Costruzione di una base di conoscenza Linked Data con tecniche di Machine Learning

---

Lorenzo Franco Ranucci

19 Aprile 2018
Versione: Versione 1

Università degli Studi di Perugia

Dipartimento di Matematica e Informatica

Tesi

Costruzione di una base di conoscenza Linked Data con tecniche di Machine Learning

Lorenzo Franco Ranucci

Relatrice Prof.ssa Valentina Poggioni
Dipartimento di Matematica e Informatica
Università degli Studi di Perugia

19 Aprile 2018

Lorenzo Franco Ranucci
Costruzione di una base di conoscenza Linked Data con tecniche di Machine Learning
Tesi, 19 Aprile 2018
Relatrice: Prof.ssa Valentina Poggioni

Università degli Studi di Perugia
Dipartimento di Matematica e Informatica
Via Luigi Vanvitelli, 1
06123 Perugia PG

Abstract

Indice

1 Introduzione

Linked Data

Come si producono i Linked Data

DBpedia

DBpedia Challenge

Struttura della tesi

Capitolo 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

2

Definizione del problema e lavori correlati

Costruzione di una base di conoscenza

Information Extraction

[**Palomares2016WikipediaKG**]

NLP

Slot Filling Task

Lavori Correlati

DBpedia

LectorPlus e Airpedia

DeepDive e Snorkel

3 Architettura della soluzione

Il lavoro di questa tesi si basa sull'intuizione che i risultati raggiunti nell'ambito della KBP mediante *Distant Supervision* 2.5.2 [**Mintz2009DistantSF**; **Exner2012EntityEF**; **Aprosio2013ExtendingTC**; **Soderland2013OpenIE**; **Cannaviccio2016AccurateFH**] e gli sviluppi raggiunti dalla nuova tecnica di *Data Programming* 2.5.3 [**Ratner2016DataPC**; **Ehrenberg2016DataPW**; **Bach2017LearningTS**; **WEB:weak_supervision**] possano essere combinati per ottenere risultati allo stato dell'arte nella costruzione o popolamento di basi di conoscenza. In particolare il fine è quello di dare una dimostrazione empirica di come una base di conoscenza Linked Data come DBpedia possa essere popolata con nuove triple, composte da predicati già definiti in un'ontologia, analizzando testi enciclopedici scritti in linguaggio naturale (Wikipedia) sfruttando l'insieme di triple già esistenti (distant supervision) ed euristiche specifiche (weak supervision).

Il paradigma di data programming consente l'utilizzo di tecniche di Machine Learning anche laddove non siano disponibili insiemi di training già etichettati a mano. Il modello di data programming è attualmente implementato nei progetti DeepDive e Snorkel. Si è scelto di utilizzare quest'ultimo principalmente perchè:

- consente la creazione di insiemi di training in maniera automatizzata utilizzando *funzioni di labelling* di vario genere, tra cui proprio la distant supervision
- l'insieme etichettato prodotto da queste funzioni viene poi raffinato da un algoritmo generativo e infine usato come input per allenare un modello discriminativo

La distant supervision necessita a sua volta di una base di conoscenza di supporto dalla quale reperire campioni di esempio. Nel progetto viene impiegata come fonte dei dati la stessa base di conoscenza che si vuole popolare: DBpedia. Come ogni base di dati Linked Data, il punto di accesso per effettuarne l'interrogazione è lo SPARQL Endpoint. Il linguaggio SPARQL viene impiegato in questo lavoro sia per il reperimento dei dati campione che per inferire metadati di interesse sui predicati Linked Data presi in esame.

In questo capitolo viene descritta la pipeline che, da un un corpus di testi di Wikipedia

e un insieme di predicati Linked Data (input), allena un modello di machine learning in grado di classificare le frasi del corpus ed estrarne nuove triple (output). La struttura è la seguente:

- analisi del corpus con tecniche di NLP che, per ogni frase rilevata, memorizzano features testuali (Sezioni 3.1 e 3.2)
- inferenza dei tipi di Named Entity (o Candidati) che esprimono il dominio e codominio di ogni predicato in input (Sezioni 3.3 e 3.4)
- labelling delle frasi candidate con distant suprevision e weak supervision (Sezioni 3.6 e 3.7)
- raffinamento dell'insieme etichettato e training del modello di classificazione (3.8)
- estrazione delle triple dalle frasi classificate positivamente (3.9)

Estrazione del *corpus*

Parsing NLP

Inferenza dei tipi di candidati

Estrazione dei candidati

Gold-labelling di valutazione

Interrogazione dei campioni della KB

Labelling Functions

Distant Supervision

DBpedia Lookup

Euristiche specifiche

Crowdsourcing

Training del modello

Estrazione delle triple

4 Test e Risultati

Test

Risultati

5 Sviluppi futuri

6 Conclusioni

Elenco delle figure

Elenco delle tabelle

Colophon

This thesis was typeset with LaTeX 2$_\varepsilon$. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at `http://cleanthesis.der-ric.de/`.

Declaration

You can put your declaration here, to declare that you have completed your work solely and only with the help of the references you mentioned.

Perugia PG, 19 Aprile 2018

Lorenzo Franco Ranucci