavoro manuale con il supporto di un'apposita interfaccia: si selezionano i partecipanti o entità, li si classificano secondo lo schema proposto dall'ontologia e si mettono in relazione tra di loro.

L'implementazione mira a semplificare questa procedura creando automaticamente dei suggerimenti di annotazioni con l'obiettivo futuro di poter automatizzare l'intero processo.

In questo capitolo è stato descritto brevemente il progetto HiStoryGraphia. Si tratta un progetto nato presso l'università degli Studi di Torino finalizzato alla costruzione di una piattaforma per la interconnessione e diffusione della conoscenza storica. L'approccio pone al centro gli utenti, che sono attivamente responsabili di arricchire l'eredità culturale del progetto. Infine, è stata fornita una presentazione dell'ontologia alla base della rappresentazione della conoscenza e un esempio di articolo storico che potrebbe essere generato dall'utenza.

Capitolo 5

Il modulo Storytelling2Knowledge

Il seguente capitolo presenta il modulo Storytelling2Knowledge per l'estrazione di conoscenza. Si inizia esponendo l'architettura generale, un sistema composto da tre sezioni, ognuna dedicata al riconoscimento di un tipo di informazione. Successivamente, si entrerà più nel dettaglio descrivendo i singoli componenti e il loro contributo alla realizzazione del grafo di conoscenza.

L'approccio di utilizzare diversi componenti per combinarne i risultati viene presentato da Martinez et al. [32] con il termine *Ensemble systems*. Una prassi comune nell'information extraction prevede l'utilizzo di uno o più software per l'estrazione di entità generiche combinati con tecniche ad hoc per estrarre quelle entità specifiche del dominio di lavoro.

Ad esempio, il sistema di estrazione entità TEXTCROWD [39] combina diversi moduli ner generici a tecniche specifiche. Come abbiamo visto nel capitolo dedicato all'estrazione di entità, i modelli generici sono addestrati a riconoscere esclusivamente tag generici quali persone, località e organizzazioni. TEXTCROWD combina i risultati di questi moduli a dei gazetteer insieme ad alcune complesse regole specifiche del dominio archeologico. Questo permette al sistema di sopperire alla debolezza di generalità dei moduli ner e di riconoscere entità specifiche, come ad esempio materiali, periodi storici, tecniche ecc.

Tuttavia, l'utilizzo congiunto di diverse risorse può portare a ottenere risultati in contraddizione, si prenda ad esempio la frase "*La Nascita di Venere è un dipinto del 1485*", diverse risorse potrebbero riconoscere entrambi i concetti di "Nascita di Venere", inteso come dipinto, e Venere, inteso come pianeta.

Servono criteri di scelta o disambiguazione. Ad esempio, il lavoro di Martinez et al. [40] presenta una implementazione di un sistema OpenIE in cui si combinano i risultati di tre diversi tool di entity linking applicando una strategia di scelta basata su voto maggioritario, ovvero l'entità riconosciuta da più tool.

Ispirati da queste ed altre opere, si propone un approccio modulare con il duplice obiettivo di estrarre un grafo di conoscenza combinando i risultati di diversi tool e contemporaneamente di valutare il contributo delle singole risorse, utilizzando un'interfaccia che permetta di controllarne l'attivazione, permettendo in futuro di introdurre nuovi componenti sempre più aggiornati o in linea con le esigenze.

5.1 Architettura

Si introduce di seguito l'architettura del modulo Storytelling2Knowledge. Esso è stato pensato al fine di renderlo facilmente modificabile, tramite aggiunta o sostituzione di componenti, seguendo l'evoluzione della piattaforma e della ricerca IE in lingua italiana.

Analizzando esempi di articoli creati da esperti storici è stato possibile individuare tre tipi di informazioni estraibili dato un documento.

La prima informazione è connessa alla struttura impostata alla creazione dei saggi. Essi sono divisi in frammenti, differenziati dal tipo di contenuto semantico trattato, e ognuno è associato a fonti di riferimento. Queste informazioni sono riconoscibili grazie a delle etichette create tramite l'utilizzo di parentesi quadre che rendono il saggio storico un documento semi strutturato. Per esempio, dato l'estratto di articolo in figura 5.1, le etichette in giallo permettono di evidenziare le sezioni che rappresentano i tre frammenti, rispettivamente, di cronologia,

CAPITOLO 5. IL MODULO STORYTELLING2KNOWLEDGE

autore e geografia artistica. Inoltre, le etichette in grigio, contengono le fonti da associare alle parti di testo.

La seconda informazione estraibile riguarda i concetti, o entità, presenti all'interno del testo: persone, luoghi, oggetti, edifici, date ecc., cioè ogni elemento associabile ad un concetto del mondo reale, che è possibile riconoscere all'interno del saggio. Per esempio, nel frammento autore, in figura 5.1, data la frase "*Gli affreschi sono assegnati, sulla base dei dati di stile a Giovanni Mazzucco, un pittore*" è presente l'entità di *Giovanni Mazzucco*.

Infine, la terza informazione riguarda le connessioni che si formano durante la narrazione: le entità presenti in un testo si connettono a concetti o altre entità tramite la struttura della frase e ne descrivono un comportamento o un aspetto. Per esempio, nel frammento geografia artistica, in figura 5.1, la frase "*Nel 1487 decora la chiesa di San Domenico a Peveragno*", contiene una relazione tra l'intervallo temporale 1487 e l'evento di *decorazione della chiesa di San Domenico*.

L'architettura del modulo Storytelling2Knowledge, riportata in figura 5.2, è composta da tre sezioni principali, ognuna incaricata di estrarre uno specifico tipo di informazione contenuta nell'articolo:

- **Wrapper:** estraie i frammenti e le fonti che costituiscono il saggio.
- **Estrattore entità:** riconosce, estraie e disambigua le entità presenti nei testi.
- **Estrattore relazioni:** riconosce le relazioni.

I tre componenti possono essere utilizzati singolarmente o combinandone i risultati. Ad esempio, se si volesse analizzare una parte di testo libero, si potrebbe scegliere di non passare dal componente wrapper e andare direttamente all'estrazione delle entità e/o relazioni.

CAPITOLO 5. IL MODULO STORYTELLING2KNOWLEDGE

[Cronologia]: sulla base della lettura dei dati di stile, la cronologia di esecuzione cade intorno al 1480 [fonte: E. Brezzi Rossetti, Giovanni Mazzucco, in La Pittura in Italia. Il Quattrocento, Milano 1987, pp. 708-709]

[AUTORE]

Gli affreschi sono assegnati, sulla base dei dati di stile a Giovanni Mazzucco, un pittore attivo in area monregalese nella seconda metà del Quattrocento. [fonte: E. Brezzi Rossetti, Giovanni Mazzucco, in La Pittura in Italia. Il Quattrocento, Milano 1987, pp. 708-709]

[GEOGRAFIA ARTISTICA]

La prima data utile per ricostruire l'attività di Giovanni Mazzucco è il 1481, quando firma gli affreschi della cappella del Santo Sepolcro a Piozzo. Nel 1487 decora la chiesa di San Domenico a Peveragno. Al 1491 risale l'ultima opera firmata da Mazzucco, realizzata nel santuario del Brichetto a Morozzo, su commissione di Biagio Fauzone. Questo nucleo di affreschi tutti firmati da Giovanni Mazzucco, sono la base di riferimento per assegnare, al pittore e alla sua bottega, altre opere, affini per i dati di stile. [fonte: L. Marino, Sulle tracce di Giovanni Mazzucco. Biografia, stile, confronti, in Il restauro della cappella di San Bernardo a Castelletto Stura, Cuneo 2007, pp.59-78]

Figura 5.1: Esempio estratto di articolo, tratto da San Bernardo Castelletto Stura

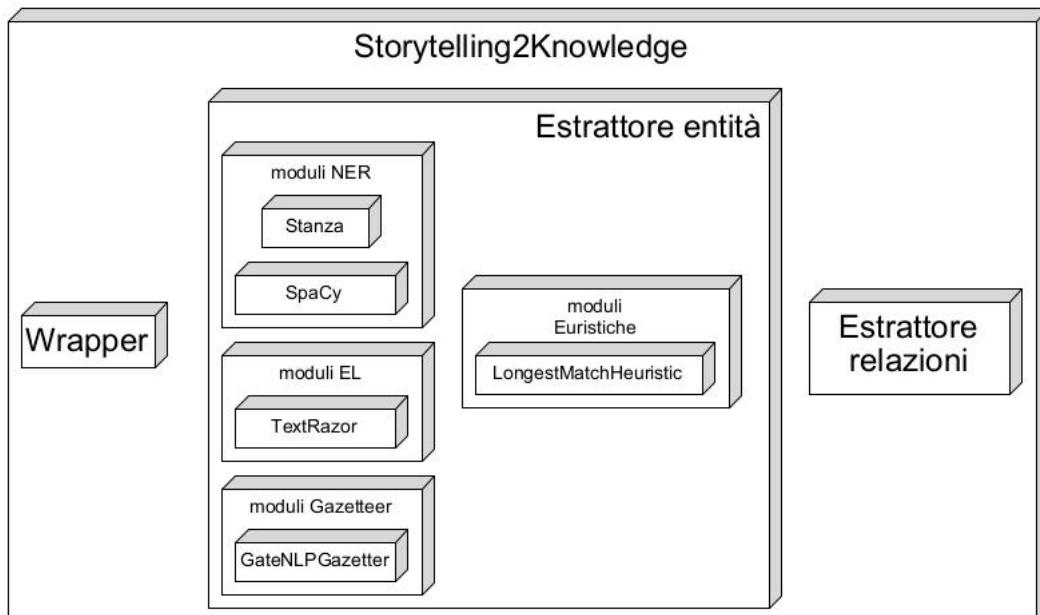


Figura 5.2: Architettura modulo Storytelling2Knowledge

5.2 Wrapper

Si inizia la descrizione dei componenti partendo dal primo, il *Wrapper*, che ha la responsabilità di estrarre gli elementi che caratterizzano la struttura del saggio: i frammenti e le fonti. Il risultato atteso è quindi l'estrazione dei vari frammenti che costituiscono il saggio, ognuno di un tipo diverso in base al contenuto, e l'associazione alle fonti da cui è tratta l'informazione.

Ispirati dai lavori di Ciravegna et al. [21] sulle potenzialità del wrapper con documenti semi strutturati, si è deciso di implementare una procedura per l'individuazione delle parti all'interno delle narrazioni create dagli utenti.

La costruzione di questi documenti prevede l'etichettatura, delimitata da parentesi quadre, sia del contenuto delle varie sezioni che li compongono, sia del riferimento alla fonte a cui sono associati i frammenti di testo.

[Articolo]: San Bernardo Castelletto Stura

[Localizzazione]: Castelletto Stura

[Diocesi]: Diocesi di Asti; dal 1430 Diocesi di Mondovì e in seguito alla riorganizzazione post-napoleonica (1817) passa sotto la Diocesi di Cuneo [fonte: L. Berra, *Riordinamento delle Diocesi di Mondovì, Saluzzo, Alba e Fossano ed erezione della Diocesi di Cuneo nel 1817*, in "Bollettino della Società per gli Studi Storici, Archeologici e Artistici della Provincia di Cuneo", 36, 1955, p. 51]

[Dipendenze]: Dal 1430 il territorio di Castelletto è assorbito in quello di più vasta pertinenza della "villanova" di Cuneo [fonte: R. Comba, *Due resoconti inediti della castellania di Cuneo (1388-1409)*, in "Bollettino della Società per gli Studi Storici, Archeologici e Artistici della Provincia di Cuneo", 67, 1972, pp. 32-33]. Nel 1619 Castelletto è infeudato ad Amedeo Ponte di Scarnafigi; nel 1661 passa a Francesco Bartolomeo Sandri Trottì, marchese di Montanera, che nel 1668 lo vende a Giovanni Battista Lamberti, famiglia a cui apparterrà fino al periodo della conquista francese [fonte: M. Ristori, *Castelletto Stura. Storia civile e religiosa*, Cuneo 1977, pp. 56-73; G. Comino, *Castelletto Stura*, 1998 <https://www.archiviocasalis.it/localized-install/content/castelletto-stura>].

[Cronologia]: sulla base della lettura dei dati di stile, la cronologia di esecuzione cade intorno al 1480 [fonte: E. Brezzi Rossetti, *Giovanni Mazzucco*, in *La Pittura in Italia. Il Quattrocento*, Milano 1987, pp. 708-709]

Figura 5.3: Estratto iniziale narrazione: in giallo i divisori delle sezioni di testo, in grigio le fonti associate ad ogni frammento

Per chiarire, si riporta in figura 5.3 l'esempio di saggio storico presentato nel ca-

CAPITOLO 5. IL MODULO STORYTELLING2KNOWLEDGE

pitolo precedente: le etichette evidenziate in giallo permettono di spezzare il testo in sezioni, il cui contenuto informativo è descritto dall'etichetta stessa. Ogni sezione è a sua volta composta da più frammenti. Si tratta di frammenti dello stesso tipo, cioè legati alla stessa etichetta in giallo, ma che si differenziano per le fonti, in figura colorate di grigio, da cui è tratta l'informazione. Ad esempio, in figura 5.3, la sezione delle dipendenze è composta da due frammenti: la frase "*Dal 1430...Cuneo*" legata alla fonte "*R.Comba..33*" e la frase "*Nel 1619...francese*" associata alla fonte "*M.Ristorto...stura*".

Dal momento che si tratta di schema fissi, cioè uguali per tutti gli articoli, si è optato per una soluzione basata su espressioni regolari. Utilizzando questa informazione è possibile costruire delle semplici regole per riconoscere le varie sezioni, estrarne i frammenti di testo e associarli a eventuali fonti di riferimento.

L'implementazione, scritta in Python, utilizza la libreria `re`¹, dedicata alla costruzione di comandi per riconoscere una specifica sequenza di caratteri interna ad un testo.

L'analisi si divide in due fasi: una prima fase riconosce le sezioni dell'articolo ed estrae l'etichetta che ne descrive il tipo, mentre, nella seconda parte, si estrae, per ogni sezione, i frammenti che la compongono.

Per l'estrazione delle sezioni si individuano le etichette che le delimitano, la strategia di scelta prevede la combinazione di due regole: la prima ricerca di parti di testo delimitate da parentesi quadre e la cui lunghezza non superi i 50 caratteri, la seconda controlla che all'inizio non sia presente la parola "fonte".

Questa scelta è voluta in caso di dimenticanza del termine "fonte" per rappresentare un riferimento. In base ai dati in possesso si è constatato che difficilmente una fonte ha un numero così ridotto di caratteri.

L'idea originale prevedeva di sostituire questa parte quando la lista di possibili etichette per i frammenti fosse stata finalizzata, tuttavia, l'attuale strategia permette di poter riconoscere automaticamente quando sia presente una nuova eti-

¹<https://docs.python.org/3/library/re.html>

CAPITOLO 5. IL MODULO STORYTELLING2KNOWLEDGE

chetta non ancora aggiunta in memoria e prevenire in caso di errori di battitura su di esse.

La seconda fase prevede di riconoscere all'interno delle sezioni divise la presenza di eventuali fonti e di separare i paragrafi in caso siano riferiti a differenti fonti come nel caso delle dipendenze in figura 5.3.

Il risultato, presentato in formato JSON², presenta una lista di frammenti, ciascuno descritto da tre attributi, di cui l'ultimo opzionale: il *tipo*, il *testo* e le *fonti*.

```
1 [  
2 ...  
3 {  
4   'source': ['L. Berra, Riordinamento ... p. 51'],  
5   'text': 'Diocesi di Asti; Dal ... sotto la Diocesi di Cuneo ',  
6   'type': 'Diocesi'},  
7 {  
8   'source': ['R. Comba, ... 32-33'],  
9   'text': 'Dal 1430 ... "villanova" di Cuneo ',  
10  'type': 'Dipendenze'},  
11 {  
12   'source': ['M. Ristorto, ... castelletto-stura'],  
13   'text': 'Nel 1619 Castelletto ... conquista francese ',  
14   'type': 'Dipendenze'},  
15 ...  
16 ]
```

Codice 5.1: Esempio di parte del risultato dell'estrazione di frammenti dall'articolo precedente

Il componente del wrapper serve a scomporre le narrazioni in frammenti classificati secondo un tipo, o classe, definito nella struttura del saggio stesso. Inoltre, l'estrazione permette di connettere la fonte da cui è tratta l'informazione del frammento. Al termine del processo si ottengono come risultato i vari frammenti,

²JavaScript Object Notation

descritti tramite: il tipo o classe semantica, il testo contenuto e una lista di fonti ad esso associate.

Questo componente contribuisce alla costruzione del grafo di conoscenza estraendo l'informazione relativa alla rappresentazione dei frammenti e delle fonti ad essi collegati.

5.3 Estrazione di entità

Il secondo componente riguarda l'estrazione di entità presenti in un testo, combinando diverse risorse esistenti per la lingua italiana. In particolare, sono presenti due tool NER per il riconoscimento di entità generiche, privi della fase di disambiguazione, un tool di entity linking (EL), che restituisce per le entità disambigue, e un gazetteer utilizzabile sia per il task di riconoscimento di entità più specifiche sia per il task di entity linking.

Inoltre, è presente una funzione di strategia per determinare il risultato migliore in caso di entità sovrapposte. Questa semplice strategia rientra nella categoria di disambiguazione basata su menzioni descritta precedentemente.

L'attività del modulo è divisa in due fasi: la prima fase consiste nell'estrarrre le entità tramite l'utilizzo delle risorse integrate, mentre la seconda si occupa di partizionare le entità in gruppi di sovrapposizione e per ognuno di essi estrarre le entità in base all'euristica impostata. Ad esempio, data la frase "*A Torino è stata esposta la Nascita di Venere*", è possibile che vengano riconosciute due entità in sovrapposizione: "*Nascita di Venere*" e "*Venere*". La seconda fase andrebbe quindi a costruire due gruppi: il primo composto solo da "*Torino*", e quindi già disambiguato, mentre il secondo, composto da [*"Nascita di Venere"*, *"Venere"*], da cui sarà estratta l'entità principale in base all'euristica impostata.

Il modulo è diviso in sotto-moduli, al fine di semplificare la gestione e l'aggiunta di nuove risorse e euristiche. Vi sono tre sotto-moduli dedicati all'estrazione di entità: un modulo per le risorse NER, che estraggono e classificano le entità

senza fase di disambiguazione, un modulo per le risorse EL, che estraggono e associano le entità a identificatori, in basi di conoscenza esterne, e un modulo per i gazetteer, che possono svolgere entrambe le funzionalità. Infine, vi è un quarto sotto-modulo per la gestione delle euristiche, che, come vedremo, serviranno per gestire, al termine dell'estrazione dei tre sotto-moduli precedenti, i casi di sovrapposizione delle menzioni.

5.3.1 Tools

Segue la presentazione delle risorse integrate per l'estrazione di entità. Le prime due risorse, *spaCy* e *Stanza*, svolgono il task NER, permettendo di estrarre entità di tipo generico. Segue la risorsa *Text Razor*, un sistema di Entity Linking, che restituisce entità disambigueate e rappresentate dall'identificatore della base di conoscenza. Infine, il gazetteer implementato che può svolgere entrambi i compiti.

SpaCy

SpaCy è una libreria open-source che contiene soluzioni sullo stato dell'arte per lo svolgimento di avanzate attività di elaborazione del linguaggio naturale in Python.



La libreria, rilasciata sotto licenza MIT³, è stata scritta in Python e Cython e supporta soluzioni per oltre 60 lingue, tra cui l'italiano.

Figura 5.4: Logo
spaCy

Per la lingua italiana sono presenti tre modelli, differenziati per le dimensioni del dataset su cui sono stati addestrati, che presentano un trade-off tra costo computazionale e precisione della prestazione. Per questo progetto si è optato per la

³Licenza di software libero creata dal Massachusetts Institute of Technology

CAPITOLO 5. IL MODULO STORYTELLING2KNOWLEDGE

soluzione che fornisse i risultati migliori. Si tratta di un modello addestrato su un dataset di 549MB denominato *it_core_news_lg*.

Questi modelli si basano su pipelines pre-allenate, una pipeline è una serie di componenti sequenziali il cui fine è l'annotazione dei testi in analisi.

Ognuno di questi componenti svolge un specifico compito o task, alcuni sono legati da una relazione di dipendenza e quindi necessitano il risultato degli altri per poter procedere al proprio compito.

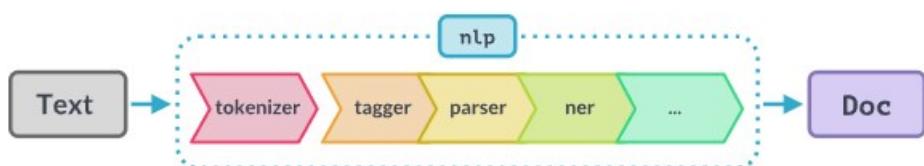


Figura 5.5: Pipeline spaCy [41]

Tra questi, siamo interessati al modulo NER che si occupa della ricerca delle *Named Entity*.

Esso si occupa di aggiungere annotazioni relative a entità nominali contenute nel testo, tali entità vengono annotate utilizzando il formato BIO.

Il modulo è addestrato utilizzando il corpus WikiNER [42], un dataset multilingua generato da Wikipedia che utilizza gli stessi tag, per entità generiche, introdotti da CoNLL-2002 (LOC, MISC, ORG, PER).

Stanza

Stanza è una libreria open-source che colleziona tool e modelli in linea con lo stato dell'arte in campo NLP per eseguire analisi linguistiche in Python.



Si tratta della libreria ufficiale del gruppo NLP di Stan-

Figura 5.6: Logo
Stanza

CAPITOLO 5. IL MODULO STORYTELLING2KNOWLEDGE

ford⁴, rilasciata sotto licenza Apache⁵.

Analogamente a spaCy, la struttura è basata su una pipeline composta da vari componenti, vedi figura 5.7.

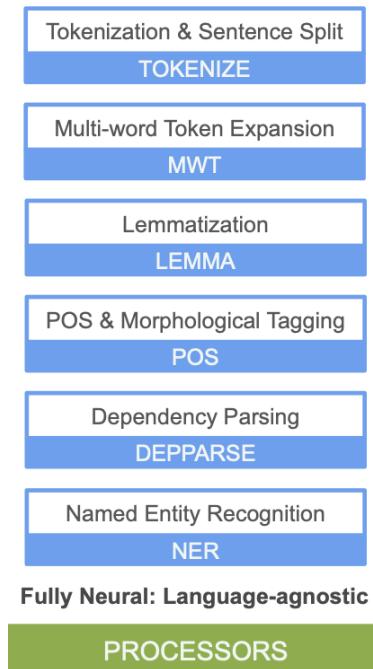


Figura 5.7: Pipeline Stanza [43]

Tra questi, il componente NER è allenato su un dataset rilasciato da FBK⁶ che utilizza tre entità generiche: PER, LOC, ORG.

La missione del modulo è il riconoscimento delle entità nominali all'interno del testo in analisi e avviene tramite un'annotazione in formato BIOES. Un esempio di annotazione, per la frase "*Nel 1487 Giovanni Mazzucco decora la chiesa di San Domenico a Peveragno*", è mostrato in figura 5.8.

⁴<https://nlp.stanford.edu/>

⁵<http://www.apache.org/licenses/LICENSE-2.0>

⁶Fondazione Bruno Kessler - <https://dh.fbk.eu/>

CAPITOLO 5. IL MODULO STORYTELLING2KNOWLEDGE

Nel	O
1487	O
Giovanni	B-PER
Mazzucco	E-PER
decora	O
la	O
chiesa	B-LOC
di	I-LOC
San	I-LOC
Domenico	E-LOC
a	O
Peveragno	S-LOC

Figura 5.8: Esempio rappresentazione in formato BIOES

TextRazor

TextRazor è una startup nata a Londra nel 2011. L'obiettivo dell'impresa è di fornire software per aiutare a costruire applicazioni di analisi testuale in maniera semplice e con prestazioni al pari dello stato dell'arte.



Figura 5.9: Logo
TextRazor

La risorsa garantisce una soluzione al task di entity linking. Permette di riconoscere, disambiguare e connettere le entità a dei sistemi di rappresentazione della conoscenza.

In particolare, si è scelto di utilizzare questa risorsa rispetto ad altre che svolgono la stessa attività, ad esempio TagMe⁷, in quanto, tra i sistemi utilizzati per rappresentare le entità estratte, è presente l'identificatore univoco di Wikidata⁸.

Wikidata è un database libero, collaborativo, multilingue e secondario che raccolge dati strutturati per fornire supporto a Wikipedia, a Wikimedia Commons, agli altri progetti del movimento Wikimedia e a chiunque nel mondo [44].

TextRazor fornisce delle librerie in diversi linguaggi, tra cui Python, finalizzate all'integrazione del servizio web offerto con le proprie applicazioni.

⁷<https://tagme.d4science.org/tagme/>

⁸https://www.wikidata.org/wiki/Wikidata:Main_Page

CAPITOLO 5. IL MODULO STORYTELLING2KNOWLEDGE

Si tratta di un servizio basato su scambi di messaggi web, che necessita una registrazione per poter ottenere una chiave di accesso e la cui versione gratuita ha un numero limitato di interazioni giornaliere.

Gazetteer

Come accennato precedentemente, i gazetteer sono la tecnica più diffusa per gestire domini specifici.

Questa soluzione permette di raccogliere un insieme di termini per estrarre entità specifiche del dominio e assegnarle ad una classe specifica o ad un identificatore in una base di conoscenza associata.

L'idea che ha portato alla costruzione di questa risorsa è la possibilità di utilizzarla per estrarre le entità relative alle opere d'arte che spesso presentano, internamente al titolo, dei nomi propri di persone, che i tool generici quali spaCy e Stanza tendono a classificare come persone. Per esempio, data l'opeara *Immacolata con i santi Sebastiano, Giuseppe, Rocco e Carlo Borromeo*, dall'articolo "Entracque – Confraternita di Santa Croce", un tool generico potrebbe considerare "Giuseppe", "Rocco" e "Carlo Borromeo" come persone, invece di riconoscere il titolo dell'opera.

Per la costruzione della risorsa si è scelto l'utilizzo del framework Python GatenLP⁹. Si tratta di un sistema che raccoglie una serie di tecniche per l'annotazione di documenti testuali, tra cui *StringGazetteer*, un tipo di gazetteer che esegue un confronto tra stringhe e cerca i termini memorizzati all'interno del testo in analisi, annotando in base al tipo di informazione per cui è configurato.

La risorsa è stata preimpostata per un futuro utilizzo, con due possibili scenari:

- Si vuole riconoscere delle entità presenti nell'ontologia di HSG e quindi associarle al corrispettivo identificativo. Essendo un progetto per l'interconnessione della conoscenza locale, è più probabile che l'entità riconosciuta si

⁹<https://gatenlp.github.io/python-gatenlp/>

riferisca allo stesso concetto nell'ontologia di HSG rispetto ad una generica come Wikidata. Per esempio, nelle narrazioni compare l'entità di *Lorenzo Gastaldi*, un pittore vissuto nel diciassettesimo secolo, mentre in Wikidata, la prima voce associata al nome Lorenzo Gastaldi è un arcivescovo vissuto nel diciannovesimo secolo.

- Si vuole riconoscere una classe specifica che i tool generici non possiedono. Una soluzione simile è presente nel lavoro di Rovera et al. [25] in cui creano file contenenti liste di termini per riconoscere colori, artefatti, periodi ecc. Per esempio, nelle narrazioni compaiono diversi riferimenti ai secoli, si potrebbe definire un gazetteer contenente i vari secoli: Cinquecento, Seicento, Settecento ecc..

Per l'implementazione è stata realizzata una classe *GateNlpGazetteer* in Python che permette di realizzare un numero a piacere di gazetteer. Essa fornisce due possibili configurazioni: la prima si utilizza per raccogliere un set di termini relativi ad una classe specifica, mentre la seconda si aspetta di ricevere oltre al termine anche l'identificativo di una base di conoscenza a discrezione ed eventualmente una classe specifica facoltativa.

In conclusione, il componente di estrazione delle entità aggrega i risultati di quattro risorse: due risorse dedicate al riconoscimento e classificazione di entità generiche, che potranno esser sfruttate per definire nuovi concetti non presenti nella base di conoscenza interna di HiStoryGraphia, una risorsa di entity linking che permetterà di introdurre entità rappresentate in Wikidata e un gazetteer da definire in base alle necessità che potrebbe essere utilizzato per definire delle classi specifiche del sistema, come ad esempio le opere d'arte o come risorsa entity linking.

5.3.2 Euristiche per la sovrapposizione

Una delle difficoltà legate alla costruzione di sistemi combinati da diverse risorse è la sovrapposizione dei risultati. È necessario introdurre euristiche o tecniche di disambiguazione per determinare tra i risultati il più corretto.

Si prenda l'esempio del dipinto "*Ritratto di Dante*", questa entità può generare le due distinte menzioni di:

- Ritratto di Dante come dipinto
- Dante inteso come persona

A tal fine si introduce una euristica che rientra nella categoria "basate su menzioni", descritta nel capitolo 3.1.2.

L'idea è di scegliere la menzione più lunga, in quanto è considerata più specifica. Nel caso precedente verrebbe selezionata quella relativa al dipinto.

Questa soluzione, se pur molto semplice, ha ottenuto risultati positivi in diversi studi, tra cui Rovera et al. [25], in cui viene utilizzata per scegliere tra diverse menzioni relative a luoghi o organizzazioni, e Anita et al. [45], dove applicano questa strategia per le entità estratte nel campo medico.

Nel modulo Storytelling2Knowledge, l'euristica si combina ad una strategia di priorità assegnata all'ordine in cui sono eseguiti i tool. Infatti, tramite l'ordine di esecuzione delle risorse, si costruisce una lista di entità che successivamente viene partizionata in gruppi di sovrapposizione; i quali al loro interno tracciano le entità dentro una coda.

In questo modo, nel caso in cui il criterio di confronto dell'euristica non permette di distinguere l'entità, si potrebbe dare la priorità ad una risorsa più precisa controllando l'ordine di esecuzione. Ad esempio, utilizzando l'euristica della menzione più lunga, nel caso in cui le due menzioni da confrontare abbiano la medesima lunghezza, si può scegliere di eseguire prima il gazetteer che ci si aspetta dia risultati più precisi rispetto ad una risorsa generica quale spaCy.

Il componente per l'estrazione delle entità ha la finalità di estrarre, classificare e eventualmente disambiguare i casi di sovrapposizione legati all'estrazione di entità. Combinando i risultati di quattro risorse con un'euristica per la gestione dei risultati di sovrapposizione, il componente permette di estrarre una lista di entità presenti nel testo, descrivendole tramite un'interfaccia interna del sistema.

Il risultato contribuisce alla costruzione del grafo di conoscenza, in quanto definisce un insieme di concetti presenti nella narrazione e li connette a risorse esterne, se ottenuti tramite entity linking, o ad una classe semantica, se ottenuti tramite NER.

5.4 Estrazione di relazioni

Si presenta di seguito l'ultimo componente del modulo Storytelling2Knowledge, che è incaricato di estrarre le relazioni all'interno di un testo. Questo componente può essere utilizzato in combinazione al componente di estrazione delle entità, passando le entità estratte e estraendo esclusivamente relazioni che le coinvolgono, o singolarmente e quindi estraendo ogni relazione riconoscibile all'interno del testo.

L'idea originale prevedeva di introdurre moduli esterni da valutare, tuttavia uno dei problemi attuali dell'analisi in lingua italiana è proprio la difficoltà nel trovare librerie libere da incorporare.

Al fine di creare un sistema il più completo possibile, si è implementato una semplice soluzione OpenIE partendo da uno script in Python di Nicolas Schrading¹⁰ [46], riadattato dalla lingua inglese a quella italiana.

La soluzione proposta permette di estrarre triple <soggetto, verbo, oggetto> da semplici frasi, come nei classici sistemi OpenIE per lingua inglese, al fine da rendere più semplice la sostituzione del modulo in futuro.

¹⁰<http://nicschrading.com/about/>

CAPITOLO 5. IL MODULO STORYTELLING2KNOWLEDGE

L'idea consiste nel partire dai risultati dell'analisi tramite il tool spaCy, che ci permette di ottenere l'albero a dipendenze delle frasi, e successivamente si utilizzano semplici regole, basate sui tag delle dipendenze, per riconoscere le relazioni SVO (soggetto-verbo-oggetto).

Per la creazione dello script è stata necessaria una fase di adattamento dall'inglese all'italiano, in quanto le dipendenze tra le parole delle due lingue hanno un uso e struttura diversa.

Inizialmente è stato necessario definire una lista di tag per riconoscere, rispettivamente, soggetto e oggetto. Tale lista è stata pensata basandosi sui tag scelti da Schrading, per la lingua inglese, e quelli presenti nel modello della lingua italiana di spaCy¹¹, combinando la scelta con esempi recuperati dalla documentazione ufficiale di Universal Dependencies¹²:

```
soggetti = "nsubj", "nsubj:pass", "csubj"  
oggetti = "obj", "obl:agent"
```

Analogamente, emulando il lavoro di Schrading, sono state adattate all'italiano tre regole, apportando le modifiche necessarie, per riconoscere le seguenti condizioni:

- presenza di **congiunzioni** di soggetti e oggetti: per esempio, data la frase "*lui e suo fratello hanno cucinato una torta e una pizza*", sono presenti due soggetti, "lui" e "fratello", e due oggetti, "torta" e "pizza". Tuttavia, negli alberi a dipendenze, il verbo è connesso solo al primo soggetto e al primo oggetto che a loro volta sono connessi ai secondi tramite una dipendenza di congiunzione. Sono state quindi introdotte regole per riconoscere tale dipendenza e quindi riconoscere la presenza di altri soggetti e oggetti nelle frasi.

¹¹https://spacy.io/models/it#it_core_news_lg

¹²<https://universaldependencies.org/it/dep/>

- presenza di **negazione**: per la rappresentazione della conoscenza, la negazione del verbo è un'informazione importante. Ad esempio, data la frase "lui non fornisce supporto", la presenza del "non" cambia il senso della frase. Sono state quindi introdotte regole per riconoscere semplici casi di negazione, come ad esempio la presenza di una connessione del verbo con il termine "non".
- presenza di frasi **passive**: la struttura a dipendenze per le frasi passive, come ad esempio "la Gioconda è stata dipinta da Leonardo", cambia rispetto a quelle attive e di conseguenza la struttura delle dipendenze. Servono regole appropriate per riconoscere queste relazioni.

In conclusione, il componente per l'estrazione delle relazioni permette di estrarre triple <soggetto, verbo, oggetto>. Per realizzarlo, è stato costruito un semplice sistema OpenIE a regole, emulando una risorsa pubblica per la lingua inglese. Il risultato ottenuto permette di definire una lista di triple soggetto-verbo-oggetto che sono successivamente filtrate per ottenere esclusivamente triple legate a entità estratte dal componente per l'estrazione di entità.

Il risultato permette di arricchire il grafo di conoscenza fornendo nuove connessioni tra le entità.

5.5 Conclusioni di capitolo

In questo capitolo è stata descritta l'implementazione del modulo Storytelling2Knowledge per l'estrazione di un grafo di conoscenza partendo da una narrazione. Il modulo prevede l'utilizzo di tre componenti, ognuna specializzata nell'estrazione di un determinato tipo di informazione: la prima, il wrapper, permette di sfruttare la semi struttura delle narrazioni per riconoscere i frammenti che le compongono e le fonti a cui sono associati. Per realizzarla è stata costruita una procedura tramite l'utilizzo di espressioni regolari. Il secondo componente

CAPITOLO 5. IL MODULO STORYTELLING2KNOWLEDGE

ha il compito di estrarre le entità. Questo componente combina i risultati di quattro risorse, di cui due per l'estrazione di entità generiche, una per l'estrazione di entità associate alla base di conoscenza Wikidata e un gazetteer che fornisce un potenziale utilizzo sia per la ricerca di entità specifiche del dominio, sia per il task di entity linking. Infine, il componente permette la gestione di eventuali casi di sovrapposizione dovuti alla combinazione dei risultati e, a tal fine, è stata presentata una semplice euristica basata sulla scelta della menzione più lunga. L'ultimo componente del modulo è dedicato all'estrazione delle entità. Data la scarsità di risorse libere per la lingua italiana, per questa parte è stato implementato un semplice sistema OpenIE a regole basandosi su un lavoro simile per la lingua inglese. Il risultato crea delle triple soggetto-verbo-oggetto che sono successivamente filtrate per ottenere esclusivamente triple di interesse per le entità estratte precedentemente.

Capitolo 6

Risultati

Questo capitolo tratta l'analisi dei risultati ottenuti in fase di sperimentazione. L'obiettivo del progetto è di valutare l'integrazione di diverse risorse, al fine di creare un grafo di conoscenza partendo da tool già esistenti. In particolare, saranno confrontati i risultati relativi alla parte di estrazione di entità, dove è stato possibile integrare diverse risorse libere per la lingua italiana. A tal fine è stata implementata un'interfaccia che permette il controllo dell'attivazione dei tool, in modo da esaminarne il contributo finale.

Prima di descrivere la sperimentazione, è necessaria una premessa: HiStoryGraphia è un progetto in fase di sviluppo, soggetto a continue modifiche, in base alle necessità della piattaforma. Il materiale a disposizione per il test e l'analisi è limitato. Inoltre, solo recentemente è iniziata la fase di arricchimento della base di conoscenza.

Nella ricerca IE, per la lingua italiana, vi è una buona diversità di tool open-source legati all'estrazione delle entità. L'attenzione della sperimentazione si è indirizzata verso il confronto di tali risorse, per determinare il contributo di ogni risorsa e cercare la migliore combinazione.

Per la sperimentazione, sono state utilizzate tre narrazioni, prodotte da esperti storici, di cui alcuni estratti sono comparsi durante gli esempi dei capitoli

CAPITOLO 6. RISULTATI

precedenti:

- Valle Gesso – Strade di transito
- San Bernardo a Castelletto Stura
- Confraternita di Santa Croce a Entracque

Questi articoli rappresentano il dataset di partenza per il test sull'estrazione di entità. Al termine dell'estrazione, vi è un confronto con un database relazionale realizzato tramite l'utilizzo della piattaforma Omeka S¹. Il database contiene una classificazione, per ognuno dei tre saggi storici, del contenuto informativo atteso al termine: la divisione in frammenti e alcune entità presenti nei testi, classificate secondo lo schema dell'ontologia HSG.

Tuttavia, la risorsa non garantisce la corretta rappresentazione dei concetti, in quanto è stata realizzata per altri obiettivi e non è stata aggiornata con l'evoluzione della piattaforma. Infatti, vi è il problema che diversi concetti non sono definiti con la stessa forma con cui sono presenti nel testo. Per esempio, per distinguere le entità classificate come "Tappa", relative alle tappe dei percorsi dell'articolo su "Valle Gesso", è stato aggiunto il termine "tappa" al nome dell'entità inserita nel database (e.g., *Colle di Tenda* è stato trasformato in *Tappa Colle di Tenda*, *valle Gesso* è stato trasformato in *Tappa valle Gesso*). Per gestire questa inconsistenza sono stati realizzati due criteri, che saranno presentati più avanti, per il confronto tra entità: match totale e parziale.

L'analisi dei risultati è stata realizzata implementando un'interfaccia, che permettesse il controllo, attivazione e disattivazione, di tool ed euristiche, avviando un test su tutti gli articoli e confrontando le entità trovate con quelle presenti nel database.

¹<https://omeka.org/s/>

CAPITOLO 6. RISULTATI

[Articolo]: Cappella di San Bernardo

[Localizzazione]: Castelletto Stura

[Diocesi]: Diocesi di Asti; dal 1430 Diocesi di Mondovì e in seguito alla riorganizzazione post-napoleonica (1817) passa sotto la Diocesi di Cuneo [fonte: L. Berra, *Riordinamento delle Diocesi di Mondovì, Saluzzo, Alba e Fossano ed erezione della Diocesi di Cuneo nel 1817*, in "Bollettino della Società per gli Studi Storici, Archeologici e Artistici della Provincia di Cuneo", 36, 1955, p. 51]

[Dipendenze] Dal 1430 il territorio di Castelletto è assorbito in quello di più vasta pertinenza della "villanova" di Cuneo [fonte: R. Comba, *Due resoconti inediti della castellania di Cuneo (1388-1409)*, in "Bollettino della Società per gli Studi Storici, Archeologici e Artistici della Provincia di Cuneo", 67, 1972, pp. 32-33]. Nel 1619 Castelletto è infedato ad Amedeo Ponte di Scarnafigi; nel 1661 passa a Francesco Bartolomeo Sandri Trottì, marchese di Montanera, che nel 1668 lo vende a Giovanni Battista Lamberti, famiglia a cui apparterrà fino al periodo della conquista francese [fonte: M. Ristori, *Castelletto Stura. Storia civile e religiosa*, Cuneo 1977, pp. 56-73; G. Comino, *Castelletto Stura*, 1998 <https://www.archiviocasalis.it/localized-install/content/castelletto-stura>].

[Cronologia]: sulla base della lettura dei dati di stile, la cronologia di esecuzione cade intorno al 1480 [fonte: E. Brezzi Rossetti, *Giovanni Mazzucco*, in *La Pittura in Italia. Il Quattrocento*, Milano 1987, pp. 708-709]

[AUTORE]

Gli affreschi sono assegnati, sulla base dei dati di stile a Giovanni Mazzucco, un pittore attivo in area monregalese nella seconda metà del Quattrocento. [fonte: E. Brezzi Rossetti, *Giovanni Mazzucco*, in *La Pittura in Italia. Il Quattrocento*, Milano 1987, pp. 708-709]

[GEOGRAFIA ARTISTICA]

La prima data utile per ricostruire l'attività di Giovanni Mazzucco è il 1481, quando firma gli affreschi della cappella del Santo Sepolcro a Piozzo. Nel 1487 decora la chiesa di San Domenico a Peveragno. Al 1491 risale l'ultima opera firmata da Mazzucco, realizzata nel santuario del Brichetto a Morozzo, su commissione di Biagio Fauzone. Questo nucleo di affreschi tutti firmati da Giovanni Mazzucco, sono la base di riferimento per assegnare, al pittore e alla sua bottega, altre opere, affini per i dati di stile. [fonte: L. Marino, *Sulle tracce di Giovanni Mazzucco. Biografia, stile, confronti*, in *Il restauro della cappella di San Bernardo a Castelletto Stura*, Cuneo 2007, pp.59-78]

[OPERE ATTRIBUITE PER CONFRONTI STILISTICI]

Le attribuzioni più convincenti sono: un frammento con la Madonna con Bambino tra sant'Antonio abate e san Giovanni Battista, nella cappella di Sant'Antonio a San Michele Mondovì (1480); il trittico con la Madonna con il Bambino e san Pietro e sant'Antonio, nella cappella di San Pietro in Roncaglia a Benevagienna (commissionato nel 1485 dai fratelli Andrea e Giuliano de Capelinis); il polittico con la Madonna col Bambino e santi nella cappella di San Bernardino a San Michele Mondovì (1488) e la decorazione nell'ex convento dei domenicani nella frazione Bertini di Roccaforte Mondovì (circa 1480). [fonte: L. Marino, *Sulle tracce di Giovanni Mazzucco. Biografia, stile, confronti*, in *Il restauro della cappella di San Bernardo a Castelletto Stura*, Cuneo 2007, pp.59-78].

[AMBITO/FORTUNA]

Un gruppo di affreschi, da datare intorno alla seconda metà del Quattrocento, appaiono vicini allo stile del Mazzucco, ma non sono direttamente opera del pittore o della sua bottega. È il caso della Crocifissione nell'antica sacrestia di Niella Tanara; la Madonna col Bambino nel Santuario del Pasco a Villanova Mondovì: sono segni della fortuna di cui gode il linguaggio del pittore. [fonte: L. Marino, *Sulle tracce di Giovanni Mazzucco. Biografia, stile, confronti*, in *Il restauro della cappella di San Bernardo a Castelletto Stura*, Cuneo 2007, pp.59-78].

[Committenza]

L'edificazione della cappella e la sua decorazione ad affresco sono realizzati su commissione del parroco del paese con la partecipazione di alcuni notabili (che identifichiamo nelle figure affrescate ai margini della lunetta con l'Incoronazione della Vergine, abbigliate in vesti moderne e più caratterizzate nella fisionomia rispetto agli altri personaggi), illustri membri della confraternita di Santa Croce, da poco istituita in paese come filiazione della più importante congregazione cuneese (fonte: la confraternita di Santa Croce di Cuneo concede un aiuto in denaro il 7 marzo 1473).

[Funzione]

La cappella di San Bernardo, nelle pratiche di devozione del paese, è una delle tappe del percorso processionale che legava i luoghi sacri del paese: la chiesa parrocchiale, l'oratorio di San Sebastiano e quindi la cappella di San Bernardo che con la sua struttura aperta accoglieva per una breve sosta la processione. A questo itinerario si aggiunge la cappella di Sant'Anna edificata dopo la peste che afflisse il territorio tra il 1520 e il 1526 (dopo la peste del 1630 la cappella assocerà al titolo di Sant'Anna, quello di San Sebastiano, nuovo santo invocato contro il flagello). Nel Seicento si aggiungeranno altre due cappelle, Sant'Anna e San Grato, sulla strada verso Cuneo, e quella intitolata a San Francesco Saverio sulla strada verso Montanera. Il paese era così tutelato da santi in ogni strada di accesso, con le cappelle distribuite sui quattro punti cardinali. [fonte: G.M. Gazzola, *San Bernardo nel tessuto rituale di Castelletto Stura. La vita devazionale di un borgo rurale dal Medioevo a oggi*, in *Il restauro della cappella di San Bernardo a Castelletto Stura*, Cuneo 2007, pp. 101-110].

[Iconografia]

Le figure dipinte sulle pareti riassumono alcuni dei temi della spiritualità propria delle confraternite dedicate alla Santa Croce: è una vera narrazione per immagini. Sulle pareti laterali, le scene della Passione del Signore, con una forte accento sulle pratiche della penitenza, richiamano l'uso della flagellazione attuata dai confratelli in alcune processioni, l'Inferno e il Paradiso, sono l'esito del Giudizio Universale che i predicatori del tempo andavano preannunciando. Ai piedi di queste scene si trovano le sequenze della Cavalcata dei vizi (in relazione con l'Inferno) e le Opere di misericordia (in relazione con il Paradiso). Le teorie dei Santi (all'interno e sulla facciata) accompagnavano le litanie che ritmavano il cammino processionale in salmi penitenziali o il "Te Deum", a seconda della festività) [fonte: F. Quasimodo, *Tra Inferno e Paradiso. L'iconografia degli affreschi*, in *Il restauro della cappella di San Bernardo a Castelletto Stura*, Cuneo 2007, pp.37-58].

Figura 6.1: Narrazione: San Bernardo a Castelletto Stura

CAPITOLO 6. RISULTATI

[Articolo]: Entracque – Confraternita di Santa Croce

[Localizzazione]

Entracque; diocesi di Mondovì

[Committente]

La chiesa della confraternita viene edificata nel 1538, segno dell'affermazione di un notabilato che controlla la vita economica e sociale della comunità e sviluppatosi tra la metà del Quattrocento e la fine del Cinquecento, a seguito dell'incremento dei traffici commerciali. La rete economica favorisce anche la circolazione di artisti.

I fratelli di Santa Croce commissionano (1658 - 1660) un ciclo di dodici tele a Lorenzo Gastaldi.

La scelta del pittore, che in quel momento si trova a Monaco, e realizza molte opere per le valli del Nizzardo e il basso cuneese, si spiega con interessi economici della comunità e alle frequentazioni liguri-nizzarde della popolazione locale. Durante il Seicento gli entracquesi vendono ovini e bovini ai macelli di Grasse, Nizza e Genova. Sovente alcuni abitanti si recavano per affari a Tenda, raggiunta attraverso i colli della Finestra e del Sabbione. A partire dagli anni Settanta del Seicento è potenziata l'attività mercantile, a cui corrisponde il tentativo di rilanciare l'antica strada del colle delle Finestre, che in seguito al potenziamento della strada del colle di Tenda aveva perso la sua importanza. [Fonte: ARNEODO, D. DEIDDA, L. VOLPE, Attività pastorizia ed evoluzione degli equilibri socio-economici a Entracque (secoli XV-XVIII), in Entracque : una comunità alpina tra Medioevo ed Età moderna, Atti della giornata di studio, Entracque, 13 Aprile 1997, a cura di R. Comba, M. Cordero, Cuneo 1997, pp. 107-143][Fonte: B. Palmero, Entracque 2008, <https://www.archiviocasalis.it/localized-install/biblio/cuneo/entracque>].

I rapporti di Entracque con la valle della Vesubie, attraverso il colle delle Finestre, sulla strada di Nizza, determinarono almeno un altro episodio di committenza: il "mastro da bosco" (falegname) Giacomo Rosso di Lantosca realizza nel 1684 un armadio per la parrocchiale di Sant'Antonino.

[Iconografia]

Si tratta di dodici teleri raffiguranti episodi della vita di Cristo e della Vergine. Una scelta che risponde ai dogmi affermati dalla Chiesa cattolica della Controriforma, l'Immacolata Concezione della Vergine, la Trinità, il riscatto del genere umano attraverso il Sacrificio di Cristo. A questo si aggiungono i riferimenti alle pratiche liturgiche svolte dai fratelli durante il Giovedì Santo, rievocate nell'Ultima Cena e soprattutto nella Lavanda dei piedi.

Il modello decorativo di riferimento è quello della Confraternita di Santa Croce a Cuneo, dove nel 1626 sono allestite le tele con i Miracoli della Vera Croce dipinti nel 1626 da Giulio e Giovanni Battista Bruno (che a loro volta rimandano ai modelli decorativi di primo Seicento delle Casacce (oratori) genovesi).

[Pittore e geografia artistica]

Lorenzo Gastaldi è un pittore originario di Triora (1625-1690); dal 1651 si trasferisce a Monaco dove si ritaglia uno spazio a corte, per poi passare a Nizza e ritornare a Triora nel 1676. Molte sono le opere da lui realizzate, destinate alle valli del Nizzardo e del basso Cuneese.

Le sue opere sono presenti a La Brigue (per la parrocchiale, Immacolata con i santi Sebastiano, Giuseppe, Rocco e Carlo Borromeo), a Peillon (per la parrocchiale, Madonna del Rosario), a Contes in Val Paillon (per la parrocchiale, Pentecoste); a Triora (Natività del Battista per l'Oratorio di San Giovanni Battista);

in valle Stura a Bersezio (per la parrocchiale, Madonna con il Bambino e santi; Sacra Famiglia); Pontebernardo (per la parrocchiale, Madonna con il Bambino, Santi e la Sindone; Moiola; Vernante (per la parrocchiale, la Sindone, 1671; Roccavione (per la cappella dei Santi Rocco e Biagio, 1676; per la parrocchiale uno stendardo per la Compagnia del Rosario, 1678); Demonte (parrocchiale, Madonna con il Bambino e i santi Marco evangelista, Chiara, Grato e Barbara); nella valle della Tinée, a Isola (per la parrocchiale, Incoronazione della Vergine e i santi Bernardo e Lorenzo). [Fonte: M. Bartolotti, Due episodi figurativi del Seicento a Entracque: l'attività del pittore Lorenzo Gastaldi e le tele dell'Apostolato nella parrocchiale, in Entracque : una comunità alpina tra Medioevo ed Età moderna, Atti della giornata di studio, Entracque, 13 Aprile 1997, a cura di R. Comba, M. Cordero, Cuneo 1997, pp. 193-214].

Figura 6.2: Narrazione: Entracque – Confraternita di Santa Croce

CAPITOLO 6. RISULTATI

[Articolo]:Valle Gesso – Strade di transito

[Quadro politico]

Per evitare gli esosi pedaggi imposti dai signori di Tenda alle merci che transitavano nei territori da loro controllati, I Savoia cercano un percorso alternativo alla strada del Col di Tenda per raggiungere Nizza. L'itinerario scelto dai Savoia è quello che attraversava la valle Gesso, e il Colle di Finestra, forse meno agevole, ma utile ad evitare i pedaggi (1200 scudi all'anno solo per il trasporto del sale) e quindi molto praticato, favorendo così, a partire dalla seconda metà del secolo XV, lo sviluppo economico della Valle e di Entracque. E' un percorso che verso la metà del Quattrocento era stato addirittura raddoppiato, con il velleitario e un po' folle progetto attuato da Paganino Dal Pozzo attraverso il colle che oggi da lui prende il nome, il Pagari, e che supera i 2800 metri di altitudine.

Va tuttavia detto che, nonostante i pedaggi imposti, la strada del colle di Tenda, attraverso la valle Roya, rimase la strada più battuta, perché più vantaggiosa rispetto agli altri itinerari sia per la minor altitudine che per la migliore percorribilità durante i mesi invernali.

Con l'annessione nel 1579 della contea di Tenda ai domini dei Savoia, la strada del sale, da Torino a Nizza per il col di Tenda, si libera dai dazi che fino a quel momento i signori di Tenda avevano inflitto al traffico delle merci e diventa la via principale per traffici e comunicazioni. E' l'avvio della crisi di Entracque, e del suo territorio, escluso dalla scelta di privilegiare la via passante per Limone e la val Roia per il trasporto di merci e per il rifornimento di sale, a scapito del più problematico collegamento attraverso gli alti passi della val Gesso. [fonte: B. Palmero, *Consenso e contrattazione politica lungo la direttrice del col di Tenda (1586-1754). I comuni della val Roya e la progettazione della strada*, in Bollettino storico Bibliografico subalpino, XCIII, 1995, pp. 507-546].

Nel 1861 la Contea di Nizza diventa francese e nasce il Regno d'Italia: sul valico viene costruita, da parte italiana, una serie di massicce opere difensive: è il Campo Trincerato del Colle di Tenda. Durante la seconda guerra mondiale, tra l'autunno del 1943 e la primavera del 1945, il territorio fortificato del valico è occupato dalle truppe tedesche e da gruppi fascisti. Il 25 aprile 1945 il tunnel e il colle sono conquistati dalla Division Française Libre del generale De Gaulle. Anche sul fronte italiano le truppe nazifasciste sono sconfitte ed è la fine della seconda guerra mondiale. Nel settembre 1947, con il trattato di Parigi, il confine tra Italia e Francia è spostato verso nord, consegnando ai francesi Briga e Tenda e buona parte del territorio del Colle. [I. Borgna, J-L. Fontana, G. Oppi, *La viabilità transfrontaliera. I colli principali nella storia*, in *Atlante transfrontaliero del Patrimonio culturale*, Villeneuve-Loubet 2013, pp. 27-32]

[Area di transito]

Gli itinerari che si diramano verso la Francia sono due:

Il primo da Borgo San Dalmazzo attraversava la valle Vermenagna e valicava il colle di Tenda in direzione di Nizza. Percorreva la valle Roya fino a Breglio e per il colle del Brouis raggiungeva Sospello; da qui valicando il Braus, scendeva a L'Escarene e sbucava a Nizza.

Questo itinerario, con alcune varianti, sarà modificato tra il 1610 e il 1714, assumendo il nome di Route Royale (Strada Reale), percorribile da carri e artiglierie. Tra il 1780 e il 1788, sono realizzati nuovi lavori di ampliamento e la via del Tenda diventa la prima strada carrozzabile che attraversa le Alpi. Le trasformazioni facilitano gli scambi commerciali a scapito del meno agevole Colle di Finestra. Dal 1815, tramontata l'occupazione francese (1792-1815), Vittorio Emanuele I fa istituire il primo servizio di trasporto pubblico da Torino a Nizza, con diligence trisettimanali. [fonte: J.-L. Fontana, *Les capitulations des forts de Nice, Villafranche et Mont-Alban*, in *Au Coeur des Alpes: Utrecht, actes du colloque de Jausiers, Colmars et Etraunes 14,15,16 Septembre 2012*, Barcellonnette 2013.].

Nel 1882, l'apertura del tunnel al di sotto del valico tronca l'importanza dell'antica via commerciale del sale. Alla sua inaugurazione, il traforo del Colle di Tenda era il tunnel stradale più lungo mai costruito (3182 m). Nel 1898 è ultimata anche una galleria ferroviaria, molto più lunga di quella stradale (8099 m). [fonte: F.Collidà, M. Gallo, A. Mola, *Cuneo-Nizza : storia di una ferrovia*, Cuneo 1982].

Il secondo percorso, parallelo alla via del Colle di Tenda, imboccava la valle Gesso e salendo ad Entracque valicava il colle delle Finestre e si immetteva nella val Vesubie, attraversava St. Martin e Lantosca. Da qui si snodavano due tragitti per Nizza: uno passante per Lévens e l'altro per Utelle.

Figura 6.3: Narrazione: Valle Gesso – Strade di transito

CAPITOLO 6. RISULTATI

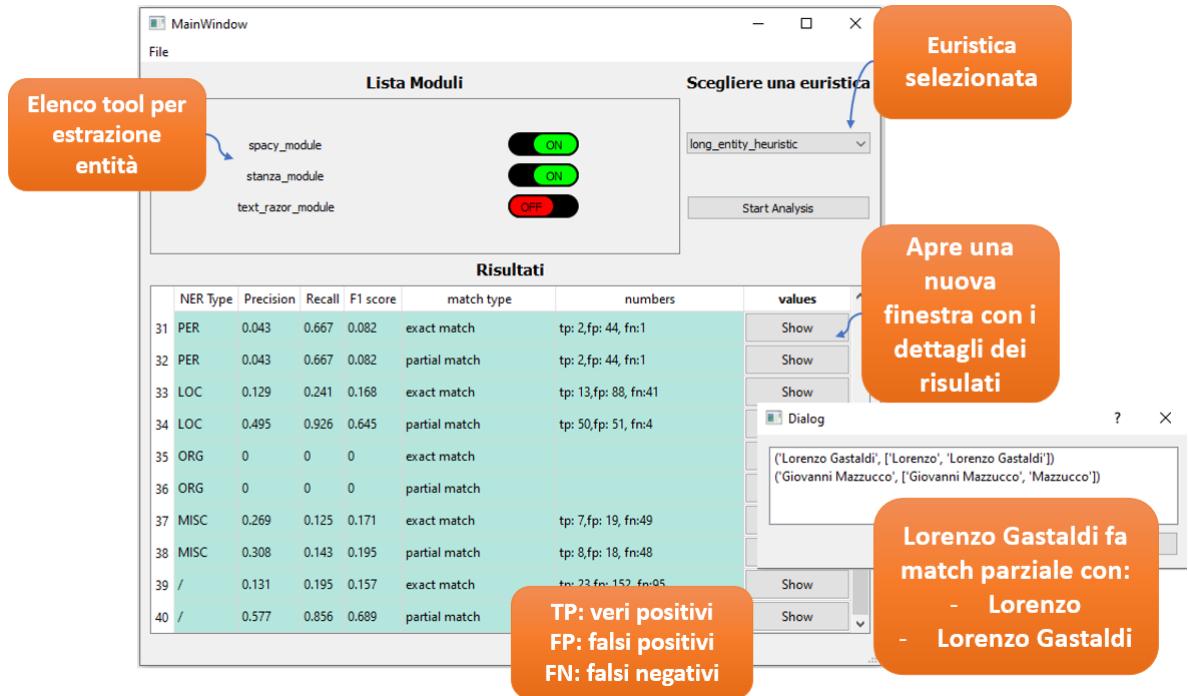


Figura 6.4: Interfaccia valutazione modulo estrazione entità.

L’interfaccia, presentata in figura 6.4, è divisa in due sezioni: nella metà superiore è presente la lista dei tool utilizzabili, ognuno affiancato da un interruttore per attivazione e disattivazione, e un menù a tendina per selezionare l’euristica da utilizzare nella gestione delle sovrapposizioni. La parte inferiore descrive i risultati ottenuti dal test tramite una tabella. Essa contiene due righe per ogni tipo di entità riconoscibile dai moduli, una con i dati risultati dal match parziale e una con i dati risultati dal match totale.

L’esecuzione del test prevede una prima fase in cui gli articoli sono passati al wrapper, che ne estrae i frammenti. Successivamente, per ogni frammento, si passa il testo al modulo *Estrattore entità*, che ne estrae le entità. Come risultato si ottengono, quindi, una lista di entità assegnate ad una classe, se estratte da un classificatore NER, o ad un identificatore, se estratte da un modulo EL.

Al momento, il gazetteer è predisposto per futuri utilizzi, di conseguenza i moduli da confrontare sono: Stanza, spaCy e TextRazor. I primi due sono moduli NER, che quindi estraggono classi generiche: Persone(PER), Località(LOC), Organizza-

CAPITOLO 6. RISULTATI

zioni(ORG) e nel caso di spaCy anche Misto(MISC), ovvero una classe generale. L'ultimo, invece, è un modulo EL, il quale restituisce identificatori in Wikidata. In questo caso, senza un mapping tra le ontologie HSG e Wikidata, diventa difficile valutare la classificazione delle entità. Per questo, è stato aggiunto un confronto che non tiene in considerazione le classi ma solo la presenza o meno dell'entità, in figura 6.4 è rappresentato dalle righe con NER Type uguale a "\".

Terminata l'estrazione, si estraggono e mappano, tramite query specifiche per ogni classe, le entità presenti nel database ad ognuna delle classi generiche estratte. In questo modo si definisce un set di entità attese, per ogni classe, su cui poter calcolare i valori statistici di valutazione.

Per determinare la correttezza dell'estrazione, il confronto tra le entità estratte e presenti del database avviene confrontando i nomi, poiché il database non traccia informazioni, quali ad esempio, la posizione dell'entità nel testo.

Tuttavia, può capitare, in particolare per le entità legate alla località, che il nome scelto nel database non rispecchi la forma con cui sono presentate nel testo. Ne consegue che diventa impossibile verificare la correttezza dell'estrazione basandosi su un confronto stretto delle due stringhe e serve una qualche forma di match meno precisa, che dia la possibilità di un confronto parziale.

Si introduce il match parziale: in questo caso due stringhe sono considerate come uguali se una delle due è contenuta nell'altra, ad esempio, se l'estrattore riconoscesse "valle Gesso", questa entità soddisfarebbe il confronto con "Tappa valle Gesso" in quanto il nome dell'entità estratta è contenuta nell'entità attesa.

L'utilizzo di un match largo, che quindi ha possibilità di sbagliare, ha il vantaggio di poter analizzare singolarmente i vari casi. Nell'interfaccia, è stato aggiunto un pulsante, che genera una finestra con i risultati associati ad ogni riga di valutazione, vedi figura 6.4. In questo modo, è possibile controllare singolarmente i match errati e creare nuove criteri di confronto parziale più vincolanti.

CAPITOLO 6. RISULTATI

L'analisi dei risultati avviene attraverso l'utilizzo di tre metriche statistiche:

- **Precisione:** misura che definisce l'accuratezza, ovvero quante entità sono state correttamente classificate rispetto al totale delle classificazioni di quella classe.

$$Precisione = \frac{VP}{VP + FP}$$

- **Recupero:** misura che definisce la sensibilità, ovvero quante entità di una data classe sono state recuperate.

$$Recupero = \frac{VP}{VP + FN}$$

- **F1 Score:** misura che combina precisione e recupero per rappresentare l'accuratezza di un test.

$$F_1 Score = \frac{Precisione \times Recupero}{Precisione + Recupero}$$

I valori presenti nelle formule matematiche, nel contesto del progetto, ovvero l'estrazione e classificazione delle entità, assumono i seguenti significati:

- Veri positivi (VP): rappresenta il numero di entità estratte dal testo che il sistema ha correttamente riconosciuto e classificato.
- Falsi negativi (FN): rappresenta il numero di entità presenti nel database ma che il sistema non è riuscito a riconoscere o classificare correttamente dal testo.
- Falsi positivi (FP): rappresenta il numero di entità spurie, ovvero che il sistema ha estratto e classificato ma non presenti nel database. Tuttavia, essendo un database integrato ma non finalizzato al test, non tutte le entità presenti nei documenti sono state classificate e dunque il numero dei falsi positivi è leggermente superiore al previsto. Per esempio, non è presente *Biagio Faugzone*, che commissiona una decorazione a Giovanni Mazzucco, nominato nel frammento di geografia artistica nel saggio di San Bernardo Castelletto

CAPITOLO 6. RISULTATI

Stura; oppure, i pittori *Giulio e Giovanni Battista Bruno* presenti nel saggio "Entracque – Confraternita di Santa Croce".

I risultati, estratti tramite l'applicazione, sono riportati nelle seguenti tabelle: sono state definite quattro tabelle, tre legate alle classi di entità estraibili dai moduli NER (Persone, Località, Misto) e una generica per valutare se l'entità è stata effettivamente trovata, indipendentemente dalla classe a cui è stata assegnata, permettendo quindi di valutare il contributo del modulo EL (TextRazor).

È stata esclusa dai risultati la classe Organizzazioni, in quanto forniva risultati nulli: nel database sono presenti cinque istanze legate a questa classe (*Confraternita Santa Croce Entracque*, *Confraternita Santa Croce Castelletto Stura*, *Confraternita Santa Croce Cuneo*, *Dinastia Savoia*, *Signori di Tenda*) ma nessun modulo le ha riconosciute come organizzazioni.

Per ogni tabella, sono riportati i valori di precisione, recupero e f1-score, relativi all'estrazione di una specifica classe e al variare della combinazione dei moduli e dell'euristica. Inoltre, sono state unite le righe associate a risultati ridondanti: in assenza dell'euristica, l'attivazione o disattivazione del modulo TextRazor non influenza l'esito dei risultati delle classi, in quanto il modulo non esegue il task di classificazione. Al contrario, in presenza dell'euristica, può capitare che i risultati dei moduli NER siano sovrascritti da un match più preciso della risorsa TextRazor. Ad esempio, l'estrazione dell'entità "*Giovanni*", di tipo Persona, potrebbe essere in sovrapposizione con l'entità "*Oratorio di San Giovanni Battista*", estratta dal modulo EL. In questo caso la presenza del modulo TextRazor, può determinare variazioni nelle statistiche.

La prima tabella presenta i dati relativi alla classe persone. La poca varianza dei dati, in particolare il Recupero, è dovuta al fatto che nel database sono presenti solo tre persone (*Lorenzo Gastaldi*, *Parroco Castelletto Stura*, *Giovanni Mazzucco*). Il basso valore di precisione è influenzato da due cause: non tutte le entità persona presenti nei testi sono rappresentate nel database. Per esempio, non sono presenti i pittori "*Giulio e Giovanni Battista Bruno*" e il falegname "*Giacomo Rosso*" tratti,

CAPITOLO 6. RISULTATI

Risultati per la classe Persone (PER)							
Precisione	Recupero	F1 Score	match	spaCy	Stanza	TextRazor	Euristica
0.048	0.667	0.089	totale	on	on	on	long
0.048	0.667	0.089	parziale	on	on	on	long
0.033	0.667	0.063	totale	on	on	on/off	/
0.033	0.667	0.063	parziale	on	on	on/off	/
0.043	0.667	0.082	totale	on	on	off	long
0.043	0.667	0.082	parziale	on	on	off	long
0.062	0.667	0.114	totale	off	on	on	long
0.062	0.667	0.114	parziale	off	on	on	long
0.050	0.667	0.093	totale	off	on	on/off	/
0.050	0.667	0.093	parziale	off	on	on/off	/
0.045	0.667	0.085	totale	on	off	on	long
0.045	0.667	0.085	parziale	on	off	on	long
0.041	0.667	0.077	totale	on	off	on/off	/
0.041	0.667	0.077	parziale	on	off	on/off	/

Tabella 6.1: Risultati classificazione al variare dei moduli ed euristiche attive per la classe Persone

rispettivamente, dai frammenti iconografia e committenza della narrazione Confraternita di Santa Croce a Entracque.

Inoltre, i moduli sono ingannati dalla presenza di diversi titoli di opere d'arte contenenti nomi propri di persone, come ad esempio "*Immacolata con i santi Sebastiano, Giuseppe, Rocco e Carlo Borromeo*" o "*Vita di Gesù*". Si tratta di opere poco conosciute, quasi sempre non presenti in basi di conoscenza come Wikipedia, i cui dati sono comunemente utilizzati per l'addestramento di queste risorse. Questo spinge i moduli a riconoscere solo i nomi interni al titolo dell'opera, classificandoli erroneamente come persone.

CAPITOLO 6. RISULTATI

Risultati per la classe Località (LOC)							
Precisione	Recupero	F1 Score	match	spaCy	Stanza	TextRazor	Euristica
0.131	0.241	0.170	totale	on	on	on	long
0.505	0.926	0.654	parziale	on	on	on	long
0.109	0.278	0.157	totale	on	on	on/off	/
0.380	0.963	0.545	parziale	on	on	on/off	/
0.129	0.241	0.168	totale	on	on	off	long
0.495	0.926	0.645	parziale	on	on	off	long
0.114	0.222	0.151	totale	off	on	on	long
0.467	0.907	0.616	parziale	off	on	on	long
0.109	0.222	0.146	totale	off	on	on/off	/
0.464	0.944	0.622	parziale	off	on	on/off	/
0.155	0.241	0.188	totale	on	off	on	long
0.595	0.926	0.725	parziale	on	off	on	long
0.148	0.241	0.183	totale	on	off	on/off	/
0.568	0.926	0.704	parziale	on	off	on/off	/

Tabella 6.2: Risultati classificazione al variare dei moduli ed euristiche attive per la classe Località

La seconda tabella riguarda il riconoscimento delle località. Questi risultati sono quelli maggiormente influenzati dal doppio criterio di match, parziale o totale. La causa è dovuta alla presenza di molte entità, che per evidenziare il fatto che fossero luoghi legati ad una tappa di un percorso, è stato modificato il nome durante il salvataggio nel database.

La successiva tabella è relativa ai risultati per la classe mista, questi dati sono relativi esclusivamente alla risorsa spaCy, in quanto è l'unica a definire tale classe. Il risultato atteso, estratto dal database per il confronto, considera come appartenenti a questa classe tutte le classi dell'ontologia HSG non riferite a persone, località o organizzazioni.

CAPITOLO 6. RISULTATI

Risultati per la classe Mista (MISC)							
Precisione	Recupero	F1 Score	match	spaCy	Stanza	TextRazor	Euristica
0.280	0.125	0.173	totale	on	on	on	long
0.320	0.143	0.198	parziale	on	on	on	long
0.269	0.125	0.171	totale	on	on/off	on/off	/
0.308	0.143	0.195	parziale	on	on/off	on/off	/
0.269	0.125	0.171	totale	on	on	off	long
0.308	0.143	0.195	parziale	on	on	off	long
0.280	0.125	0.173	totale	on	off	on	long
0.320	0.143	0.198	parziale	on	off	on	long

Tabella 6.3: Risultati classificazione al variare dei moduli ed euristiche attive per la classe Mista

L'ultima tabella riporta i dati riguardo il riconoscimento delle entità estratte, indipendentemente dalla classe a cui sono associate.

La finalità del modulo è di aiutare l'estrazione manuale, suggerendo entità estratte automaticamente. Il riconoscimento di un'entità erroneamente classificata, è comunque un'informazione in più rispetto al non trovarla. L'utente ha la possibilità di correggere la classificazione con quella appropriata.

Inoltre, i dati che seguono permettono di analizzare il contributo dei moduli EL, che, altrimenti, non sarebbe possibile valutare basandosi sulle classi.

CAPITOLO 6. RISULTATI

Risultati riconoscimento entità senza classificazione)							
Precisione	Recupero	F1 Score	match	spaCy	Stanza	TextRazor	Euristica
0.114	0.263	0.159	totale	on	on	on	long
0.419	0.966	0.585	parziale	on	on	on	long
0.094	0.271	0.139	totale	on	on	on	/
0.334	0.966	0.497	parziale	on	on	on	/
0.131	0.195	0.157	totale	on	on	off	long
0.577	0.856	0.689	parziale	on	on	off	long
0.117	0.203	0.149	totale	on	on	off	/
0.493	0.856	0.625	parziale	on	on	off	/
0.112	0.254	0.155	totale	off	on	on	long
0.418	0.949	0.580	parziale	off	on	on	long
0.096	0.254	0.139	totale	off	on	on	/
0.358	0.949	0.520	parziale	off	on	on	/
0.114	0.263	0.159	totale	on	off	on	long
0.421	0.966	0.586	parziale	on	off	on	long
0.100	0.271	0.146	totale	on	off	on	/
0.356	0.966	0.521	parziale	on	off	on	/
0.141	0.195	0.164	totale	on	off	off	/
0.620	0.856	0.719	parziale	on	off	off	/
0.110	0.144	0.125	totale	off	on	off	/
0.617	0.805	0.699	parziale	off	on	off	/
0.106	0.237	0.147	totale	off	off	on	/
0.424	0.949	0.586	parziale	off	off	on	/

Tabella 6.4: Risultati riconoscimento delle entità al variare dei moduli ed euristiche attive, senza la classificazione

CAPITOLO 6. RISULTATI

Dai dati raccolti, sono emerse le seguenti considerazioni:

- I risultati relativi al riconoscimento delle persone restano uguali al variare del tipo di confronto, totale o parziale, utilizzato. Tuttavia, questo dato può derivare dal basso numero di persone presenti nel database.
- L'utilizzo di modelli per entità generiche non sono sufficienti per domini specifici. L'ontologia HSG, definisce molte classi che non sono riconoscibili, in quanto non rientrano nelle tre classi generiche Persone, Organizzazioni e Località. Ad esempio, gli intervalli temporali o le opere d'arte.
- Esiste un trade-off, tra precisione e recupero, legato all'utilizzo dell'euristica. L'attivazione tende a migliorarne la precisione, tuttavia riduce il valore di recupero.

Essendo il modulo finalizzato al suggerimento di entità, si potrebbe utilizzare la combinazione di tutti i componenti senza euristica. Incrementando il valore di recupero, aumenterebbero le entità riconosciute e si darebbe la libertà all'utente di selezionare tra i suggerimenti solo quelli corretti.

Inoltre, l'aggiunta al gazetteer, attualmente predisposto ma privo di dati, delle principali opere d'arte, se combinato con l'euristica, permetterebbe di migliorare i risultati legati al riconoscimento di persone.

Infine, è stata costruita una seconda interfaccia, destinata a utilizzi futuri, per la valutazione manuale dei risultati, in particolare per i tool EL. La parte superiore dell'interfaccia è dedicata alla scrittura del saggio. Successivamente, è possibile avviare l'analisi, attualmente impostata con tutte le risorse attive e l'utilizzo dell'euristica della menzione più lunga. Nella metà inferiore viene visualizzata, al termine dell'analisi, una lista di entità estratte e il rispettivo identificatore, rappresentato in questo caso dal link a Wikidata. L'utente ha a disposizione due pulsanti per segnalare se l'estrazione è corretta o meno.

CAPITOLO 6. RISULTATI

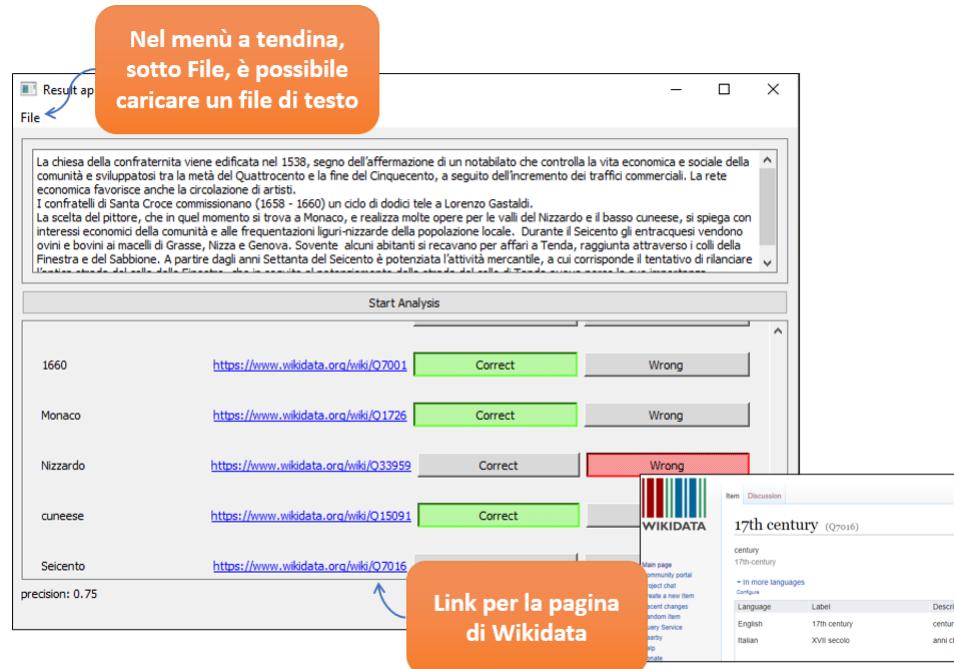


Figura 6.5: Interfaccia valutazione manuale.

6.1 Conclusioni di capitolo

In questo capitolo sono stati riportati i dati raccolti in seguito alla sperimentazione. A tal fine, è stata implementata un'interfaccia per gestire l'attivazione dei tool e per scegliere l'euristica da utilizzare. Successivamente, le entità estratte sono state confrontate con i dati presenti in un database relazionale. Tuttavia, la risorsa è stata realizzata per altri obiettivi: non tutte le risorse presenti nelle narrazioni sono memorizzate nel database e alcuni nomi sono stati modificati rispetto alla forma con cui sono presenti nei testi. Per gestire questa inconsistenza sono stati utilizzati due approcci per il confronto: il match totale e parziale. In seguito, sono state calcolate tre metriche per valutare i risultati ottenuti: precisione, recupero e f1-score.

In conclusione, è emerso un trade-off legato all'utilizzo dell'euristica, che in caso di attivazione migliora il valore della precisione ma decrementa il valore legato al recupero. Inoltre, l'utilizzo di risorse generiche non è sufficiente per riconoscere entità più specifiche come le opere d'arte o gli intervalli temporali.

CAPITOLO 6. RISULTATI

Infine, è stata presentata una seconda interfaccia, che permette una valutazione manuale, in particolare per verificare la correttezza dei risultati EL.

Capitolo 7

Conclusioni e sviluppi futuri

In questo lavoro è stato proposto uno studio legato all'estrazione della conoscenza finalizzato all'integrazione con il progetto HiStoryGraphia.

Il progetto, ancora nella fase iniziale, raccoglie e analizza lo stato dell'arte per la ricerca IE in lingua italiana. Si sono confrontate diverse risorse, legate all'estrazione delle entità, per studiarne l'efficienza, singolarmente o combinandone i risultati, nel contesto storico.

Il modulo implementato è progettato in maniera modulare per rendere semplice l'aggiunta o la sostituzione di nuove risorse, al fine di rendere possibile una continua e futura analisi al passo con l'evoluzione della ricerca in lingua italiana e degli obiettivi della piattaforma. Inoltre, tramite l'applicazione è anche possibile analizzare quali sono le classi di entità più problematiche, per via del dominio specifico, permettendo di costruire soluzioni dedicate, come ad esempio il gazetteer già predisposto.

In conclusione, il sistema proposto crea una base di partenza per una soluzione di estrazione della conoscenza continuamente migliorabile e facilmente replicabile in altri domini.

Un prossimo miglioramento potrebbe essere l'introduzione di altre risorse mul-

CAPITOLO 7. CONCLUSIONI E SVILUPPI FUTURI

tilingua per il riconoscimento di entità, come ad esempio il progetto Heideltimer¹ per l'estrazione di intervalli temporali più accurati.

Legato al task dell'estrazione delle relazioni, sarebbe interessante costruire alcune regole per il Semantic Role Labelling. Il progetto HiStoryGraphia contiene un insieme di eventi e situazioni che difficilmente sarebbero riconoscibili da semplici relazioni binari. Ad esempio, il concetto di committenza prevede il coinvolgimento di attori quali: agente, destinatario, oggetto, luogo, data ecc. Una soluzione basata su frame sarebbe più adatta a riconoscere queste situazioni.

Inoltre, il progetto HiStoryGraphia ha iniziato recentemente la popolazione della base di conoscenza. In futuro, con l'aumentare dei dati al suo interno, sarebbe possibile implementare soluzioni ad hoc, tramite la costruzione di dataset per l'addestramento di modelli su classi e relazioni specifiche dell'ontologia HSG.

¹<https://github.com/HeidelTime/heideltimer>

Elenco delle figure

1.1	Estratto narrazione: Entracque – Confraternita di Santa Croce	3
1.2	Parte del grafo di conoscenza relativo a 1.1	3
2.1	Il ciclo di vita dell'informazione storica [2]	7
2.2	Classificazione dei documenti storici in base al livello di struttura [2] . . .	10
2.3	Esempio classificazione in base alla struttura [9]	12
2.4	Word cloud per "Jack and Jill went up the hill" e "Humpty Dumpty sat on wall" generata tramite: https://www.wordclouds.com	14
2.5	(a) Franklin Delano Roosevelt, 32° presidente degli Stati Uniti [11] (b) Theodore Roosevelt, 26° presidente degli Stati Uniti [12]	16
3.1	Pagina web contenente i prefissi telefonici di alcuni paesi [22]	22
3.2	Risultato e procedura per ottenere l'informazione contenuta nella figura	
3.1	[22]	23
3.3	Esempi NER. Classi: PER (Persona), LOC (Località), ORG (Organizzazione) (a) Esempio ufficiale CoNLL-2003 [28] (b) Esempio in lingua italiana tramite la risorsa Stanza [29]	25
3.4	Esempi rappresentazioni in formati BIO e BIOES [30]	26
3.5	Esempio NED [31]	28

ELENCO DELLE FIGURE

3.6	Voci di un gazetteer per riconoscere un concetto in diverse lingue [33]	31
3.7	Esempio Open IE [35]	33
3.8	Esempio albero a dipendenze per la frase "Mario Biondi ama la pizza"	34
4.1	Logo HSG	36
4.2	Estratto narrazione. In giallo i divisori che identificano i diversi frammenti di testo, in grigio le fonti associate ad ogni parte di testo	38
4.3	Rappresentazione concettuale HSG	38
4.4	Rappresentazione dettagliata dei concetti di macro livello in HSG	39
4.5	Esempio articolo: Confraternita di Santa Croce a Entracque	40
4.6	Esempio rappresentazione concettuale dell'articolo in figura 4.5	41
5.1	Esempio estratto di articolo, tratto da San Bernardo Castelletto Stura	46
5.2	Architettura modulo Storytelling2Knowledge	46
5.3	Estratto iniziale narrazione: in giallo i divisori delle sezioni di testo, in grigio le fonti associate ad ogni frammento	47
5.4	Logo spaCy	51
5.5	Pipeline spaCy [41]	52
5.6	Logo Stanza	52
5.7	Pipeline Stanza [43]	53
5.8	Esempio rappresentazione in formato BIOES	54
5.9	Logo TextRazor	54
6.1	Narrazione: San Bernardo a Castelletto Stura	64
6.2	Narrazione: Entracque – Confraternita di Santa Croce	65

ELENCO DELLE FIGURE

6.3	Narrazione: Valle Gesso – Strade di transito	66
6.4	Interfaccia valutazione modulo estrazione entità.	67
6.5	Interfaccia valutazione manuale.	76

Elenco delle tabelle

6.1	Risultati classificazione al variare dei moduli ed euristiche attive per la classe Persone	71
6.2	Risultati classificazione al variare dei moduli ed euristiche attive per la classe Località	72
6.3	Risultati classificazione al variare dei moduli ed euristiche attive per la classe Mista	73
6.4	Risultati riconoscimento delle entità al variare dei moduli ed euristiche attive, senza la classificazione	74

Elenco dei codici

5.1 Esempio di parte del risultato dell'estrazione di frammenti dall'articolo precedente	49
---	----

Bibliografia

- [1] O. Boonstra, L. Breure, and P. Doorn, "Past, present and future of historical information science," *Historical Social Research*, vol. 29, no. 2, pp. 4–132, 2004.
- [2] A. Meroño-Peña, A. Ashkpour, M. Erp, K. Mandemakers, L. Breure, A. Scharnhorst, S. Schlobach, and F. Harmelen, "Semantic technologies for historical research: A survey," *Semantic Web*, vol. 6, pp. 539–564, 10 2014.
- [3] P. Quaresma and M. J. B. Finatto, "Information extraction from historical texts: a case study," in *DHandNLP@PROPOR*, 2020.
- [4] S. Denbo, "Linking the past: history and the semantic web," 2014.
<https://www.historians.org/publications-and-directories/perspectives-on-history/october-2014/linking-the-past>.
- [5] O. Signore, "Introduzione al semantic web," 01 2008.
- [6] T. Berners-Lee, "Axioms of web architecture: Metadata." Aviable at <https://www.w3.org/DesignIssues/Metadata.html> (2022/04/24), 1997.
- [7] H. Nguyen and T. Cao, "Enriching ontologies for named entity disambiguation," 05 2022.
- [8] "Primary, secondary and tertiary sources: Types of information sources."
<https://libguides.jcu.edu.au/scholarly-sources>.
- [9] J. Cardoso, *Developing Dynamic Packaging Applications Using Semantic Web-Based Integration*. 01 2005.

BIBLIOGRAFIA

- [10] R. Brath, "Literal encoding: Text is a first-class data encoding," *CoRR*, vol. abs/2009.02374, 2020. <https://arxiv.org/abs/2009.02374>.
- [11] L. A. Perskie, "Gift of beatrice perskie foxman and dr. stanley b. foxman." Available at <https://www.flickr.com/photos/fdrlibrary/8145288140/> (2022/04/24), August 21, 1944. License: <https://creativecommons.org/licenses/by/2.0/>.
- [12] P. Bros, "Theodore roosevelt, pres. u.s., 1858-1919." Available at <http://loc.gov/pictures/resource/ppmsca.35645/> (2022/04/24), 1904. License: <https://creativecommons.org/publicdomain/mark/1.0/>.
- [13] Y. Kolikant and S. Pollack, "Collaborative, multi-perspective historical writing: The explanatory power of a dialogical framework," *Dialogic Pedagogy: An International Online Journal*, vol. 7, 09 2019.
- [14] Wikipedia, "Libro bianco (palestina)," 2021. [Online; controllata il 1-marzo-2022].
- [15] D. Jurafsky and J. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, vol. 2. 02 2008.
- [16] S. Singh, "Natural language processing for information extraction," *CoRR*, vol. abs/1807.02383, 2018. [http://arxiv.org/abs/1807.02383](https://arxiv.org/abs/1807.02383).
- [17] D. Wimalasuriya and D. Dou, "Ontology-based information extraction: An introduction and a survey of current approaches," *J. Information Science*, vol. 36, pp. 306–323, 05 2010.
- [18] K. Kaiser and S. Miksch, "Information extraction a survey," 2005.
- [19] N. Guarino, "Formal ontologies and information systems," 06 1998.
- [20] S. J. Russell and P. Norvig, *Artificial Intelligence: a modern approach*. Pearson, 3 ed., 2009.

BIBLIOGRAFIA

- [21] F. Ciravegna, A. Dingli, Y. Wilks, and D. Petrelli, "Adaptive information extraction for document annotation in amilcare," in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '02*, (New York, NY, USA), p. 451, Association for Computing Machinery, 2002. <https://doi.org/10.1145/564376.564492>.
- [22] N. Kushmerick, D. S. Weld, and R. B. Doorenbos, "Wrapper induction for information extraction," in *IJCAI*, 1997.
- [23] A. Gentile, Z. Zhang, and F. Ciravegna, "Self training wrapper induction with linked data," pp. 285–292, 09 2014.
- [24] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. Trippe, J. Gutierrez, and K. Kochut, "A brief survey of text mining: Classification, clustering and extraction techniques," 07 2017.
- [25] M. Rovera, F. Nanni, S. P. Ponzetto, and A. Goy, *Domain-specific Named Entity Disambiguation in Historical Memoirs*, pp. 287–291. 01 2017.
- [26] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, pp. 3–26, 2007.
- [27] K. Adnan and R. Akbar, "An analytical study of information extraction from unstructured and multidimensional big data," *Journal of Big Data*, vol. 6, 10 2019.
- [28] Aviable at <https://www.clips.uantwerpen.be/conll2003/ner/> (2022/04/24).
- [29] Aviable at <http://stanza.run/> (2022/04/24).
- [30] Aviable at [https://en.wikipedia.org/wiki/Inside-outside-beginning_\(tagging\)](https://en.wikipedia.org/wiki/Inside-outside-beginning_(tagging)) (2022/04/24).
- [31] Aviable at <https://blogs.oracle.com/ai-and-datasience/post/named-entity-disambiguation-with-knowledge-graphs> (2022/04/24).

BIBLIOGRAFIA

- [32] J. Martinez-Rodriguez, A. Hogan, and I. Lopez-Arevalo, "Information extraction meets the semantic web: A survey," *Semantic Web*, vol. 11, pp. 1–81, 10 2018.
- [33] S. Busemann and H.-U. Krieger, "Resources and techniques for multilingual information extraction," 01 2004.
- [34] N. Viani, C. Larizza, V. Tibollo, C. Napolitano, S. Priori, R. Bellazzi, and L. Sacchi, "Information extraction from italian medical reports: An ontology-driven approach," *International Journal of Medical Informatics*, vol. 111, 12 2017.
- [35] S. N. Group, "Stanford open information extraction." <https://nlp.stanford.edu/software/openie.html>.
- [36] R. Guarasci, E. Damiano, A. Minutolo, and M. Esposito, "When lexicon-grammar meets open information extraction: a computational experiment for italian sentences," in *CLiC-it*, 2019.
- [37] C. Niklaus, M. Cetto, A. Freitas, and S. Handschuh, "A survey on open information extraction," *CoRR*, vol. abs/1806.05599, 2018. <http://arxiv.org/abs/1806.05599>.
- [38] V. Lombardo, G. Spione, and L. Provero. http://www.di.unito.it/~vincenzo/FTP_Historygraphia/Progetto_HSG_def.pdf.
- [39] A. Felicetti, D. Williams, I. Galluccio, D. Tudhope, and F. Niccolucci, "Nlp tools for knowledge extraction from italian archaeological free text," pp. 1–8, 10 2018.
- [40] J. Martinez-Rodriguez, I. Lopez-Arevalo, and A. Rios-Alvarado, "Openie-based approach for knowledge graph construction from text," *Expert Systems with Applications*, vol. 113, 07 2018.
- [41] Explosion. <https://spacy.io/usage/spacy-101#whats-spacy>.

BIBLIOGRAFIA

- [42] J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran, "Learning multilingual named entity recognition from Wikipedia," 10 2017.
- [43] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python natural language processing toolkit for many human languages," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020.
- [44] Wikipedia, "Wikidata:introduction," 2021. [Online; controllata il 1-marzo-2022].
- [45] A. Anita, A. Corazza, F. Isgrò, and S. Silvestri, "Unsupervised entity and relation extraction from clinical records in italian," *Computers in Biology and Medicine*, vol. 72, pp. 265–275, 01 2016.
- [46] NSchrading, "Introduction to spacy for nlp and machine learning." https://github.com/NSchrading/intro-spacy-nlp/blob/master/subject_object_extraction.py.
- [47] F. Ciravegna, "Adaptive information extraction from text by rule induction and generalisation," in *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'01, (San Francisco, CA, USA), p. 1251–1256, Morgan Kaufmann Publishers Inc., 2001.
- [48] F. Ciravegna, S. Chapman, A. Dingli, and Y. Wilks, "Learning to harvest information for the semantic web," pp. 312–326, 05 2004.
- [49] S. Chapman, A. Dingli, and F. Ciravegna, "Armadillo: harvesting information for the semantic web," p. 598, 01 2004.
- [50] K. Varathan and S. Geetha, "Information extraction -a text mining approach," vol. 2007, pp. 1111 – 1118, 01 2008.
- [51] B. G. Robertson, ""fawcett": A toolkit to begin an historical semantic web," 2009.

BIBLIOGRAFIA

- [52] F. Hogenboom, F. Frasincar, U. Kaymak, and J. FMG, "An overview of event extraction from text," *CEUR Workshop Proceedings*, vol. 779, 01 2011.
- [53] B. Haslhofer, W. Robitza, C. Lagoze, and F. Guimbretiere, "Semantic tagging on historical maps," 04 2013.
- [54] R. Anantharangachar, S. Ramani, and S. Rajagopalan, "Ontology guided information extraction from unstructured text," *International journal of Web & Semantic Technology*, vol. 4, 02 2013.
- [55] F. Boschetti, A. Cimino, F. dell'Orletta, G. E. Lebani, Lucia, Passaro, P. Picchi, G. Venturi, S. Montemagni, and A. Lenci, "Computational analysis of historical documents : An application to italian war bulletins in world war i and ii," 2014.
- [56] R. Shah and S. Jain, "Ontology-based information extraction: An overview and a study of different approaches," *International Journal of Computer Applications*, vol. 87, 01 2014.
- [57] W. Shen, J. Wang, and J. Han, "Entity linking with a knowledge base: Issues, techniques, and solutions," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 27, pp. 443–460, 02 2015.
- [58] M. P. Di Buono, "Information extraction for ontology population tasks. an application to the italian archaeological domain," *International Journal of Computer Science: Theories and Applications.*, vol. 3, pp. 40–50, 04 2015.
- [59] V. Mohan, "Ontology based information extraction -a survey," 10 2016.
- [60] G. Moretti, R. Sprugnoli, S. Menini, and S. Tonelli, "Alcide: Extracting and visualising content from large document collections to support humanities studies," *Knowledge-Based Systems*, vol. 111, 08 2016.
- [61] P. Basile, G. Semeraro, and A. Caputo, "Entity linking for the semantic annotation of italian tweets," *Italian Journal of Computational Linguistics*, vol. 2, pp. 87–99, 06 2016.

BIBLIOGRAFIA

- [62] D. Calvanese, P. Liuzzo, A. Mosca, J. Remesal, M. Rezk, and G. Rull, “Ontology-based data integration in epnet: Production and distribution of food during the roman empire,” *Engineering Applications of Artificial Intelligence*, vol. 51, 02 2016.
- [63] A. Vlachidis, A. Bikakis, D. Kyriaki-Manessi, I. Triantafyllou, and A. Antoniou, “The crosscult knowledge base: A co-inhabitant of cultural heritage ontology and vocabulary classification,” pp. 353–362, 09 2017.
- [64] P. Naderi, H. A. Rahmani, S. Azizi, and L. Safari, “A study of recent contributions on information extraction,” 05 2018.
- [65] A. Konys, “Towards knowledge handling in ontology-based information extraction systems,” *Procedia Computer Science*, vol. 126, pp. 2208–2218, 01 2018.
- [66] A. Goy, D. Magro, and A. Baldo, “A semantic web approach to enable a smart route to historical archives,” *Journal of Web Engineering*, vol. 18, pp. 287–318, 01 2019.
- [67] M. Kokla and E. Guilbert, “A review of geospatial semantic information modeling and elicitation approaches,” *ISPRS International Journal of Geo-Information*, vol. 9, p. 146, 03 2020.
- [68] A. Goy, D. Colla, D. Magro, C. Accornero, F. Loreto, and D. Radicioni, “Building semantic metadata for historical archives through an ontology-driven user interface,” *Journal on Computing and Cultural Heritage*, vol. 13, pp. 1–36, 08 2020.
- [69] M. Favaro, M. Biffi, and S. Montemagni, *Risorse linguistiche di varietà storiche di italiano: il progetto TrAVaSI*, pp. 178–186. 01 2020.
- [70] P. Cassotti, L. Siciliani, P. Basile, M. de Gemmis, and P. Lops, “Extracting Relations from Italian Wikipedia using Unsupervised Information Extraction,” in *Proceedings of the 11th Italian Information Retrieval Workshop 2021 (IIR 2021)*

BIBLIOGRAFIA

- (V. W. Anelli, T. Di Noia, N. Ferro, and F. Narducci, eds.), CEUR-WS, 2021.
<http://ceur-ws.org/Vol-2947/paper2.pdf>.
- [71] J. C. Harrison, "What is history & why study it?." <https://web.archive.org/web/20140201183734/http://www.siena.edu/pages/3289.asp> (2022/04/24).
- [72] TextRazor. <https://www.textrazor.com/>.
- [73] Explosion. <https://spacy.io/models/it>.