Intelligenza Artificiale per l'Estrazione e la Gestione della Conoscenza da Documenti Testuali Biomedicali

Mario Ciampi, Angelo Esposito, Francesco Gargiulo, Mario Sicuranza, Stefano Silvestri e Giuseppe De Pietro

Istituto di Calcolo e Reti ad Alte Prestazioni del Consiglio Nazionale della Ricerche - ICAR-CNR
Via Pietro Castellino 111 - 80131, Napoli, Italia
{mario.ciampi, angelo.esposito, francesco.gargiulo, mario.sicuranza, stefano.silvestri, giuseppe.depietro}@icar.cnr.it

Abstract

Una rilevante area di ricerca dell'Istituto di Calcolo e Reti ad Alte Prestazioni del Consiglio Nazionale delle Ricerche (ICAR-CNR) è focalizzata sull'estrazione e gestione della conoscenza medica contenuta nella documentazione di dominio biomedicale. Gli approcci adottati riguardano in particolare la definizione e realizzazione di metodologie innovative basate su tecniche di Machine Learning e Big Data Analytics, applicate a grandi moli di documenti medici scritti in linguaggio naturale. Attualmente, le competenze maturate hanno permesso di instaurare numerose collaborazioni scientifiche con enti pubblici e privati e l'approvazione di molteplici progetti di ricerca industriale, nei quali trovano applicazione i risultati delle ricerche scientifiche enfatizzandone l'utilità e l'importanza.

1 Introduzione

Il significativo *know-how* nell'ambito dell'Intelligenza Artificiale (IA), nonché di quello della Big Data Analytics (BDA), posseduto dall'ICAR-CNR ha reso possibile la partecipazione dell'Istituto sia come proponente che come consulente in numerosi progetti di ricerca scientifica e industriale applicati al settore dell'e-Health.

L'attuale disponibilità di grandi moli di dati sanitari, infatti, richiede la definizione di metodologie ad-hoc capaci di sfruttarne il potenziale. Uno dei problemi da superare in tal senso riguarda l'estrazione di informazioni da documenti testuali semi-strutturati o non strutturati, contenenti parti narrative in linguaggio naturale. Tale task presenta attualmente delle difficoltà, causate della complessità del linguaggio naturale e dalla mancanza di sistemi specifici per il dominio biomedico e per la lingua italiana. Per migliorare le performance delle metodologie e tecniche attualmente disponibili in questo ambito, l'Istituto ha prodotto importanti risultati dal punto di vista scientifico e applicativo, supportando grazie ad essi lo sviluppo di progettualità di rilevanza nazionale e regionale.

Più nel dettaglio, le attività di ricerca svolte hanno riguardato la progettazione e l'implementazione di strumenti di estrazione di informazioni da testo in linguaggio naturale, specifici per documenti sanitari. Sono state definite tecniche innovative di classificazione del testo, di riconoscimento di entità di dominio biomedicale, di correlazione semantica. Tutti i risultati della ricerca hanno trovato applicazione nelle progettualità descritte nel seguito del documento.

2 Metodologie e Ambiti Applicativi

Di seguito sono descritte le metodologie scientifiche di IA sviluppate nell'ambito dello svolgimento dei progetti di ricerca e i rispettivi ambiti di applicazione. L'unità di ricerca dell'ICAR-CNR ha definito e realizzato tecniche basate su Machine Learning (ML) per la creazione di un motore di ricerca semantico di articoli medico/scientifici [Gargiulo *et al.*, 2017b; Gargiulo *et al.*, 2017a]. Nel dettaglio, tali strumenti sfruttano la proprietà di composizionalità semantica dei Word Embedding [Mikolov *et al.*, 2018] e un addestramento della rete neurale specifico su testo di dominio biomedicale. Tali tecniche hanno lo scopo di favorire la correlazione tra dati di Fascicolo Sanitario Elettronico (FSE) e la letteratura scientifica.

Inoltre, sono stati realizzati sistemi intelligenti ibridi basati su ML [Gargiulo *et al.*, 2018a] capaci di supportare la validazione automatica delle regole di conformità descritte in linguaggio naturale nelle *Implementation Guides* facenti parte dello standard CDA R2 di HL7 Italia per documenti di Referto di Medicina di Laboratorio, Profilo Sanitario Sintetico e Lettera di Dimissione Ospedaliera da far confluire nell'architettura per il FSE [Ciampi *et al.*, 2017],

Tecniche di ML basate su reti convoluzionali sono state applicate per l'implementazione di sistemi per la classificazione di tipo *eXtreme Multilabel Text Classification* di testo biomedico [Gargiulo *et al.*, 2018a]. In questo caso, la sfida affrontata è stata relativa ad un problema multi-classe, con una cardinalità molto ampia (dell'ordine delle decine di migliaia).

Nell'ambito del *Named Entity Recognition* (NER) nel dominio Biomedicale (vedi Figura 1), sono state implementate metodologie e modelli di word e char embedding [Lample *et al.*, 2016] per l'annotazione di un corpus in italiano finalizzato all'addestramento di sistemi di ML supervisionati basati su Reti Neurali Ricorrenti.

Per fornire analisi statistiche su grandi moli di dati eterogenei necessari per il conseguimento dei risultati previsti da specifici progetti di ricerca, sono stati studiati processi di Big Data Analytics basati su moderne tecnologie (quali Spark, SparkSQL e MongoDB) [Gargiulo *et al.*, 2018b].

Figura 1: Esempio di *Named Entity Recognition* di un documento medico.

3 Ricadute Progettuali

Le metodologie illustrate in precedenza sono state applicate in numerosi progetti di ricerca, di cui nel seguito si riportano i più significativi.

Nell'ambito dei progetti *"Realizzazione di servizi e strumenti a favore delle Pubbliche Amministrazioni per l'attuazione del Fascicolo Sanitario Elettronico"* e *"Ottimizzazione del workflow e dei processi del Fascicolo Sanitario Elettronico"*, oggetti di convenzioni operative tra Agenzia per l'Italia Digitale (AgID) e CNR, ICAR-CNR è impegnato nella definizione e progettazione dell'architettura del framework nazionale per l'interoperabilità tra i sistemi regionali di FSE, in conformità al D.L. 179/2012 e ss.mm.ii. Tali argomenti sono discussi nell'ambito di un tavolo nazionale al quale partecipano AgID, Ministero della Salute, Ministero dell'Economia e delle Finanze, Regioni e Province Autonome e CNR. Nonostante le indicazioni tecniche suggeriscano la generazione di documenti strutturati in formato standard (HL7 CDA R2), risulta fondamentale gestire il notevole pregresso costituito da documenti narrativi. Pertanto si rendono necessarie metodologie innovative per l'estrazione automatica di informazioni sanitarie da documentazione biomedicale.

Nel contesto del progetto *"BDA4PHR - Big Data Analytics for Personal Health Record"*, approvato a valere sul Bando MISE grandi progetti di R&S - PON I&C 2014-20, è stata definita e realizzata una piattaforma integrata, innovativa ed avanzata nei campi salute e benessere. Tale piattaforma si basa su servizi di *Big Data Analytics* (BDA) associati a meccanismi di raccolta, estrazione e gestione di dati sanitari su piattaforma *Cloud*.

Il progetto *"RARE.PLAT.NET - Innovazioni diagnostiche e terapeutiche per tumori neuroendocrini, endocrini e per il glioblastoma attraverso una piattaforma tecnologica integrata di competenze cliniche, genomiche, ICT, farmacologiche e farmaceutiche"*, approvato a valere sul Bando POR Campania FESR 2014-2020, coinvolge l'analisi dell'enorme quantità informativa sanitaria al fine di supportare il miglioramento della gestione clinica di pazienti affetti da neoplasie rare come i tumori aggressivi ipotalamo-ipofisari, i carcinomi corticosurrenalici, i tumori neuroendocrini ed i glioblastomi. Tale analisi richiede l'uso di tecniche di estrazione di concetti sanitari e relazioni da banche dati, nonché di tecniche di BDA.

Riferimenti bibliografici

[Ciampi *et al.*, 2017] Mario Ciampi, Mario Sicuranza, Angelo Esposito, Roberto Guarasci, e Giuseppe De Pietro. A technological framework for ehr interoperability: Experiences from italy. In Carsten Röcker, John O'Donoghue, Martina Ziefle, Markus Helfert, e William Molloy, editors, *Information and Communication Technologies for Ageing Well and e-Health*, pages 80–99, Cham, 2017. Springer International Publishing.

[Gargiulo *et al.*, 2017a] Francesco Gargiulo, Stefano Silvestri, e Mario Ciampi. A big data architecture for knowledge discovery in PubMed articles. In *2017 IEEE Symposium on Computers and Communications, ISCC 2017*, pages 82–87, Heraklion, Greece, 2017. IEEE.

[Gargiulo *et al.*, 2017b] Francesco Gargiulo, Stefano Silvestri, Mariarosaria Fontanella, Mario Ciampi, e Giuseppe De Pietro. A deep learning approach for scientific paper semantic ranking. In *International Conference on Intelligent Interactive Multimedia Systems and Services*, pages 471–481, Vilamoura, Portugal, 2017. Springer.

[Gargiulo *et al.*, 2018a] Francesco Gargiulo, Stefano Silvestri, e Mario Ciampi. A clustering based methodology to support the translation of medical specifications to software models. *Applied Soft Computing*, 71:199 – 212, 2018.

[Gargiulo *et al.*, 2018b] Francesco Gargiulo, Stefano Silvestri, Gennaro Oliva, e Mario Ciampi. Integration and performances of spark on a pbs-based hpc environment. Technical Report RT-ICAR-NA-2018-01, CNR - Istituto di Calcolo e Reti ad Alte Prestazioni - ICAR-CNR, 2018.

[Lample *et al.*, 2016] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, e Chris Dyer. Neural architectures for named entity recognition. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego California, USA, 2016. ACL.

[Mikolov *et al.*, 2018] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, e Armand Joulin. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018*, Miyazaki, Japan, 2018. ELRA.