Un'analisi del mondo del lavoro e un modello predittivo per potenziali nuove occupazioni

Gabriella Pasi[1], Mirko Cesarini[2], Stefania Marrara[1], Fabio Mercorio[2], Marco Viviani[1], Mario Mezzanzanica[2], and Marco Pappagallo[1]

[1] DISCo, Università degli Studi di Milano-Bicocca,
Edificio U14, Viale Sarca 336, 20126 Milano, Italy
[2] DISMeQ, Università degli Studi di Milano-Bicocca,
Edificio U7, Via Bicocca degli Arcimboldi 8, 20126 Milano, Italy

Sommario Questo articolo presenta un approccio basato su Language Model per l'identificazione di potenziali nuove occupazioni in annunci di lavoro, ovvero professioni non codificate dalla tassonomia standard europea ISCO. Poter riconoscere tali nuove occupazioni permette di filtrare il flusso di annunci che verrà poi classificato dal sistema WoLMIS, sviluppato dall'Università degli Studi di Milano-Bicocca per l'Agenzia europea CEDEFOP, secondo la tassonomia ISCO. Ciò facilita il lavoro degli esperti europei incaricati di estendere ISCO aggiungendo nuove professioni. L'approccio è stato testato su un dataset di annunci in lingua inglese ottenendo risultati promettenti.

1 Introduzione

Negli ultimi decenni un numero crescente di aziende e di persone in cerca di lavoro si affida al Web per far entrare in contatto domanda e offerta. Se opportunamente recuperata e analizzata, l'enorme quantità di offerte di lavoro disponibile oggi sui portali online dedicati è in grado di fornire informazioni dettagliate e preziose circa le dinamiche e le tendenze del mercato del lavoro sul Web. Queste informazioni possono essere particolarmente utili a diverse categorie di operatori, pubblici e privati, che svolgono un ruolo nel mercato del lavoro europeo, offrendo analisi di trend, valutazioni delle politiche per il mercato del lavoro e supporto all'orientamento (Vocational ad Educational Training - VET).

In questo contesto si situa WoLMIS,[3] un sistema prototipale progettato e sviluppato nell'ambito di una gara d'appalto europea con l'obiettivo di raccogliere e classificare automaticamente offerte di lavoro pubblicate sul Web in cinque paesi europei: Italia, Regno Unito, Repubblica Ceca, Irlanda e Germania, nelle rispettive lingue. In particolare, WoLMIS classifica automaticamente gli annunci estratti dal Web sulla base della tassonomia standard per le occupazioni esistenti ISCO (*International Standard Classification of Occupation*), al fine di costruire una base di conoscenza relativa al mercato del lavoro europeo.

[3] Il progetto è frutto di una collaborazione internazionale tra cinque paesi europei sotto la guida del *Centro di Ricerca* CRISP, con la collaborazione del *Dipartimento di Statistica e Metodi Quantitativi* (DISMeQ) e del *Dipartimento di Informatica, Sistemistica e Comunicazione* (DISCo) dell'Università degli Studi di Milano-Bicocca.

Il lavoro presentato in questo articolo si inserisce nell'attività di ricerca legata al progetto WoLMIS, e ne rappresenta una possibile estensione. Il nucleo principale di WoLMIS è il sistema di classificazione degli annunci di lavoro, che adotta tecniche di apprendimento automatico (machine learning). Per sua natura, il classificatore basato su machine learning presenta dei limiti per quanto riguarda la classificazione di nuove occupazioni, vale a dire occupazioni per le quali non esiste un'adeguata classe (o codice) ISCO predefinita. Tale problema è di particolare interesse, dato che sempre più settori come il turismo, i servizi bancari, le assicurazioni, o la moda, si sono trasformati negli ultimi anni seguendo l'evoluzione digitale, creando nuove figure professionali che non trovano adeguata descrizione nella tassonomia ISCO esistente. Alcune professioni sono state create ex-novo, mentre altre sono il risultato della fusione di due o più occupazioni esistenti. Ciò richiede il possesso di competenze trasversali (ad esempio per la professione di *Data Scientist*).

L'obiettivo di questo lavoro, quindi, è quello di estendere il sistema di classificazione proposto da WoLMIS, definendo una metodologia basata su tecniche statistiche e di analisi testuale per l'identificazione di potenziali nuove occupazioni nel mercato del lavoro.

2 Un approccio basato su Language Model per il riconoscimento di nuove occupazioni

Alla base dell'idea proposta in questo lavoro c'è la considerazione che le offerte di lavoro sono, principalmente, documenti non strutturati. Di conseguenza, attraverso un'analisi del linguaggio in essi utilizzato è possibile non solo determinare le peculiarità di ogni singola professione esistente, ma anche comprendere se si tratta o meno di annunci relativi a professioni non codificate dalla tassonomia ISCO. Partendo da questa intuizione, l'approccio proposto si fonda su una modellazione del linguaggio utilizzato nel testo di un'offerta di lavoro mediante *Language Model* [2]. I Language Model sono un modello probabilistico, che offre una rappresentazione formale di un documento attraverso distribuzioni di probabilità di parole in un linguaggio.

Dopo una fase iniziale di pre-processing degli annunci di lavoro (rimozione di stopword e riconoscimento di sostantivi e aggettivi tramite Pos-Tagging [1]), è stato costruito un *modello basato su bigrammi* [3] per ogni professione esistente poiché, da una fase preliminare di analisi, si è notato che mediamente i nomi delle occupazioni, sia nuove che attuali, sono costituiti principalmente da due parole; l'idea si è rivelata corretta e l'approccio basato sui bigrammi ha ottenuto risultati migliori rispetto ad una soluzione di test sviluppata usando semplici unigrammi. Per quanto riguarda la metodologia di smoothing, è stato utilizzato l'algoritmo *add-k smoothing* [4] in quanto, a livello computazionale, si è dimostrata la soluzione più opportuna tra quelle disponibili in letteratura, presentando comunque buoni risultati in termini di efficacia.

La creazione di un Language Model per ciascuna delle professioni codificate dalla tassonomia ISCO, ha permesso il riconoscimento di quei documenti il cui testo non è generato da nessuno dei Language Model generati. Questi documenti rappresentano quindi delle potenziali nuovi occupazioni. L'approccio presentato è stato testato su annunci in lingua in inglese ma è estendibile alle altre lingue presenti in WoLMIS. Nella fase di creazione dei Language Model delle professioni già codificate ISCO è stato utilizzato un dataset fornito dal Dipartimento di Statistica e Metodi Quantitativi contenente una serie di annunci acquisiti dal Web (tipicamente costituiti da un identificatore, un titolo e una o più descrizioni). Il dataset è costituito da circa 36.000 annunci.

Le fasi principali dell'approccio proposto sono riassumibili come segue:

1. Filtraggio del dataset iniziale. Il dataset filtrato contiene esclusivamente documenti relativi ad annunci che descrivono professioni già codificate dalla tassonomia.
2. Ripartizione del dataset in un insieme di training (75%) e uno di test (25%).
3. Apprendimento supervisionato. Questa fase si suddivide in:
   (a) costruzione di un documento per ciascuna professione esistente, raggruppando gli annunci identificati dallo stesso codice ISCO all'interno del training set;
   (b) rappresentazione formale di ciascun documento creato per ogni codice ISCO tramite Language Model (modello basato su bigrammi);
   (c) classificazione dei singoli annunci del dataset di test, utilizzati come query, identificando per ciascun offerta lavorativa il Language Model che più probabilmente genera il testo di ogni singolo annuncio.

   Lo scopo di questa fase è quello di determinare i valori dei parametri relativi allo smoothing e determinare il comportamento dei Language Model creati in termini di accuratezza.
4. Calcolo degli intervalli, ad un livello di confidenza del 99%, entro i quali si è assunto che un annuncio sia generato dal dato Language Model, uno per ciascuna professione esistente.
5. Test eseguito su un insieme di annunci riguardanti sia occupazioni già esistenti nella tassonomia ISCO, sia occupazioni non ancora codificate in essa (ovvero potenziali nuove professioni).

L'approccio proposto basato su Language Model è in grado di discriminare le potenziali nuove professioni da quelle esistenti associando, ad ogni annuncio filtrato, una lista di occupazioni già codificate ISCO più simili come descrizione e competenze richieste. Tale approccio pone le basi per un ulteriore sviluppo di WoLMIS durante la fase di realizzazione del sistema di classificazione per tutti i 28 paesi dell'Unione Europea.

3 Risultati sperimentali

Per la valutazione sperimentale del lavoro proposto è stato sviluppato un prototipo in Java. Per esigenze di valutazione è stato utilizzato un ulteriore dataset fornito da WoLMIS, contenente annunci di lavoro riguardanti sia professioni già codificate ISCO che non. Questo ulteriore campione di test contiene al più cinque offerte di lavoro per ciascuna occupazione esistente (codificata ISCO), più un centinaio di annunci relativi a occupazioni non codificate.

Come visibile in Tabella 1, i risultati ottenuti mostrano una buona separazione tra annunci riguardanti potenziali nuove occupazioni (colonna "NEW"), indicate con l'etichetta *Unknown* (quinta riga della tabella), e quelli relativi ad occupazioni esistenti. La colonna "LM#1" indica quanti annunci hanno trovato una corretta attribuzione nel Language Model identificato dal tool come il più probabile. Si può notare che mediamente, su cinque annunci di professioni esistenti, al più uno non viene correttamente riconosciuto e pertanto identificato come "NEW". Al contrario, tra gli annunci relativi a professioni non codificate, ben 108 vengono correttamente riconosciuti come "NEW". La fase di sperimentazione ha mostrato le potenzialità del sistema sviluppato, grazie all'identificazione di offerte di lavoro come potenziali nuove professioni. L'uso dei limiti inferiori degli intervalli di confidenza come valore soglia, può essere una buona soluzione nel distinguere queste nuove tipologie di occupazioni in un insieme misto di offerte di lavoro.

Occupation_L4	LM#1	NEW
Data entry clerks	4	1
Dispensing opticians	4	1
Hotel receptionists	4	1
Sales and marketing managers	4	1
Unknown	9	**108**
Advertising and marketing professionals	5	0
Athletes and sports players	1	0
Cashiers and ticket clerks	5	0
Chemists	5	0

Tabella 1. Alcuni risultati sperimentali

4 Conclusioni

In questo articolo è stato proposto un approccio per il riconoscimento di occupazioni non codificate rispetto ad una tassonomia standard (ISCO) nell'ambito del *Web Labour Market*, e per questo definite come *nuove occupazioni*. L'approccio è basato sulla definizione di un classificatore che implementa un modello generativo, mediante l'utilizzo di Language Model. Tale approccio rappresenta una possibile evoluzione del sistema WoLMIS, una piattaforma di classificazione di annunci di lavoro reperiti sul Web basata su tecniche di apprendimento automatico.

L'approccio proposto è in grado di discriminare tra tipologie di occupazioni diverse con una minima percentuale di errore. Così, non soltanto è possibile dotare il sistema WoLMIS della funzionalità di filtro offerta dal suddetto approccio, incrementando la sua efficacia nella classificazione, ma viene offerta la possibilità di facilitare il compito di analisi e identificazione di potenziali nuove occupazioni con cui estendere ISCO da parte di esperti, fornendo loro un insieme di annunci da analizzare prefiltrato e pertanto considerevolmente meno numeroso rispetto all'intero dataset.

La fase di sperimentazione ha mostrato le potenzialità dell'approccio proposto, grazie all'identificazione di annunci relativi a potenziali nuove professioni (ad esempio, *Data Scientist, Mystery Shopper* e *Data Analyst*). Questo pone le basi per un ulteriore sviluppo del sistema durante la fase progettuale che prevede la realizzazione del sistema di classificazione su tutti i 28 paesi dell'Unione Europea.

Riferimenti bibliografici

1. David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.
2. Fei Song and W Bruce Croft. A general language model for information retrieval. In *Proceedings of the eighth international conference on Information and knowledge management*, pages 316–321. ACM, 1999.
3. Munirathnam Srikanth and Rohini Srihari. Biterm language models for document retrieval. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 425–426. ACM, 2002.
4. Jinsong Su, Deyi Xiong, Yang Liu, Xianpei Han, Hongyu Lin, Junfeng Yao, and Min Zhang. A context-aware topic model for statistical machine translation. In *ACL (1)*, pages 229–238, 2015.

An Analysis of the Job Market and a Predictive Model for Potential New Jobs

Gabriella Pasi[1], Mirko Cesarini[2], Stefania Marrara[1], Fabio Mercorio[2], Marco Viviani[1], Mario Mezzanzanica[2], and Marco Pappagallo[1]

[1] DISCo, Università degli Studi di Milano-Bicocca,
Edificio U14, Viale Sarca 336, 20126 Milano, Italy
[2] DISMeQ, Università degli Studi di Milano-Bicocca,
Edificio U7, Via Bicocca degli Arcimboldi 8, 20126 Milano, Italy

Abstract. In this paper we present an approach based on Language Model to identify potential new jobs in job posts and new job types which have not been encoded in the European ISCO standard taxonomy. Identifying these new jobs calls for filtering the posts and eventually identifying them with reference to the ISCO taxonomy through the WoLMIS system developed by the University of Milan Bicocca for the CEDEFOP European Agency. This helps the European experts in extending ISCO by adding new job types. The overall approach has been tested on a dataset for the English language with very promising results.