HomeWork 3
Ingegneria dei Dati

Progetto

Analisi approfondita del seguente articolo:

Schema-Agnostic Entity Matching using Pre-trained Language Models

Autori

Kai-Sheng Teong
School of Information Technology
Monash University Malaysia
kai.teong@monash.edu

Lay-Ki Soon
School of Information Technology
Monash University Malaysia
soon.layki@monash.edut

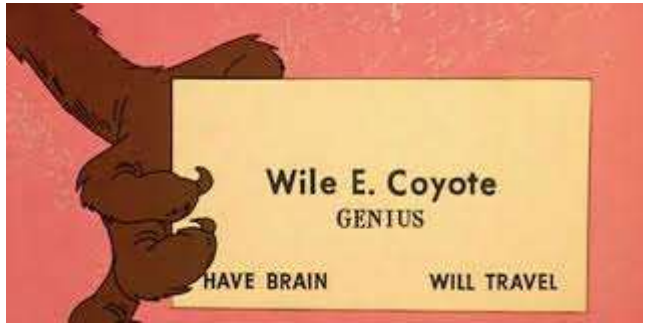
Tin Tin Su
Jeff ery Cheah School of Medicine and Health Sciences
Monash University Malaysia
intin.su@monash.edu

Team di sviluppo

ACME

Membri

Alessandro di Mario, Pietro Barone, Guido Petrolito, Douglas Ruffini.



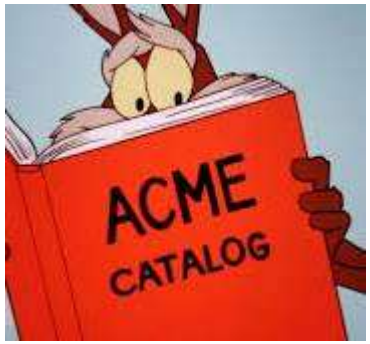
ROMA TRE
UNIVERSITÀ DEGLI STUDI

HomeWork 3
Ingegneria dei Dati

Ambito di ricerca: Integrazione dei dati. Fornisce un accesso unificato ai dati residenti su più fonti.

Scopo: Quello di creare un archivio unico di ricerca.

Prospettiva dell'integrazione virtuale: Supportare le query su uno schema mediato.



ROMA TRE
UNIVERSITÀ DEGLI STUDI

HomeWork 3
Ingegneria dei Dati

Presupposti sbagliati: Molti studi su EM (Corrispondenza tra Entità) presuppongono che per le specifiche dello schema, sia più conveniente abbinare coppie di record a livello di attributi. Sfortunatamente, le tabelle sottoposte a EM potrebbero non avere uno schema allineato, dove spesso, i metadati della tabella e degli attributi non sono congruenti.

ROMA TRE
UNIVERSITÀ DEGLI STUDI

HomeWork 3
Ingegneria dei Dati

Sfida: Sopperire il più possibile al processamento umano, il quale etichetta un piccolo set di coppie di record (match/non match) per insegnare a BERT come applicare la sua conoscenza generale al dominio specifico dell'integrazione dati.

Soluzione proposta: Utilizzare un modello linguistico pre-addestrato. Il pre-addestramento è l'apprendimento su miliardi di pagine web e libri (testo generico), prima di vedere i dati delle fonti. Serve a dare a BERT una comprensione enciclopedica del linguaggio.

L'approccio proposto mira a affrontare il problema dell'EM a schema non allineato trattando le coppie di record candidati in modo simile a un compito di classificazione di coppie di frasi in NLP.

Encoder: È la parte del Transformer responsabile di prendere una sequenza di input e trasformarla in una serie di rappresentazioni numeriche (vettori) che catturano il significato e le relazioni contestuali di ciascuna parola.

Meccanismo di Attenzione: La chiave del Transformer è l'**attenzione**

Agnosticismo: BERT non ha bisogno di sapere che le colonne si chiamano "Colore" o "Categoria"; analizza il significato delle stringhe di testo Nero e Dark (anche se sono in colonne diverse) per trovare la corrispondenza.

ROMA TRE
UNIVERSITÀ DEGLI STUDI

HomeWork 3
Ingegneria dei Dati

Metodo: BERT è stato pre-addestrato utilizzando un ampio corpus di testo non etichettato

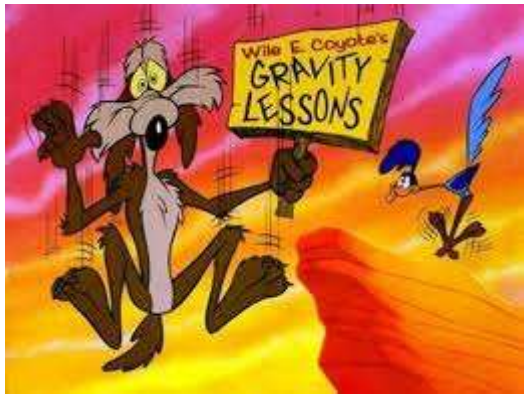
Fase 1: mascheramento casuale di una certa percentuale delle parole di input e addestramento del modello a predire quelle parole mascherate in base al contesto della frase con un processo chiamato Masked Language Modelling (MLM).

Fase 2: Next Sentence Prediction (NSP), addestra BERT a predire se la seconda frase viene effettivamente dopo la prima frase.

Presunzione sul metodo: Il lavoro di sperimentazione su dataset testuali e sporchi presuppone che questa struttura stratificata lo classifica come un modello di Deep Learning (una sottocategoria del Machine Learning).

Risultati: Questo lavoro, conferma i presupposti fornendo i risultati dell'utilizzo di modelli linguistici pre-addestrati su dati strutturati corrispondenti.

Conseguenze: Trattando ogni tupla come una frase, possiamo sfruttare le potenzialità di BERT come soluzione per l'EM indipendente dallo schema.



Confronto con altri metodi.

DeepMatcher e Magellan non sono modelli linguistici (PLM). Non hanno una fase di pre-addestramento. Si concentrano sull'addestramento supervisionato specifico per la Corrispondenza di Entità per il compito di classificazione (Match/No Match).

Magellan: si basa principalmente su metodi di Machine Learning Tradizionale (non Deep Learning) e sulle euristiche (Programmazione Tradizionale). Una volta estratte le feature numeriche (Euristica), queste vengono utilizzate per addestrare un modello di ML tradizionale.

DeepMatcher è stato uno dei primi framework a sostituire i classificatori ML tradizionali (DeepLearning) con Reti Neurali Profonde per il matching. DeepMatcher elimina la necessità per l'uomo di definire esplicitamente le feature. La rete neurale impara da sola le combinazioni di caratteri e parole che indicano un match.

Risultati:

Type	Dataset	Precision (P)	Recall (R)
Structured	Beer	73.68	100
	Amazon-Google	76.96	67.09
	Fodors-Zagats	100	100
	iTunes-Amazon	90.00	100
	DBLP-ACM	98.43	98.65
	DBLP-GoogleScholar	94.74	95.89
	Walmart-Amazon	84.86	81.35
Dirty	iTunes-Amazon	90.00	100
	DBLP-ACM	98.21	99.10
	DBLP-GoogleScholar	94.50	94.77
	Walmart-Amazon	80.35	72.02
Textual	Abt-Buy	91.49	83.50

Type	Dataset	Ours	DeepMatcher	Magellan
Structured	Beer	**84.8**	72.7	78.8
	Amazon-Google	**71.7**	69.3	49.1
	Fodors-Zagats	**100**	**100**	**100**
	iTunes-Amazon	**94.7**	88.5	91.2
	DBLP-ACM	**99.5**	98.4	98.4
	DBLP-GoogleScholar	**95.3**	94.7	92.3
	Walmart-Amazon	**83.1**	66.9	71.9
Dirty	iTunes-Amazon	**94.7**	79.4	46.8
	DBLP-ACM	**98.7**	98.1	91.9
	DBLP-GoogleScholar	**94.6**	93.8	82.5
	Walmart-Amazon	**83.4**	53.8	37.4
Textual	Abt-Buy	**87.3**	62.8	43.6

I risultati mostrano che la soluzione proposta supera costantemente i due metodi selezionati di almeno il 9% in termini di media (SOTA). Questi risultati sono stati condotti su Google Colaboratory: CPU Intel Xeon , RAM 12 GB, GPU Nvidia Tesla K80.

ROMA TRE
UNIVERSITÀ DEGLI STUDI

HomeWork 3
Ingegneria dei Dati

Grazie a tutti per l'attenzione.