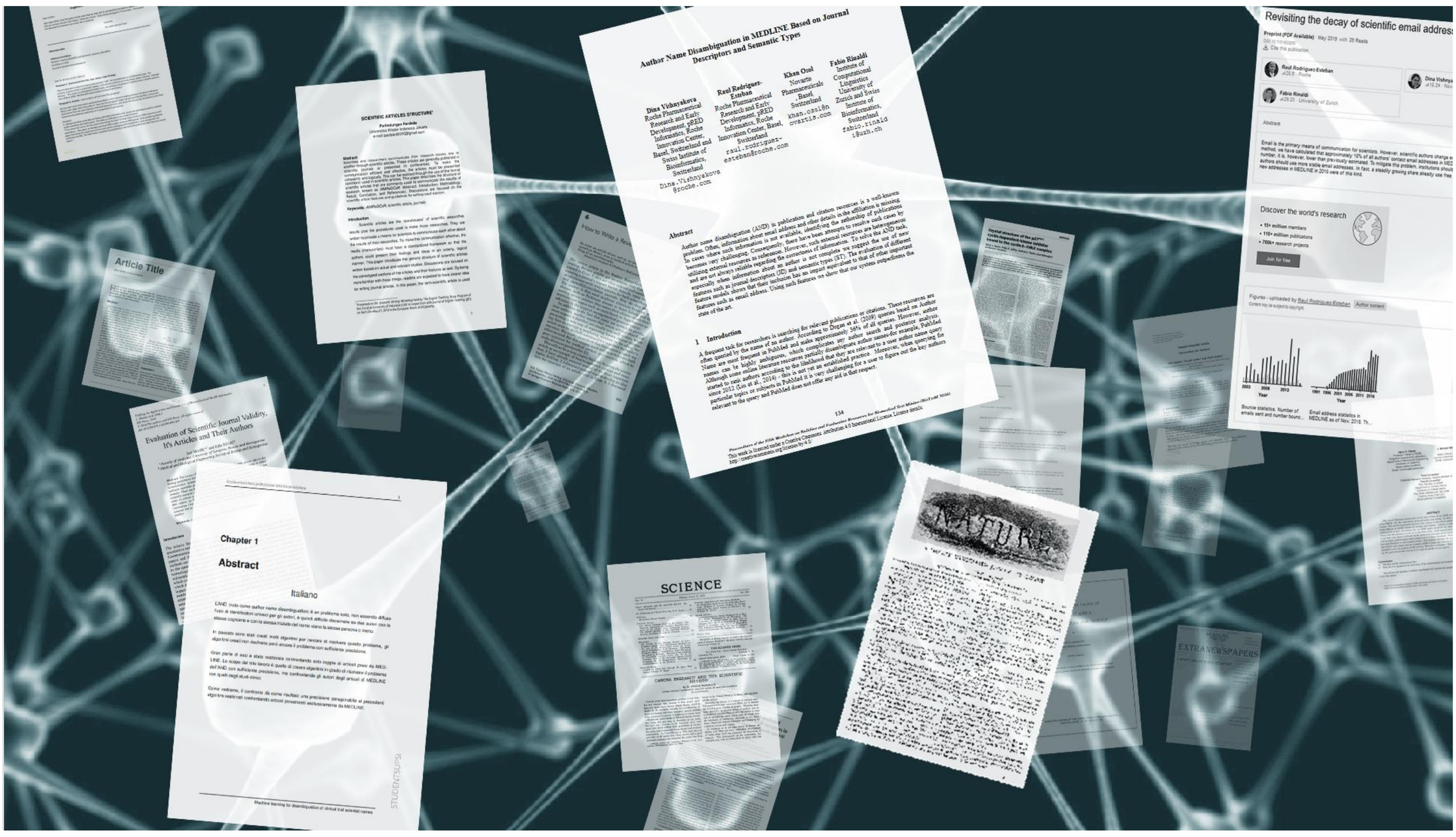
Scuola universitaria professionale della Svizzera italiana
Dipartimento tecnologie innovative

SUPSI

Machine learning for disambiguation of clinical trial scientist names

Studente/i	Relatore	Correlatore	Committente
Matteo Bresciani	Fabio Rinaldi	Luca Maria Gambardella	Hoffman-La Roche

Corso di laurea	N° Progetto	Anno	Data
Ingegneria Informatica	C10070	2018/2019	27/08/2019



Abstract

L'AND (author name disambiguation) è un problema noto, non essendo ancora diffuso l'uso di identificatori univoci per gli autori, è quindi difficile discernere se due autori con lo stesso cognome e con lo stessa iniziale del nome siano la stessa persona o meno.

In questo studio ci siamo concentrati sulla comparazione fra articoli scientifici e studi clinici.

In passato sono stati creati molti algoritmi per cercare di risolvere questo problema, non risolvendolo però con sufficiente precisione. Gran parte di essi è stata realizzata confrontando solo coppie di articoli presi da MEDLINE. Lo scopo del mio lavoro è quello di creare algoritmi in grado di risolvere il problema dell'AND con sufficiente precisione, ma confrontando gli autori degli articoli di MEDLINE con quelli degli studi clinici.

Come vedremo, il confronto dà come risultato una precisione paragonabile ai precedenti algoritmi.

Obiettivi

Il primo obiettivo è quello di utilizzare tecniche di **text mining** per estrarre informazioni utili dagli articoli scientifici e dagli studi clinici.

Il secondo è quello di utilizzare tecniche di **machine learning** in grado di riuscire a risolvere il problema con una precisione soddisfacente.

L'author name disambiguation è un problema conosciuto, ma davvero poche ricerche sono state fatte in passato per risolverlo in modo soddisfacente tra coppie di dati non omogenei tra loro.

Spesso le coppie di dati provengono dallo stesso database e sono strutturate nello stesso modo, in questo progetto, invece, si tratta di dati strutturati diversamente. Ci poniamo quindi l'obiettivo di risolvere l'AND in modo soddisfacente anche con questo ulteriore grado di difficoltà.

Conclusione

Nonostante la diversa struttura degli articoli scientifici rispetto ai rapporti delle prove cliniche, le tecniche di text mining e di text categorization hanno dato buoni risultati.

Rispetto alle precedenti ricerche, in questo progetto sono stati aggiunti nuovi attributi, i doc2vec ed è stata sfruttata la libreria di ontoGene.

Tali features hanno permesso di raggiungere, se non addirittura superare, risultati ottenuti da ricerche precedenti, nonostante il grado di difficoltà aggiuntivo.

Da questi risultati possiamo dedurre che l'AND è risolvibile con precisioni soddisfacenti, indipendentemente dalle fonti dei dati.

Le features che estrapolano informazioni dal testo degli articoli e dal testo delle prove cliniche (in particolare journal descriptors, semantic types e doc2vec) sono risultati i più validi.