Intelligent Document Processing per l'Estrazione di Informazioni da Documenti Complessi

Francesco Visalli, Antonio Patrizio, Massimo Ruffolo
Altilia.ai, Piazza Vermicelli, c/o Technest, Università della Calabria, Rende (CS), 87036, Italia
francesco.visalli@altiliagroup.com, antonio.patrizio@altiliagroup.com,
massimo.ruffolo@altiliagroup.com

Abstract

I progressi nel campo dell'Intelligenza Artificiale stanno consentendo il raggiungimento di traguardi impensabili fino a qualche anno fa. In questo lavoro presenteremo la Altilia Intelligent Automation™, piattaforma innovativa di Intelligent Document Processing (IDP), che fa ampio uso di tecniche allo Stato dell'Arte nel mondo del deep learning. Affronteremo dunque alcune delle problematiche che ci si trova ad affrontare per l'estrazione di dati da documenti anche molto complessi, con il supporto di un caso d'uso reale: i report di sostenibilità. Presenteremo infine una possibile pipeline di IDP, per la risoluzione di alcuni dei problemi che tratteremo, immersa nel contesto del deep learning.

1 Introduzione

I progressi effettuati negli ultimi anni, nel campo dell'Intelligenza Artificiale (IA), hanno spalancato le porte ad una quantità di applicazioni inimmaginabili fino a qualche tempo fa. Basti pensare a tecnologie pervasive come, ad esempio, gli assistenti vocali [Ram *et al.*, 2018; Sigtia *et al.*, 2020] o alle sempre più concrete self-driving car [Shalev-Shwartz *et al.*, 2017].

Nell'ambito del deep learning, i recenti lavori basati su meccanismi di attenzione [Vaswani *et al.*, 2017], come ad esempio il Bidirectional Encoder Representations from Transformers (BERT) [Devlin *et al.*, 2018], hanno dato il via, nel mondo del Natural Language Processing/Understanding (NLP/NLU), ad un rivolgimento come quella che fu, per il settore della Computer Vision (CV), la cosiddetta "rivoluzione ImageNet[1]", iniziata nel 2012 con AlexNet [Krizhevsky *et al.*, 2012].

Tuttavia, molte sfide restano ancora aperte, soprattutto legate all'industrializzazione dei risultati di ricerca sopracitati, al fine di consentire una fruizione massiva di tali tecnologie specialmente da parte di utenti "non addetti ai lavori". In accordo con le stime di Statista[2], la quantità totale di dati creati, catturati, copiati e consumati globalmente, è stimata abbia raggiunto, nel 2021, i 74 zettabytes. Di questi, secondo Gartner, più dell'80-85%, a livello enterprise, è non strutturato[3]. I modelli di machine learning sono sempre più performanti e capaci di catturare l'informazione non strutturata presente all'interno di questi dati per trasformarla in contenuti strutturati e insights da esplorare. Ciononostante, meno del 2% di tali dati è processato oggigiorno da algoritmi di AI. Ne deriva che la strada che porta alla riduzione del lavoro manuale, all'automazione di processi e al risparmio di costi, è molto promettente ma ancora da percorrere.

2 Altilia Intelligent Automation™

Altilia è una deep-tech company, fondata nel 2010 come spin-off di ricerca del CNR. Prima PMI innovativa in Calabria, è formata oggi da un team di oltre 45 persone, impegnate nella creazione di Altilia Intelligent Automation™, una piattaforma no/low-code offerta in modalità SaaS, per Intelligent Document Processing (IDP), che automatizza processi aziendali document-intensive mediante algoritmi di Intelligenza Artificiale.

La tecnologia di Altilia democratizza l'adozione dell'AI per IDP su scala nelle organizzazioni moderne di qualsiasi dimensione e settore aiutandole a realizzare l'automazione intelligente di processi (Intelligent Process Automation). Essa, in particolare, aiuta le persone a utilizzare l'AI per insegnare ai robot software a leggere e comprendere documenti e testi per estrarre dati e conoscenze in modo rapido, sicuro e preciso per automatizzare in modo efficiente i processi operativi e decisionali.

Attraverso la tecnologia di Altilia anche gli utenti di business e gli analisti, privi di competenze profonde di machine learning, possono costruire, istruire e applicare i propri modelli di AI, orchestrandoli, con semplici azioni point-and-click, in flussi di lavoro robotici che eseguono singole attività o interi processi. Grazie ai più recenti metodi di AI, a brevetti di proprietà[4] e a intense attività di R&S interne, Altilia integra il potere dell'intelligenza artificiale con la RPA per automatizzare anche quelle attività che richiedono capacità di giudizio e di contestualizzazione tipicamente umane.

[1] https://www.image-net.org/

[2] https://www.statista.com/statistics/871513/worldwide-data-created/

[3] https://www.forbes.com/sites/forbestechcouncil/2017/06/05/the-big-unstructured-data-problem/

[4] https://patents.google.com/patent/US9582494B2/

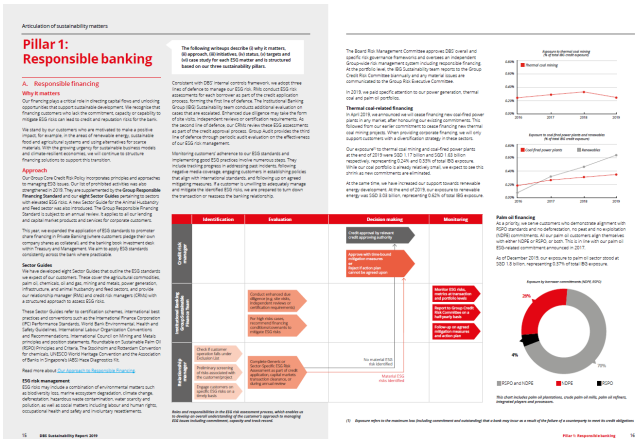

Figura 1: Esempio di documento ESG.

Altilia si rivolge ad aziende di medie e grandi dimensioni di ogni area industriale per soluzioni di Intelligent Document Processing, Intelligent Process Automation e Decision Intelligence. I clienti principali di Altilia ricadono tra i cosiddetti Fortune 500, e appartengono soprattutto al settore dei servizi finanziari.

3 Caso d'Uso: Report di Sostenibilità

L'Intelligent Document Processing consente, dunque, di sbloccare informazioni chiave racchiuse all'interno di documenti e di trasformarne il contenuto da non strutturato a strutturato. I documenti in oggetto possono essere di diversa natura: documenti di identità, documenti di reddito, buste paga, ordini di acquisto, bilanci, etc. Ne deriva, quindi, una difficoltà di estrazione che varia in base alla tipologia documentale. In questo lavoro presenteremo una panoramica delle maggiori problematiche, corredata da un insieme di possibili soluzioni, che ci si trova ad affrontare in questi contesti, prendendo in esame una delle classi documentali più complesse, oggetto di ricerca all'interno di Altilia: i report di sostenibilità.

I rapporti di sostenibilità, o ESG (da Environmental, Social e Governance), sono documenti che le grandi company producono annualmente e trattano tematiche ambientali, sociali, legate alla struttura del board e del managament. L'estrazione di informazioni da questi report, al fine di consentire una corretta profilazione ESG, sta diventando un tema sempre più caldo negli ultimi anni. Di recente, ad esempio, si sente molto parlare di investimenti "green", accessibili solo da società che dimostrino di rispettare determinate metriche legate all'ambiente.

Come anticipato, gli ESG rappresentano una delle classi più complesse di documenti, sia per forma che per contenuto. Quando parliamo di documenti complessi ci riferiamo a documenti che pesano spesso diverse decine di MB ed arrivano a superare il centinaio di pagine, documenti con layout variegati e difficili da interpretare, nei quali l'informazione di interesse può trovarsi in diversi elementi e risulta spesso disomogenea. In Figura 1) viene riportato un esempio di documento ESG, nello specifico si tratta di un documento multi colonna con diversi elementi di layout, sia testuali che grafici.

Per capire la complessità dell'informazione che si può voler estrarre, prendiamo in considerazione il dato riferito all'energia consumata e facciamoci guidare dalle tabelle per la spiegazione (Figura 2). Innanzitutto, bisogna stabilire la granularità dell'informazione: l'energia consumata può essere, infatti, riportata per tipologia di energia, per energia prodotta e poi consumata, per energia acquistata, per energia da fonti rinnovabili e non rinnovabili, etc. Più l'informazione diventa di grana fine, più diventa difficile da reperire e quindi da estrarre. Nello specifico, terremo in considerazione il totale dell'energia consumata, e dunque la grana più grossa possibile. In Figura 2 si può notare, evidenziato in magenta, il dato riferito all'energia consumata. Il valore numerico deve essere accompagnato dal tipo di unità di misura, che vediamo evidenziato, sempre in Figura 2, in verde. Da notare come, già in questo caso, l'informazione sia in posti diversi della tabella, e di come sia riportata una volta in MJ ed un'altra in GJ. Infine, la coppia *<energia consumata, unità di misura>*, deve essere associata all'informazione di colonna (o di riga, in base alla tipologia di tabella). Da notare, ancora una volta, come l'header delle due tabelle risulti disomogeneo in quanto, in un caso, il valore è riportato per anno, in un altro, per dipartimento. Se volessimo rendere omogenea l'informazione, e riportarla per anno (evidenziato in blu scuro in Figura 2), bisognerebbe fare la somma di riga della tabella (b). La medesima informazione può essere espressa in modi differenti. In Figura 3 e Figura 4, troviamo riportato il dato relativo al totale dell'energia consumata rispettivamente in testo e grafico. Discorsi analoghi a quelli affrontati fin'ora, valgono anche per queste tipologie di elementi di layout.

Energy consumption within the organisation (GRI 302-1, 302-2)	Unit	2017	2018	2019
From non-renewable sources	**GJ**	**73,857**	**86,947**	**87,871**
Methane gas	m3	1,777,114	2,292,236	2,334,974
Diesel fuel (for heating and generators)	l	133,819	10,445	9,818
Petrol (for the fleet)	l	22,068	15,579	16,882
Diesel (for the fleet)		77,448	95,768	94,621
From renewable sources	**GJ**	**385**	**348**	**339**
Photovoltaic (self-generated electricity)	kWh	106,840	96,805	94,283
Purchased electricity	**GJ**	**28,321**	**19,861**	**18,843**
from non-renewable sources	kWh	7,866,948	253,373	216,280
certified from renewable sources		-	5,263,489	5,017,969
Electricity sold				
Self-generated electricity sold to the grid	kWh	-	8,172	64,273
Total consumption	GJ	102,562	107,156	107,053
From non-renewable sources	GJ	102,178	87,859	88,649
From renewable sources		385	19,297	18,404

(a) Tabella che riporta il totale di energia consumata in base agli anni.

ENERGY CONSUMPTION AND TYPE AT CPDC PLANTS IN 2018

Item (MJ)	Toufen Plant	Dashe Plant	Hsiaokang Plant
Externally purchased electricity	6,321,388	203,486,934	533,586,608
Diesel	437,809	0	1,219,552
Gasoline	0	0	161,642
LPG	0	0	242,252
Natural Gas	857,986	753,199,266	40,734,431
Heavy Oil / Fuel Oil	25,292,252	130,483,724	23,726,292
Coal	6,503,467,277	0	0
Steam for internal use	7,352,339,389	737,483,088	892,642,488
Steam for external sale	0	97,618,234	0
Electricity for external sale	569,325,623	0	0
Total Energy Consumption	13,319,390,477	1,727,034,777	1,492,313,266

(b) Tabella che riporta il totale di energia consumata in base ai dipartimenti.

Figura 2: Esempi di tabelle relative alla quantità totale di energia consumata con informazione disomogenea.

4 Pipeline di Intelligent Document Processing

Per addentrarci nel trattamento delle problematiche finora esposte, cominciamo con l'elencare quali sono i possibili step di una pipeline di IDP:

- preprocessing;
- document layout analysis;
- indexing;
- information extraction;
- postprocessing.

La fase di preprocessing riguarda il trattamento preliminare dei documenti al fine di consentire una corretta estrazione delle informazioni. Una delle problematiche in cui ci si può imbattere è legata al fatto che i software di produzione dei documenti PDF non scrivano correttamente il PostScript ad essi associato. Ne risulta quindi che l'ordine di lettura visuale degli elementi è diverso da quello con cui sono stati codificati. Correggere l'ordine di lettura di tali PDF è fondamentale per consentire alle fasi successive di lavorare correttamente. Altri task che riguardano questa fase sono relativi, ad esempio, all'applicazione di algoritmi di OCR/ICR (Optical/Intelligent Character Recognition); infatti, non tutti i documenti sono nativamente digitali e corredati di testo. Tipicamente, questo genere di algoritmi, lavora in due fasi: prima vengono individuate le coordinate relative ai box contenti il testo (a diversa granularità, in base al modello: parola, carattere, etc.) all'interno del documento e successivamente viene riconosciuto [Visalli *et al.*, 2021]. Tali reti possono poi presentare ulteriori funzionalità, a corredo, come ad esempio il miglioramento delle immagini, la rimozione del rumore o meccanismi di allineamento del testo rispetto all'asse verticale.

La fase di document layout analysis è propedeutica alla scelta di applicazione degli algoritmi di estrazione. Testi e tabelle, ad esempio, potrebbero richiedere algoritmi di estrazioni differenti. In questo step si annoverano algoritmi di CV che sfruttano feature visive, tipicamente basate su modelli di object detection [Zhong *et al.*, 2019; Zheng *et al.*, 2020], algoritmi di NLP/NLU che sfruttano feature testuali [Hong *et al.*, 2021], e reti che combinano entrambe le feature, aggiungendone di ulteriori, come quelle spaziali [Xu *et al.*, 2019; Li *et al.*, 2019].

La fase di indexing risulta fondamentale quando ci si trova davanti a documenti aventi un gran numero di pagine, come gli ESG appena descritti. È uno step importante sia in termini di efficienza che di efficacia. Una buona indicizzazione seguita quindi da una fase di retrieving consente, infatti, di ridurre il numero di elementi su cui andare ad applicare i modelli di information extraction, concentrandosi solo sui passage più promettenti, riducendo così le tempistiche di esecuzione e la probabilità di generare falsi positivi. Le tecniche di indicizzazione si classificano in due categorie: quelle che utilizzano vettori sparsi, come i classici metodi basati ad esempio su TF/IDF e le sue varianti come BM25 [Robertson e Zaragoza, 2009], e le più recenti tecniche neurali che sfruttano vettori densi [Karpukhin *et al.*, 2020]. Il focus si sposta da similitudini a livello sintattico a similitudini a livello semantico.

In 2019 we concluded the sale of the mass market segment, as well as certain assets required for its operation. In this way we modified the parameter used to calculate energy intensity. In previous years it was related to the number of mass market subscribers; however, since 2018 we have taken the company's revenues as a reference.

In 2019, 48.28 GJ were consumed for every billion in revenue, considering $12.784 billion and a total energy consumption of 617,251 GJ. This calculation only includes consumption within Axtel.

Figura 3: Esempio di testo relativo alla quantità totale di energia consumata.

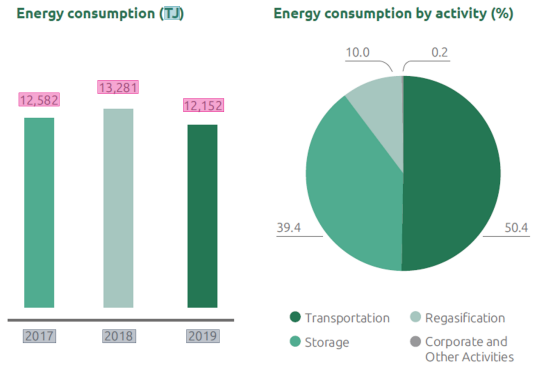

Figura 4: Esempio di grafico relativo alla quantità totale di energia consumata.

Una volta individuati i passage di interesse si passa dunque alla vera e propria fase di information extraction. Qui lo stato dell'arte è rappresentato dai language model [Devlin *et al.*, 2018; Liu *et al.*, 2019; Brown *et al.*, 2020] ai quali viene applicata una fase di "fine-tuning" in base al task specifico che si vuole risolvere: entity recognition, question answering, text classification, etc. Una particolare menzione va fatta per le graph neural network che di recente hanno trovato ampia applicazione nel campo del document understanding. Infatti, data la loro natura, i grafi riescono a rappresentare in maniera perfetta le relazioni che ci sono tra i vari elementi che appartengono ad alcune tipologie di classi documentali come le tabelle o i form [Lohani *et al.*, 2018; Qasim *et al.*, 2019].

Sono relative alla fase di postprocessing tutte quelle operazioni necessarie per trasformare il dato nel formato strutturato desiderato. Ad esempio, riferendoci ancora una volta agli esempi di Figura 2, una volta identificate separatamente le entità relative al totale dell'energia consumata per anno, l'unità di misura e gli anni, vanno collegate insieme costruendo le triple <*energia consumata, unità di misura, anno*>. Questa operazione può essere svolta tramite euristiche o attraverso modelli di machine learning.

5 Conclusioni

In questo lavoro è stata presentata la Altilia Intelligent Automation™, piattaforma innovativa di Intelligent Document Processing (IDP), che fa ampio uso di tecniche all'avanguardia di Intelligenza Artificiale, capace di leggere e comprendere documenti e testi per estrarre dati e conoscenze in modo rapido, sicuro e preciso, automatizzando in maniera efficiente

i processi operativi e decisionali. Sono state dunque introdotte le principali problematiche che ci si trova ad affrontare per automatizzare l'estrazione di dati, quando si lavora con documenti complessi, utilizzando come caso d'uso quello dei rapporti di sostenibilità. Infine, è stata presentata una possibile pipeline di IDP, calata nel contesto dello Stato dell'Arte del deep learning, per la risoluzione di alcune delle suddette problematiche.

Riferimenti bibliografici

[Brown *et al.*, 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, e Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, e Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[Hong *et al.*, 2021] Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, e Sungrae Park. BROS: A layout-aware pre-trained language model for understanding documents. *CoRR*, abs/2108.04539, 2021.

[Karpukhin *et al.*, 2020] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, e Wen-tau Yih. Dense passage retrieval for open-domain question answering. *CoRR*, abs/2004.04906, 2020.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, e Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012.

[Li *et al.*, 2019] Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, e Zhoujun Li. Tablebank: Table benchmark for image-based table detection and recognition. *CoRR*, abs/1903.01949, 2019.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, e Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[Lohani *et al.*, 2018] Devashish Lohani, Abdel Belaïd, e Yolande Belaïd. An invoice reading system using a graph convolutional network. In *ACCV Workshops*, volume 11367 of *Lecture Notes in Computer Science*, pages 144–158. Springer, 2018.

[Qasim *et al.*, 2019] Shah Rukh Qasim, Hassan Mahmood, e Faisal Shafait. Rethinking table recognition using graph neural networks. In *ICDAR*, pages 142–147. IEEE, 2019.

[Ram *et al.*, 2018] Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, e Art Pettigrue. Conversational AI: the science behind the alexa prize. *CoRR*, abs/1801.03604, 2018.

[Robertson e Zaragoza, 2009] Stephen E. Robertson e Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, 2009.

[Shalev-Shwartz *et al.*, 2017] Shai Shalev-Shwartz, Shaked Shammah, e Amnon Shashua. On a formal model of safe and scalable self-driving cars. *CoRR*, abs/1708.06374, 2017.

[Sigtia *et al.*, 2020] Siddharth Sigtia, Pascal Clark, Rob Haynes, Hywel Richards, e John Bridle. Multi-task learning for voice trigger detection. In *ICASSP*, pages 7449–7453. IEEE, 2020.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, e Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[Visalli *et al.*, 2021] Francesco Visalli, Antonio Patrizio, e Massimo Ruffolo. A two step fine-tuning approach for text recognition on identity documents. In *ICAART (2)*, pages 837–844. SCITEPRESS, 2021.

[Xu *et al.*, 2019] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, e Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. *CoRR*, abs/1912.13318, 2019.

[Zheng *et al.*, 2020] Xinyi Zheng, Douglas Burdick, Lucian Popa, e Nancy Xin Ru Wang. Global table extractor (GTE): A framework for joint table identification and cell structure recognition using visual context. *CoRR*, abs/2005.00589, 2020.

[Zhong *et al.*, 2019] Xu Zhong, Jianbin Tang, e Antonio Jimeno-Yepes. Publaynet: largest dataset ever for document layout analysis. *CoRR*, abs/1908.07836, 2019.