Università degli Studi di Bari
"Aldo Moro"

Dipartimento di Informatica

Dottorato di Ricerca in Informatica e Matematica XXXI ciclo

Proposta di Progetto di Ricerca

Concept Understanding and Suggestion in Interactive Information Seeking

Dottorando

Dott. Gaetano Rossiello

Tutor

Prof. Giovanni Semeraro

1 Titolo della ricerca

Concept Understanding and Suggestion in Interactive Information Seeking

2 Area di ricerca

- Apprendimento Automatico e Data Mining
  - Text Mining
    - Natural Language Processing
    - Information Retrieval
    - Information Filtering

3 Obiettivi

La comprensione del linguaggio naturale rimane tuttora uno dei problemi aperti nel campo della Artificial Intelligence (AI). Il Natural Language Understanding (NLU), anche noto con il nome di *machine reading comprehension* o *text understanding*, è stato classificato come problema AI-Complete [40][49], ovvero uno di quei problemi che appaiono semplici per gli esseri umani, ma per i quali ad oggi non si conosce una soluzione (trattabile) in termini computazionali. Il text understanding consiste nel leggere un testo espresso in linguaggio naturale, determinarne il significato attribuendo un significato a termini, frasi e paragrafi in esso presenti ed effettuare inferenze su questi elementi al fine di elicitarne le proprietà esplicite o implicite [34]. In particolare, una delle problematiche più salienti nel modellare una rappresentazione testuale è quella di catturare le relazioni semantiche tra concetti. Per risolvere questo task, in letteratura sono state proposte molteplici metodologie, alcune delle quali accedono a basi di conoscenza esterne, altre invece costruiscono spazi distribuzionali semantici [16] analizzando il contenuto della raccolta di testi senza far uso di conoscenza pregressa.

Sulla base degli studi pregressi, l'obiettivo del progetto di ricerca è proporre nuovi modelli che, attraverso rappresentazioni a più livelli di astrazione (*deep*),

siano in grado di modellare differenti tipi di relazioni tra i concetti estratti da una raccolta di documenti. Inoltre questi modelli devono essere in grado di combinare ed integrare, in modo unificato, varie fonti di informazioni derivanti dalla conoscenza comune (*commonsense knowledge*), accedendo a basi di conoscenza ed ontologie, e da profili utente, analizzando contesti a breve e lungo termine (*short/long term context*). Lo scopo è proporre una piattaforma innovativa di supporto a varie attività correlate con l'accesso all'informazione, in grado di predire/anticipare i bisogni informativi degli utenti (*information need*) di un sistema per l'accesso all'informazione, suggerendo concetti in relazione al contesto in cui il sistema viene utilizzato.

4 Motivazioni

Il sovraccarico cognitivo (*information overload*) è un problema noto nell'odierna società digitale ed è causato dall'esplosione della mole di informazioni prodotta attraverso il World Wide Web. Per le informazioni di tipo testuale, il problema è ancor più significativo per via del maggior carico cognitivo richiesto per la lettura di un testo. Questo ha portato alla necessità di sviluppare sistemi sempre più intelligenti per adiuvare un utente nel processo di scelta (*decision-making*) delle informazioni rilevanti. Negli anni, per svolgere questo compito, la ricerca e l'industria di settore hanno proposto modelli e sistemi di Information Retrieval (IR) e Information Filtering (IF) sempre più efficaci ed accurati.

Sebbene i due paradigmi condividano lo stesso obiettivo, ovvero soddisfare il bisogno informativo di un utente, IR ed IF sono stati spesso oggetto di studio da parte di comunità distinte di ricercatori. Mentre in passato alcuni studi [3][15] hanno delineato formalmente i confini tra IR ed IF, sostenendo l'idea di una possibile convergenza tra i due paradigmi, recentemente, Ed. Chi, ricercatore presso Google, ha indicato [10] la direzione che sta intraprendendo l'industria di settore: integrare i sistemi di IR ed IF ripensando i modelli di interazione dell'utente e rivisitando i concetti stessi di rilevanza e serendipità. Il motivo di questa convergenza è da ricercasi nella sempre più rapida crescita di strumenti che assistono gli utenti anticipandone il bisogno informativo.

Meccanismi come Query Auto-Completion (QAC), Query Suggestion (QS), Entity Recommendation (ER), Hashtag Suggestion (HS), ma anche "assistenti personali intelligenti', basati su tecniche di apprendimento automatico, come Siri della Apple, Google Now, Microsoft Cortana, Nina della Nuance, e sistemi come IBM Watson si collocano a metà strada tra i paradigmi di ricerca e filtraggio dell'informazione. Il meccanismo *push*, tipico dei Recommender System (RS), viene attivato contestualmente al meccanismo *pull* del classico paradigma di ricerca.

Questo trend è confermato dal workshop SWIRL 2012 [1], dove ricercatori del settore hanno delineato le nuove frontiere e sfide della futura ricerca in ambito IR. Il minimo comune denominatore tra le varie proposte di ricerca è quello di investigare nuovi modelli e metodologie per predire e anticipare il bisogno informativo dell'utente. Inoltre, se si pensa anche alle nuove modalità con le quali un utente accede all'informazione, ad esempio dispositivi utilizzati in mobilità, ci si trova in uno scenario quasi paradossale: da un lato le informazioni prodotte crescono in modo esponenziale, dall'altro gli utenti hanno la possibilità di esplicitare sempre meno le proprie necessità.

Il presente progetto di ricerca si colloca in questo contesto. Verranno studiate e confrontate metodologie esistenti allo scopo di sviluppare meccanismi più accurati in grado di assistere l'utente in tutte le fasi del processo di ricerca dell'informazione, proponendo nuovi modelli in grado di identificare, comprendere per poi suggerire pezzi di informazione a grana fine (concetti) estratti da contenuti testuali.

5 Stato dell'arte

L'obiettivo prefissato nel presente progetto di ricerca comporta un ampio studio multidisciplinare e trasversale nell'ambito del Text Mining e più in generale del Machine Learning. In questa sezione verrà fornita una panoramica generale sulla letteratura recentemente prodotta nei vari ambiti di ricerca che coinvolgono (o che possono coinvolgere) l'intero processo di suggerimento di concetti, partendo dalle tecniche di identificazione ed estrazione di informa-

zioni da contenuti testuali fino alle metodologie utilizzate per modellarne e catturarne la semantica.

5.1 Information Extraction

L'Information Extraction (IE) è quell'area di ricerca appartenente al Natural Language Processing (NLP) che si occupa di sviluppare modelli capaci di estrarre informazioni strutturate da contenuti testuali non strutturati [13].

Un'adeguata rappresentazione dei dati testuali non strutturati è il punto di partenza per una corretta formulazione di ogni problema nell'area di ricerca del Text Mining [43][6]. Il modello Bag-of-Word (BoW) [39] è la rappresentazione più diffusa nell'ambito del Text Mining. Indipendentemente dai modelli (vettoriali o probabilistici) adottati nei vari ambiti, una rappresentazione *shallow* dei soli termini come insieme di *feature* non permette di cogliere adeguatamente il significato insito in una porzione di testo. In termini probabilistici, il motivo è da ricercarsi nell'assunzione di indipendenza condizionale tra i termini imposta dalle rappresentazioni basate sul modello BoW o sue estensioni.

Il processo di estrazione delle informazioni, nello specifico dei concetti, è il primo passo fondamentale per perseguire gli obiettivi del progetto *de quo*. Senza addentrarsi nelle varie definizioni in ambito filosofico, in questa sede per *concetto* si intende una porzione di informazione testuale, composto da uno o più simboli (caratteri o parole), che nel senso comune hanno un significato autonomo ed immutabile (o, più precisamente, che cambia lentamente nel tempo). Quindi, ad esempio, un concetto può essere indistintamente un termine, un sintagma nominale, una entità, un topic, una URL, oppure un (hash)tag.

Le tecniche a stato dell'arte di Named Entity Recognition (NER) adottano modelli chiamati Conditional Random Field (CRF) [12]. Le tecniche che permettono di estrarre sintagmi (*chunk*) dato un testo sono definite di *shallow parsing* o *chunking*. In letteratura sono stati proposti modelli supervisionati di Maximum Entropy (ME) [23] per questo scopo.

Recentemente, in ambito Semantic Web e Linked Open Data (LOD), con la rapida espansione e popolazione di basi di conoscenza strutturate (*knowledge base*) come DBpedia, YAGO e Google Knowledge Graph, nuove tecniche dette di Entity Linking [11] permettono di identificare e disambiguare [19] entità in un testo, associandone una URL che le identifica univocamente.

Per gli obiettivi di questo progetto, il vantaggio di rappresentare documenti testuali come Bag-of-Concept (BoC) [38] invece che come BoW è duplice. Avendo a disposizione una conoscenza pregressa più ricca, è possibile costruire modelli semantici più raffinati al fine di comprendere in modo più accurato il significato del testo. Inoltre, la suddivisione del testo in concetti auto-consistenti permette di identificare e quindi costruire un repository di oggetti atomici da utilizzare come suggerimenti all'utente nel processo di ricerca dell'informazione.

5.2 Distributional Semantics

Ludwig Wittgenstein in [48], suggerisce che "il significato delle parole è determinato dal loro utilizzo". Basandosi su questa intuizione, l'area di ricerca della Distributional Semantics (DS) [16][21][37] studia metodi e modelli per quantificare le similarità semantiche tra elementi linguistici analizzandone le proprietà distribuzionali all'interno di una raccolta di documenti. La semantica distribuzionale utilizza l'algebra lineare come strumento di rappresentazione in cui ogni elemento linguistico in una raccolta di dati è proiettato in uno spazio vettoriale. La similarità semantica tra due elementi linguistici viene calcolata attraverso misure di similarità, come quella del coseno, tra le rispettive rappresentazioni vettoriali nello spazio.

I metodi di costruzione degli spazi vettoriali e la loro riduzione di dimensionalità sono fattori che possono fortemente influenzare l'efficacia di questi modelli. In letteratura sono stati proposti diversi metodi per costruire spazi distribuzionali, come ad esempio il Latent Semantic Analysis (LSA) [26], che applica una decomposizione a valori singolari (SVD) sulla matrice di correlazione dei termini, con lo scopo di far emergere le informazioni latenti da una raccolta di testi. Un altro approccio chiamato Explicit Semantic Ana-

lysis (ESA) [14] costruisce uno spazio distribuzionale proiettando i termini nello spazio vettoriale costruito su una base di conoscenza esterna come, ad esempio, Wikipedia. Il metodo Random Indexing (RI) [36] invece, costruisce uno spazio vettoriale ridotto utilizzando la tecnica incrementale nota con il nome di *random projection* [45]. Il vantaggio di questa tecnica risiede proprio nella proprietà intrinseca di incrementalità rispetto alle altre tecniche. Infine, un modello proposto recentemente, chiamato Word2Vec [29][30], utilizza una rete neurale per costruire vettori di valori reali, detti *word embedding*, di dimensione ridotta rispetto a quella dell'intero vocabolario.

5.3 Query Suggestion

Varie definizioni di Query Suggestion (QS) si posso ritrovare in letteratura. In alcuni lavori QS è definito come sinonimo di Query Recommendation (QR), ossia quei meccanismi attivati dopo che l'utente ha già formulato interamente il proprio bisogno informativo e che suggeriscono query affini a quella iniziale. In altri lavori, QS è definito come tecnica di Query Auto-Completion (QAC), ovvero quei meccanismi che forniscono interattivamente una lista di possibili completamenti di query ad ogni carattere digitato nel box di ricerca.

In letteratura sono stati proposti due tipi di approcci differenti al QS, quello basato su *query log* e quello cosiddetto *corpus-based*.

L'approccio basato su query log, consiste nel suggerire o completare la query (parziale) analizzando insiemi di query precedentemente formulate dagli utenti. Questo approccio è assimilabile al concetto di "*wisdom of the crowd*" tipico dei metodi di Collaborative Filtering nei Recommender System. Una panoramica dei metodi esistenti è presente in [41]. Uno svantaggio di questo approccio è rappresentato dalla difficoltà di suggerire query rare presenti nella long tail, ossia query formulate da pochi utenti. Recentemente, sono stati proposti dei lavori che fanno uso di tecniche Deep Learning per superare questo limite. In [42] si adottano Recurrent Neural Network (RNN) per predire il cosiddetto *user intent* in base alle query precedentemente formulate. Al contrario, in [33] viene proposto un modello *deep*, denominato Convolutional

Latent Semantic Model (CLSM), per suggerire nuove query partendo da quelle presenti nei query log.

Gli approcci corpus-based proposti in letteratura sono molto più rari. Questi metodi generano automaticamente query analizzando il contenuto di una raccolta di testi e sono adatti in quei contesti dove non si ha a disposizione una base di query log sufficiente grande. Il primo tentativo è stato affrontato in [2], anche se la tecnica non suggerisce intere query, ma singoli termini. In [7] è proposto un modello probabilistico corpus-based con l'intento di completare la query parziale con *n-gram* estratti dal testo. Un framework che suggerisce *phrasal concept*, applicato al dominio della letteratura scientifica, è proposto in [22].

5.4 Entity Recommendation

Entity Recommendation (ER) è una nuova e promettente area di ricerca che studia modelli per suggerire entità collegate a basi di conoscenza. In un recente tutorial [27] tenuto da Hao Ma, ricercatore presso Microsoft, viene fornita una panoramica ben strutturata sulla letteratura prodotta in questo ambito. Vengono delineate varie applicazioni possibili di ER. La prima è intesa come metodo per calcolare la similarità tra due entità. In letteratura sono state proposte varie misure di *entity relatedness*, come, ad esempio, co-occorrenze di entità [47] in una base di conoscenza testuale come Wikipedia.

Il suggerimento di entità può essere formulato come un classico problema di *recommendation*, in cui gli oggetti da suggerire (item) in base ad un dato profilo utente corrispondono alle entità. Il framework proposto in [50] permette di suggerire entità analizzando i cosiddetto *click-through* (il comportamento dell'utente nell'interazione con un sistema) degli utenti e accedendo a basi di conoscenza come Freebase.

Infine, ER può essere considerato come problema di Question Answering (QA), in cui un ipotetico sistema restituisce una entità come risposta ad una interrogazione di un utente [44][28].

5.5 Deep Learning

Durante l'apertura del laboratorio di ricerca Facebook a Parigi, a Giugno 2015, il suo direttore Yann LeCun afferma: "*The next big step for Deep Learning is natural language understanding*". In un recente articolo pubblicato nel Gennaio 2016 sulla rivista Computational Linguistic [1], Christopher Manning parla di "*The Deep Learning Tsunami*", riferendosi alla grande quantità di pubblicazioni su Deep Learning in ambito Natural Language Processing prodotte nel solo anno 2015.

La letteratura recentissima dimostra come modelli di deep learning, opportunamente addestrati, siano in grado di produrre risultati strabilianti in compiti di comprensione del testo. Ad esempio, in [51] si adottano modelli di Convolutional Neural Network (CNN) in grado di comprendere il testo partendo dal livello dei singoli caratteri sino ad arrivare a comprendere concetti astratti. Modelli di Memory Network [46] sono in grado di implementare sistemi di Question Answering che rispondono a domande complesse. Mentre, in [18], modelli di Long Short-Term Memory (LSTM) che adottano tecniche di *data augmentation* in fase di training sono utilizzati per la lettura e comprensione automatica del testo.

Per le finalità di questo progetto, la vasta e recente letteratura sul Deep Learning verrà studiata approfonditamente, investigando i modelli e le architetture deep più adatte a risolvere il problema della comprensione del testo e del suggerimento di concetti.

6 Approccio al problema

Indipendentemente da quale sia il task o il dominio applicativo, un modello di suggerimento di concetti può essere definito in termini di probabilità condizionata $P(s|C)$, ovvero la probabilità di un suggerimento candidato s dato un contesto C. Prima di approntare una qualsiasi soluzione al problema, si renderà necessario definire formalmente s e C.

[1] http://www.mitpressjournals.org/doi/full/10.1162/COLI_a_00239#.VrjPd3UrLCI

Come anticipato in 5.1, un suggerimento candidato s è un concetto, ovvero una qualsiasi combinazione di uno o più simboli (caratteri o parole) che, in un contesto C, viene utilizzata per riferirsi ad un elemento che possiede un significato ben preciso ed autonomo. C è definito dal contesto riferito ad un utente impegnato in un processo di ritrovamento dell'informazione. Un contesto può essere costituito dal un insieme di informazioni, come ad esempio *keyword* già digitate nel box di ricerca, storia delle *query* precedentemente effettuate, pagine web visitate e/o preferenze dell'utente inferite dalle interazioni con i social media.

La fase successiva riguarderà lo sviluppo di una metodologia in grado inferire un modello dei concetti e delle loro relazioni, desumibili dalle varie fonti di informazione a disposizione. Un possibile approccio per calcolare le probabilità $P(s|C)$ è basato sull'utilizzo di modelli grafici probabilistici come i Factor Graph (FG) [25][4]. Il vantaggio nell'adottare un approccio basato su FG, risiede nella loro flessibilità nel modellare dipendenze condizionali tra variabili aleatorie. Infatti, la probabilità congiunta può essere fattorizzata attraverso un prodotto di funzioni. Questo permette di adottare un approccio *divide et impera* nella stima delle probabilità, soprattutto se le variabili aleatorie rappresentano dati provenienti da sorgenti di informazioni differenti. Con questo approccio i singoli fattori, che rappresentano le dipendenze tra una o più variabili aleatorie, possono essere calcolati utilizzando varie misure (ad esempio, misure di prossimità di concetti all'interno di una raccolta di testi, o misure di similarità semantica che adottano modelli distribuzionali semantici, oppure anche misure su grafi come Random Walk [9] o Page Rank [35], utilizzate per calcolare la *relatedness* tra entità). Recentemente, con il crescente interesse riguardo il *Deep Learning* (DL) [5][20], i gruppi di ricerca di maggior rilievo del settore stanno studiando la possibilità di integrare e combinare architetture deep con modelli grafici probabilistici. Alcuni risultati suggeriscono che l'integrazione può essere effettuata in due modi distinti. In primo luogo, è possibile considerare le architetture *deep* come fattori in un factor graph [32][31], oppure usare queste architetture per apprendere le strutture stesse di un modello grafico probabilistico.

La definizione della struttura della rappresentazione dei concetti e dei

fattori nell'ipergrafo, consentirà di disporre di un motore inferenziale probabilistico con il quale sarà possibile effettuare interrogazioni e stimare le probabilità dei concetti da suggerire.

A tale scopo, prima di definire qualsiasi nuovo approccio, si renderà necessario effettuare uno studio preliminare e approfondito per costruire delle basi solide nell'ambito del *Statistical Machine Learning* (SML) [17] e dei *Probabilistic Graphical Model* (PGM) [8][24]. Inoltre, sarà necessario studiare approfonditamente anche la vasta letteratura recentemente prodotta in ambito *Deep Learning*, soprattutto riguardo le applicazioni nell'area NLP.

7 Ricadute applicative

Un modello che sia in grado di comprendere e suggerire concetti da raccolte di testi può trovare applicazione in scenari sia scientifici sia industriali. Tale modello può essere utilizzato come meccanismo di Query Auto-Completion, con lo scopo di aiutare un utente ad esprimere il proprio bisogno informativo suggerendo una lista di auto-completamenti coerenti con la query parzialmente digitata. Il modello potrebbe anche essere utilizzato come tecnica di Query Recommendation, al fine di suggerire ad un utente un insieme di query semanticamente correlate a quella immessa in un motore di ricerca. In ambito social network, il modello potrebbe trovare la sua utilità come meccanismo per suggerire *hashtag* durante la scrittura di un *tweet*.

Un modello di suggerimento di concetti può trovare la sua utilità anche come piattaforma di supporto ad altri task. Ad esempio, può essere adottato come meccanismo di Query Expansion per migliorare le performance di un sistema di Information Retrieval, o ancora può essere adoperato come framework in un Recommender System cognitivo (Content-Based), in cui la similarità tra item può essere calcolata attraverso una rappresentazione più espressiva offerta dalla modellazione di differenti tipi di relazioni tra concetti.

Anche l'industria può trovare interesse nella ricerca condotta in questo progetto. Grandi imprese di settore, come Google, Yahoo! e Microsoft di recente hanno iniziato ad esplorare nuovi modelli per migliorare l'esperienza dell'utente, proponendo soluzioni come Entity Recommendation in cui entità

collegate a Knowledge Base sono suggerite in base ad un contesto di ricerca. Inoltre imprese come Apple, IBM e Nuance, stanno investendo molto sulla ricerca e lo sviluppo di sistemi di Conversational Question Answering come Siri, Watson e Nina.

Una piattaforma capace di analizzare efficacemente il contenuto testuale di una raccolta di documenti può trovare la sua utilità anche in piccole e medie imprese (PMI) ed enti pubblici, in cui si ha la necessità di adottare strumenti automatici per poter accedere in modo intelligente all'informazione. Sistemi come Document Clustering e Classification di bandi pubblici o di documentazione prodotta internamente, possono trovare giovamento da una metodologia in grado di estrarre e mettere in relazione i concetti estratti da documenti.

Risultano inoltre di particolare interesse anche i possibili risvolti che l'applicazione di dette soluzioni tecnologiche possono avere in ambito di ricerca di anteriorità da riferirsi sia ad analisi su IPR (Intellectual Property Rights) che su artefatti, stato dell'arte e soluzioni disponibili sul mercato in forma di letteratura con un impatto diretto ad esempio nell'uso dei strumenti di procurement innovativo come il precommercial procurement o il procurement of innovative solutions.

8 Fasi del progetto

Le attività principali previste in questo progetto sono *ricerca*, *sviluppo* e *valutazione*. Queste tre attività verranno portate avanti in parallelo nel corso dei tre anni e sarà necessario schedularle al meglio per poter ottimizzare il processo di sviluppo di modelli innovativi e la conseguente valutazione dei risultati.

Una possibile suddivisione delle attività nei tre anni può essere delineata come segue:

- Primo anno

  - Studio approfondito nell'area del Machine Learningm, in particolare negli ambiti del Deep Learning e dei Probabilistic Graphical Models.
  - Studio approfondito dei modelli di Distributional Semantics.
  - Ricerca e studio approfondito della letteratura su Query Suggestion ed Entity Recommendation.
  - Partecipazione a summer school internazionali, conferenze, workshop e doctoral consortium.
  - Sviluppo delle abilità critiche attraverso attività di peer review.

- Secondo anno
  - Definizione, sviluppo e valutazione dei primi prototipi per il calcolo della similarità semantica tra concetti.
  - Integrazione e valutazione dei prototipi proposti in sistemi di suggerimento di concetti.
  - Pubblicazioni dei risultati in riviste e conferenze nazionali ed internazionali.
  - Confronto e condivisione dei risultati con gruppi di ricerca affini.

- Terzo anno
  - Integrazione e valutazione del modello proposto in compiti di ricerca affini, quali Information Retrieval e Recommender System.
  - Integrazione e valutazione del modello per la risoluzione di problemi reali in ambito industriale.
  - Periodo di Internship presso gruppi di ricerca internazionali di rilievo affini alla ricerca condotta.
  - Stesura della tesi di dottorato.

9 Valutazione del progetto

I risultati ottenuti dai vari prototipi sviluppati durante l'intero progetto di ricerca saranno valutati attraverso le modalità più adatte all'ambito applicativo in cui i modelli saranno adottati. Verranno condotte sperimentazioni seguendo protocolli formali che assicurano la replicabilità degli esperimenti oltre a stabilire la valenza scientifica dei modelli proposti.

Per ogni tipo di problema affrontato, saranno identificati i cosiddetti *gold standard*, dataset utilizzati in letteratura, al fine di confrontare le metodologie proposte con i metodi presenti in letteratura, adottando le misure più appropriate al problema in oggetto. Qualora non siano presenti dataset per un determinato problema, saranno predisposti valutazioni *in vivo* attraverso progettazione di *user study* e AB test (*randomized controlled experiments*).

10 Referenti esterni

Il referente scientifico esterno verrà designato da InnovaPuglia S.p.A., ente che finanzia la borsa di dottorato.

Possibili referenti scientifici esterni saranno identificati nel corso dei tre anni di dottorato, durante la partecipazione a summer school, conferenze, workshop e doctoral consortium. In via preliminare, in base alla linea di ricerca finora predisposta, verranno elencati dei profili di rilievo che lavorano su argomenti di ricerca affini con i quali, ove possibile, verrà instaurata una collaborazione:

- **Edgar Meij**, ricercatore presso Yahoo! Labs a Londra, autore di pubblicazioni di rilievo in ambito Entity Linking e Ranking.
- **Francesco Ricci**, professore presso l'Università di Bolzano, autore di pubblicazioni di rilievo in ambito Recommender System.
- **Bhaskar Mitra**, Microsoft, Cambridge UK, si occupa di Web Search, Information Retrieval and Machine Learning,

- **Christopher Manning**, professore di linguistica ed informatica presso la Stanford University, autore di pubblicazioni recenti nell'ambito modelli Deep Learning applicati al NLP.
- **Hao Ma**, ricercatore presso Microsoft Research, autore di pubblicazioni in ambito Entity Recommendation.

Riferimenti bibliografici

[1] J. Allan, B. Croft, A. Moffat, and M. Sanderson. Frontiers, challenges, and opportunities for information retrieval: Report from swirl 2012 the second strategic workshop on information retrieval in lorne. *SIGIR Forum*, 46(1):2–32, May 2012.

[2] H. Bast and I. Weber. Type less, find more: Fast autocompletion search with a succinct index. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 364–371, New York, NY, USA, 2006. ACM.

[3] N. J. Belkin and W. B. Croft. Information filtering and information retrieval: Two sides of the same coin? *Commun. ACM*, 35(12):29–38, Dec. 1992.

[4] M. Bendersky and W. B. Croft. Modeling higher-order term dependencies in information retrieval using query hypergraphs. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 941–950, New York, NY, USA, 2012. ACM.

[5] Y. Bengio. Learning Deep Architectures for AI. *Found. Trends Mach. Learn.*, 2(1):1–127, Jan. 2009.

[6] M. W. Berry and M. Castellanos. *Survey of Text Mining II: Clustering, Classification, and Retrieval*. 1 edition.

[7] S. Bhatia, D. Majumdar, and P. Mitra. Query suggestions in the absence of query logs. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 795–804, New York, NY, USA, 2011. ACM.

[8] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[9] L. V. Bogachev. Random Walks in Random Environments. *ArXiv e-prints*, July 2007.

[10] E. H. Chi. Blurring of the boundary between interactive search and recommendation. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, pages 2–2, New York, NY, USA, 2015. ACM.

[11] H.-J. Dai, C.-Y. Wu, R. Tsai, and W. Hsu. From entity recognition to entity linking: a survey of advanced entity linking techniques. In *The 26th Annual Conference of the Japanese Society for Artificial Intelligence*, pages 1–10, 2012.

[12] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[13] D. Freitag. Machine learning for information extraction in informal domains. *Mach. Learn.*, 39(2-3):169–202, May 2000.

[14] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[15] H. Garcia-Molina, G. Koutrika, and A. Parameswaran. Information seeking: Convergence of search, recommendations, and advertising. *Commun. ACM*, 54(11):121–130, Nov. 2011.

[16] Z. S. Harris. *Papers on Syntax*, chapter Distributional Structure, pages 3–22. Springer Netherlands, Dordrecht, 1981.

[17] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[18] K. M. Hermann, T. Kociský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. *CoRR*, abs/1506.03340, 2015.

[19] J. Hoffart, S. Seufert, D. B. Nguyen, M. Theobald, and G. Weikum. Kore: Keyphrase overlap relatedness for entity disambiguation. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 545–554, New York, NY, USA, 2012. ACM.

[20] Y. B. Ian Goodfellow and A. Courville. Deep learning. Book in preparation for MIT Press, 2016.

[21] J. Karlgren and M. Sahlgren. From words to understanding, 2001.

[22] Y. Kim, J. Seo, W. B. Croft, and D. A. Smith. Automatic suggestion of phrasal-concept queries for literature search. *Inf. Process. Manage.*, 50(4):568–583, 2014.

[23] R. Koeling. Chunking with maximum entropy models. In *Proceedings of the 2Nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning - Volume 7*, ConLL '00, pages 139–

141, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[24] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

[25] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE TRANSACTIONS ON INFORMATION THEORY*, 47:498–519, 1998.

[26] T. K. Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284, 1998.

[27] H. Ma and Y. Ke. An introduction to entity recommendation and understanding. In *Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18-22, 2015 - Companion Volume*, pages 1521–1522, 2015.

[28] E. Meij, K. Balog, and D. Odijk. Entity linking and retrieval for semantic search. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 683–684, New York, NY, USA, 2014. ACM.

[29] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[30] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119, 2013.

[31] P. Mirowski, M. Ranzato, and Y. LeCun. Dynamic auto-encoders for semantic indexing. In *Proceedings of the NIPS 2010 Workshop on Deep Learning*, pages 1–9, 2010.

[32] P. W. Mirowski and Y. LeCun. Dynamic factor graphs for time series modeling. In W. L. Buntine, M. Grobelnik, D. Mladenic, and J. Shawe-Taylor, editors, *ECML/PKDD (2)*, volume 5782 of *Lecture Notes in Computer Science*, pages 128–143. Springer, 2009.

[33] B. Mitra and N. Craswell. Query auto-completion for rare prefixes. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1755–1758, New York, NY, USA,

2015. ACM.

[34] P. Norvig. Inference in text understanding. In *Proceedings of the Sixth National Conference on Artificial Intelligence - Volume 2*, AAAI'87, pages 561–565. AAAI Press, 1987.

[35] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172, Brisbane, Australia, 1998.

[36] M. Sahlgren. An introduction to random indexing. In *Methods and applications of semantic indexing workshop at the 7th international conference on terminology and knowledge engineering, TKE*, volume 5, 2005.

[37] M. Sahlgren. The distributional hypothesis. *Italian Journal of Linguistics*, 20(1):33–54, 2008.

[38] M. Sahlgren and R. Cöster. Using bag-of-concepts to improve the performance of support vector machines in text categorization. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[39] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, Nov. 1975.

[40] D. Shahaf and E. Amir. Towards a theory of ai completeness. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, pages 150–155. AAAI, 2007.

[41] F. Silvestri. Mining query logs: Turning search usage data into knowledge. *Found. Trends Inf. Retr.*, 4:1–174, Jan. 2010.

[42] A. Sordoni, Y. Bengio, H. Vahabi, C. Lioma, J. Grue Simonsen, and J.-Y. Nie. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 553–562, New York, NY, USA, 2015. ACM.

[43] K. L. Sumathy and M. Chidambaram. Article: Text mining: Concepts, applications, tools and issues. an overview. *International Journal of Computer Applications*, 80(4):29–32, October 2013. Full text available.

[44] H. Sun, H. Ma, W.-t. Yih, C.-T. Tsai, J. Liu, and M.-W. Chang. Open domain question answering via semantic enrichment. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 1045–1055, New York, NY, USA, 2015. ACM.

[45] S. S. Vempala. *The random projection method*, volume 65 of *DIMACS series in discrete mathematics and theoretical computer science*. Providence, R.I. American Mathematical Society, 2004. Appendice p.101-105.

[46] J. Weston, A. Bordes, S. Chopra, and T. Mikolov. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. *CoRR*, abs/1502.05698, 2015.

[47] I. Witten and D. Milne. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, pages 25–30, 2008.

[48] L. Wittgenstein and G. E. M. Anscombe. *Philosophical investigations*, volume 255. Blackwell Oxford, 1958.

[49] R. V. Yampolskiy. AI-Complete, AI-Hard, or AI-Easy: Classification of Problems in AI. In *The 23rd Midwest Artificial Intelligence and Cognitive Science Conference*, pages 21–22.

[50] X. Yu, H. Ma, B.-J. P. Hsu, and J. Han. On building entity recommender systems using user click log and freebase knowledge. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 263–272, New York, NY, USA, 2014. ACM.

[51] X. Zhang and Y. LeCun. Text understanding from scratch. *CoRR*, abs/1502.01710, 2015.