Raccolta accurata di fatti da testo in linguaggio naturale di Wikipedia

Ricreare Lector utilizzando DBpedia

M. Faretra, G. Marini, A. Martinelli

25 febbraio 2017

RIASSUNTO

Molti approcci sono stati tentati nell'ultimo periodo per estrarre informazione da Wikipedia sotto forma di fatti (entità, relazione, entità) per il popolamento di Knowledge Graphs, in particolare sfruttando le informazione contenute nelle sue *infoboxes*. Tuttavia queste strutture dati riportano solo una piccola parte delle informazioni contenute negli articoli. Infatti nel testo libero si concentra la maggior parte dei fatti estraibili, la rilevazione di essi però risulta più problematica, trovandosi all'interno di un testo in linguaggio naturale, sicuramente più variegato e meno strutturato degli infoboxes, questi fatti non sono facilmente riconoscibili e dunque estraibili per aumentare la base di conoscenza. In questo lavoro si cerca di quantificare il numero e la precisione di nuovi fatti estratti da questi articoli con il supporto di un KG già popolato. In particolare, la nostra valutazione è stata effettuata utilizzando DBpedia, un KG gratuito, che ci ha portato a... (discuss to be implemented, un paio di numeri sui risultati che poi verranno discussi meglio)

1 INTRODUZIONE

L'incremento dei Knowledge Graph in questi ultimi tempi è stata di particolare interesse scientifico e ha evidenziato la concreta possibilità di aumentare la base di conoscenza che queste strutture possono offrire.

Il progetto DBpedia fornisce informazioni e fatti in 125 linguaggi differenti. La più grande base di conoscenza è estratta dalla versione inglese e consiste in 1,3 miliardi di fatti che descrivono 6 milioni di entità.

La conoscenza estraibile dal testo libero potrebbe essere decisamente più consistente e aumentare di molto il Knowledge Graph. L'approccio utilizzato va a scalare sfruttando i fatti già contenuti nel KG stesso, è quindi dipendente anche da essi: DBpedia offre una varietà di grafi da poter utilizzare, per questo lavoro sono stati utilizzati due grafi, il primo realizzato a partire dalle informazioni estratte dagli infoboxes fidati, dunque molto pulite; mentre il secondo è stato realizzato a partire sempre dagli infoboxes ma accettando informazioni anche meno pulite rispetto alle prime. I due grafi sono stati utilizzati parallelamente in due processi distinti di estrazione dei fatti.

L'utilizzo di Wikipedia è ampiamente diffuso tra i vari KG esistenti data la grande affidabilità che ormai garantisce, tuttavia le infoboxes fino a pochi anni fa erano ancora molto poco diffuse e solo nell'ultimo decennio esse si trovano in più della metà degli articoli.

Se invece si va a considerare la conoscenza presente nel testo libero, si può immaginare che da esso (ovviamente presente in ogni articolo) si possa estrarre informazione non presente nel KG poiché esso riporta una serie di informazioni che probabilmente non sono presenti nell'infobox, ad esempio riguardanti entità che non sono il soggetto dell'articolo.

Il nostro approccio al problema considera pattern del tipo "[entità] frase [entità]", ad esempio: "Francesco Totti was born in Rome" mette in relazione le due entità "Francesco Totti" e "Rome" utilizzando la frase "was born in" che descrive, in questo caso, un'istanza della relazione "birthPlace".

In questo articolo descriviamo l'approccio di estrazione di conoscenza da testo libero di Wikipedia, e i risultati in termini di aumento dei fatti presenti in DBpedia.

Per l'estrazione abbiamo generato dal dump in input, un insieme di frasi in cui le entità erano già state riconosciute, triple candidate a rappresentare informazione valida, verificando successivamente se la frase della tripla candidata potesse o meno esprimere una relazione. (discuss sommaria spiegazione del processo e qualche numero di risultati)

Il resto dell'articolo è organizzato in vari capitoli: nel Capitolo 2 vengono analizzate in maggior dettaglio le risorse utilizzate; nel Capitolo 3 si presenta l'approccio utilizzato per estrarre nuovi fatti; nel Capitolo 4 vengono mostrati i risultati ottenuti; nel Capitolo 5 si trovano cenni a lavori correlati e su cui ci siamo basati (discuss se vogliamo scriverci qualche cagata su Lector da cui siamo partiti); infine, nel Capitolo 6 vengono presentate le conclusioni sul lavoro effettuato ed eventuali sviluppi futuri.

2 RISORSE

Wikipedia.

Wikipedia è un'enciclopedia online libera e collaborativa, che attualmente comprende circa 5.3 milioni di articoli nella sua versione inglese. Questa modalità di collaborazione garantisce una grande qualità e affidabilità sulle informazione ed anche una certa omogeneità nell'esprimere determinati concetti che va a facilitare l'estrazione di fatti utilizzando proprio questi pattern ricorrenti.

Ogni articolo in Wikipedia fa riferimento ad una entità principale, che può rappresentare una persona, un luogo, un oggetto, un fatto, ecc..., identificata con un *wiki ID*. Nel testo le informazioni sono codificate in linguaggio naturale, inoltre le entità secondarie eventualmente presenti in un determinato articolo sono rappresentate usando dei *wikilinks*, una sintassi specifica di Wikipedia per evidenziare un particolare concetto e offrire un link all'articolo che lo descrive.

Nel nostro particolare caso le frasi di Wikipedia su cui lavorare ci sono state fornite dal docente già processate dagli strumenti messi a disposizione dall' NLP Group di Stanford, in particolare dallo strumento di Named Entity Recognition, queste erano nella forma di frasi con le entità presenti già etichettate in una particolare formato, ad esempio un frammento di interesse di una frase può essere:

- entità soggetto: [[Barack_Obama|m.02mjmr]]
- frase: "met"
- entità oggetto: [[Donald_Trump|m.0cqt90]]

Come si può vedere le entità sono ben identificate e quindi di facile estrazione dalle frasi, inoltre possiedono anche un id che identifica l'entità all'interno di Freebase, non rilevante per questa trattazione.

DBpedia.

DBpedia è frutto di un lavoro collaborativo da parte di una moltitudine di utenti per estrarre informazioni strutturate da Wikipedia e renderle disponibili sul Web in maniera strutturata.

Ogni entità in DBpedia è identificata da un URI del tipo:

http://dbpedia.org/resource/Name

dove "Name" è preso dall'URL del relativo articolo di Wikipedia, che ha la forma:

http://en.wikipedia.org/wiki/Name.

In questo modo ogni risorsa è legata direttamente ad un articolo di Wikipedia. I dati sono suddivisi in dataset diversi e dati accessori, in particolare:

- dati accessori relativi ai tipi delle entità, ovvero una serie di coppie "URL entità - URL tipo", che associa ad esempio all'entità "Barack Obama" il tipo "Persona";
- dati accessori relativi agli schemi, contenenti informazioni riguardo dominio e codominio di una relazione, oltre a dati sui tipi e supertipi delle entità utilizzati da DBpedia che analizzeremo in seguito;
- due dataset rappresentanti ognuno un KG, il primo costruito a partire da fonti più affidabili del secondo. Il secondo è stato utilizzato per filtrare i fatti estratti a partire dal primo.

Lector.

Per estrarre nuovi fatti ed aumentare il grafo è stato seguito un procedimento simile a quello utilizzato originariamente dall'originale Lector, tuttavia si è scelto di variare leggermente alcune fasi del processo estrattivo, oltre che variare anche alcune scelte relativamente allo scoring delle frasi. Queste modifiche, laddove presenti saranno evidenziate nel seguito della trattazione.

3 APPROCCIO

Il dataset iniziale di frasi, come detto precedentemente, ci è stato fornito con le entità etichettate. Inizialmente si è provato ad applicare un approccio basato su euristiche per il riconoscimento di frasi di tipo lista, basato sia su il lavoro svolto dal gruppo di NLP di Stanford sia su un lavoro precedentemente svolto da alcuni colleghi. Questo approccio purtroppo si è rivelato non scalabile su una quantità di frasi come quella a disposizione.

Si è passati poi ad un approccio basato solamente su semplici euristiche, come ad esempio il controllo dei tipi delle entità candidate ad essere quelle nella lista, il controllo molto semplice richiedeva che tutte le entità nella lista fossero dello stesso tipo, questa è solo un'esempio delle euristiche applicate. Questo approccio si è rivelato efficiente ma non preciso, la quantità di informazione estratta risultava essere poco rilevante rispetto alla mole delle frasi, ulteriori analisi hanno rivelato inoltre che purtroppo questa informazione era quasi sempre non corretta (e.g. l'entità primaria legata alle entità nella lista non erano correlate in nessuna maniera).

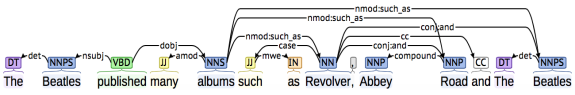

Figura 1: Esempio di una frase di tipo lista

Si è quindi preferito perdere una piccola parte dell'informazione presente nelle frasi favorendo la semplicità, data la grande mole di dati.

Il processamento delle frasi fino ad arrivare all'estrazione dei nuovi fatti si svolge in diversi passaggi:

1. Pulizia delle frasi originali di Wikipedia, con estrazione di triple entità-frase-entità;
2. Verifica della relazionalità delle frasi che legano due entità;
3. Etichettatura delle triple "presenti" e "non presenti" nel grafo di DBpedia, eseguito due volte, una per ogni grafo;
4. Scoring delle frasi ottenute dal passo precedente;
5. Estrazione dei nuovi fatti, eseguita per entrambi i grafi in modo indipendente.

3.1 Pulizia delle frasi originali

Il dump iniziale delle frasi ci è stato fornito dal docente con le entità delimitate da doppie parentesi quadre e arricchite con l'id relativo su *Freebase*, superfluo per questo lavoro. Il dato all'interno di esse ci fornisce il *wiki ID* che è stato utile per reperire l'entità sul grafo di DBpedia, che usa proprio questi identificatori per denotare i nodi. In alcuni casi questi identificatori fanno riferimento ad un *redirect* che punta all'effettivo *wiki ID* che abbiamo reperito tramite una mappatura offerta dal docente, successivamente abbiamo provveduto a sostituire i *redirect* con l'effettivo *wiki ID*.

Abbiamo quindi suddiviso le frasi iniziali in una lista di triple comprendenti tutte le coppie di entità e il testo compreso tra esse, le triple sono state prese in maniera sequenziale, ad esempio se in una frase comparivano 5 entità, le triple sono state estratte nel seguente modo: *ent1-frase1-ent2*, *ent2-frase2-ent3*, ecc.

Le frasi tra le entità sono state quindi passate ad un filtro per trovare quelle relazionali. Quelle che riescono a superare il test vengono usate per scremare le triple estratte al passo precedente, prendendo solamente quelle triple che contengono una delle frasi che ha passato il test relazionale. Questa operazione snellisce di molto la quantità di triple estratte garantendo delle triple che quasi sicuramente esprimono una relazione tra due entità.

3.2 Etichettatura

A questo punto, abbiamo considerato le coppie di entità di ogni tripla per verificarne l'effettiva presenza all'interno del grafo di DBpedia, sia in quello "fidato" che in quello "non fidato".

Le triple cosiddette "etichettate" rappresentano fatti già presenti nel KG e sono utili per ricavare una mappatura tra le frasi che legano le entità e le relazioni ad esse accomunabili.

Quelle "non etichettate" sono invece candidate a rappresentare quella parte di conoscenza in più che il nostro lavoro tenta di estrarre, ampliando l'informazione della base di conoscenza.

Le frasi hanno subito una elaborazione molto leggera per pulire le frasi e ricondurre quelle più particolari a forme più generali in modo da non perdere informazione (e.g.: *was born in* e *was Born in* sono state assimilate alla stessa frase)

3.3 Scoring

Ogni frase che lega due entità *"etichettate"* a questo punto è associata ad una o più relazioni definite da DBpedia, ma ovviamente queste corrispondenze potrebbero non rispettare l'effettiva semantica o delle frasi stesse, o della relazione a cui sono associate. Abbiamo quindi assegnato ad ogni frase un punteggio per ottenere le frasi che secondo lo score definito, possono esprimere in maniera corretta la relazione a cui sono associate, considerando principalmente il numero di volte che esse sono associate con la relazione in esame e il numero di relazioni in cui compare la frase. In questo modo abbiamo cercato di penalizzare frasi che esprimono concetti molto generici, ovvero che compaiono in molte relazioni, poiché non danno molta fiducia sulla correttezza dell'associazione.

Il punteggio viene assegnato tenendo conto del numero di volte che la frase è collegata alla relazione i-esima ($c(p, r_i)$), del conteggio totale delle sue occorrenze in tutte le relazioni analizzate e trovate nel processo di etichettatura ($\sum_{j \in R} c(p, r_j)$), unendo questi due fattori si ottiene la probabilità originale di Lector:

$$P(r_i|p) = \frac{c(p, r_i)}{\sum_{j \in R} c(p, r_j)}$$

Nel processo di definizione del nuovo score è stato inserito un nuovo fattore moltiplicativo che tiene conto del numero di relazioni a cui è associata la frase in esame ($c(R|p)$).

Il nuovo score risulta essere dunque:

$$score(p, r_i) = c(p, r_i) \cdot P(r_i|p) \cdot \frac{1}{c(R|p)}$$

In questo modo una frase otterrà un punteggio maggiore quanto più compare associata alla relazione in questione rispetto al numero totale di occorrenze, penalizzando inoltre le frasi in base al numero di relazioni a cui sono associate. Questo score penalizza le frasi associate con molte relazioni, tuttavia se la frase è associata con la relazione in esame un numero molto alto di volte allora il primo e il terzo fattore tendono a bilanciarsi, questo è il caso delle frasi che esprimono relazioni ontologiche (e.g. "was a", "is a", ecc...), queste sono associate con molte relazioni, tuttavia il loro "count" per alcune di queste è talmente alto che riescono a ottenere un punteggio molto alto all'interno della relazione in esame.

Un'altra differenza con lo score originale è la rimozione del logaritmo del "count" a favore di un fattore lineare, questo per enfatizzare il ruolo di quest'ultimo nel computo dello score finale. Inoltre non è impostata nessuna soglia sulla probabilità per l'ottenimento delle *top-K* frasi, mentre nell'originale Lector la soglia per permettere ad una frase di far parte delle *top-K* frasi di una relazione era 0.5

3.4 Estrazione dell'informazione

Una volta ottenuto lo score per ogni frase filtriamo per ogni relazione le 20 frasi (o tutte se ce ne sono meno) con punteggio più alto, queste vengono poi utilizzate per estrarre effettivamente i nuovi fatti.

A questo punto del procedimento si è reso necessario controllare i tipi delle entità dei fatti candidati a essere estratti. Nello schema fornitoci dal docente le entità note a DBpedia vengono associate ad un tipo, quello più specifico per l'entità (e.g. all'entità Barack_Obama è associato il tipo *OfficeHolder* il quale però è un sottotipo di *Person*, il quale è un sottotipo di Agent, e così via fino a raggiungere la radice dei tipi *Thing*), le entità dunque sono state arricchite con tutti i loro tipi, si è in pratica scorso l'albero dei tipi dal basso verso l'alto partendo dal tipo associato a ogni entità ottenendo per ogni entità tutti i tipi associati, radice *Thing* inclusa.

Oltre a tipi offerti dalle entità è stato necessario controllare e modificare anche i tipi richiesti dalle relazioni. Molte delle relazioni in DBpedia richiedono che il soggetto e l'oggetto rispettino particolari tipi (e.g la relazione *birthPlace* richiede che il soggetto sia un *Person* mentre l'oggetto sia un *Place*),

tipo di grafo utilizzato	*#fatti presenti*	*#nuovi fatti estratti*
grafo con fatti "trusted"	14,913,819	933,484
grafo con fatti "untrusted"	11,478,361	1,309,598

Tabella 1: Estrazione dei nuovi fatti in relazione ai fatti già esistenti

Relazione	*#"trusted"*	*#n. "trusted"*	*#n. "untrusted"*	#n. comuni	#eval.	precisione
birthPlace	1,131,887	56,554	107,030	56,126	100	87%
deathPlace [a]	303,537	35,392	22,154	17,260	100	16%
nationality	123,492	3,614	20	0	100	100%
team	778,837	9,977	11,520	917	100	89%
almaMater	119,573	77,944	65,113	63,246	100	98%
spouse	35,450	6,378	7,081	6,324	54	50%
parent [b]	28,308	6,397	8	8	100	63%
child [c]	14,503	856	0	0	127	57,4%
ethnicity	7,864	1,776	1,450	1,303	104	89,4%
religion	47,683	877	900	596	103	71,8%
award	86,279	3,087	0	0	101	94%
party	81,526	3,486	3,614	3,421	65	95,3%

[a] La relazione deathPlace soffre molto del filtraggio, i fatti ottenuti a partire dal grafo "untrusted" non risultano essere imprecisi, tutt'altro, ne risulta dunque una bassissima precisione dovuta al fatto che le triple valutate sono proprio un sottoinsieme sporco dei nuovi fatti "trusted", legati principalmente dalla frase *moved to*.

[b] Le relazioni *parent* e *child* hanno dei problemi legati all'omonimia tra gli antenati e i successori, gli strumenti di NER quasi mai riescono a discernere l'antenato dal successore se questi hanno lo stesso nome, in questi casi abbiamo un fatto in cui soggetto e oggetto puntano entrambi allo stesso *wiki ID* costringendoci a etichettare il fatto come falso.

[c] Vedi nota *b*.

Tabella 2: Dati relativi alle 12 relazioni di interesse in Lector originale, i fatti valutati sono stati presi eliminando dai nuovi fatti "trusted" i nuovi fatti comuni

per ogni relazione si è preso il *domain* della relazione ovvero il tipo richiesto al soggetto e il *range* ovvero il tipo richiesto all'oggetto e si sono associati alla relazione. Qualora uno dei due non fosse stato definito allora il tipo scelto per l'associazione è stato *Thing*, questa scelta è stata fatta in seguito ad una serie di test, nello schema fornitoci laddove non ci fosse stato il tipo del *domain* e/o *range* un'analisi sul portale di DBpedia ha rivelato che il tipo associato al campo mancante era sempre *Thing* ovvero la radice dei tipi.

Lo stesso discorso è stato applicato per le relazioni che non hanno né *domain* né *range* specificato, che sono: *religion, authority, builder, category, cpu, currency, division, format, gender, hasVariant, honours, influenced, influencedBy, isPartOf, jurisdiction, localAuthority, mainInterest, management, namedAfter, operatedBy, picture, position, predecessor, related, series, similar, source, sportGoverningBody, successor e webcast.* Per queste relazioni sia *domain* che *range* sono stati fissati a *Thing*, questa scelta poteva portare ad estrarre fatti errati poiché a questo punto non c'è nessuna limitazione sul tipo delle entità legate da una delle suddette relazioni, tuttavia un'analisi svolta sulla relazione *religion* ci dimostra che la precisione anche senza vincolo dei tipi richiesti risulta essere molto alta.

Da notare che grazie al passaggio precedente eseguito sulle entità è possibile che il test del tipo richiesto da una relazione passi, nel caso in cui l'entità su cui si esegue il test offra un sottotipo di quel-

lo richiesto (se è offerto *Scientist* e viene richiesto *Person* allora il test passa), ma fallisca nel caso in cui l'entità su cui si esegue il test offra solamente un supertipo di quello richiesto (se è offerto *Person* e viene richiesto *Scientist* il test fallisce), questo è ottenuto senza ulteriori modifiche a *domain* e/o *range* delle relazioni.

Alla luce di quanto detto precedentemente, nella fase di estrazione un nuovo fatto viene estratto se la frase che lega le due entità della tripla è presente nelle *top-K* frasi di una delle relazioni in esame e se i tipi dell'oggetto e del soggetto sono compatibili con quelli richiesti dalla relazione associata alla frase del passo precedente.

4 VALUTAZIONE

Come possiamo vedere dalla tabella in 3.4 il numero dei fatti nuovi estratti a partire dal grafo considerato meno affidabile è superiore a quello relativo al numero dei nuovi fatti estratti a partire dal grafo più affidabile, questo è dovuto ad una serie di fattori: alcune relazioni del grafo "fidato" sono assimilabili a più relazioni del grafo non fidato, dunque è possibile utilizzare un numero maggiore di *top-K* per estrarre ulteriori fatti; il numero di fatti non etichettati è maggiore nel primo caso rispetto al secondo, dunque è più probabile estrarre un maggior numero di fatti nuovi semplicemente perché si hanno più fatti da estrarre.

I 570,976 fatti comuni ai due insiemi di nuovi fatti estratti sono stati utilizzati per filtrare i nuovi fatti estratti a partire dal grafo fidato. Per la valutazione dunque ci siamo concentrati sulle dodici relazioni analizzate originariamente da Lector, ovvero: *birthPlace, deathPlace, nationality, team, almaMater, spouse, parent, child, ethnicity, religion, award, party.* Il numero di fatti filtrati è dunque 362,508 e su questi si è concentrata la nostra analisi.