10.12871/CLICIT2014116

Errori di OCR e riconoscimento di entità nell'Archivio Storico de La Stampa

Andrea Bolioli
CELI Torino
abolioli@celi.it

Eleonora Marchioni
CELI Torino
marchioni@celi.it

Raffaella Ventaglio
CELI Torino
ventaglio@celi.it

Abstract

Italiano. In questo articolo presentiamo il progetto di riconoscimento delle menzioni di entità effettuato per l'Archivio Storico de La Stampa e una breve analisi degli errori di OCR incontrati nei documenti. L'annotazione automatica è stata effettuata su circa 5 milioni di articoli, nelle edizioni dal 1910 al 2005.

English. *In this paper we present the project of named entity recognition (NER) carried out on the documents of the historical archive of La Stampa and we show a short analysis of the OCR errors we had to deal with. We automatically annotated the authors of the articles, mentions of persons, geographical entities and organizations in approximately 5 million newspaper articles ranging from 1910 to 2005.*

1 Introduzione

In questo articolo descriveremo sinteticamente il progetto di annotazione automatica di menzioni di entità effettuato sui documenti dell'Archivio Storico de La Stampa, cioè il riconoscimento automatico delle menzioni di persone, entitá geografiche ed organizzazioni (le "named entities") effettuato su circa 5 milioni di articoli del quotidiano, seguito al progetto più ampio di digitalizzazione dell'Archivio Storico. [1]

Anche se il progetto risale ad alcuni anni fa (2011), pensiamo che possa essere d'interesse in quanto molti dei problemi affrontati e alcune delle metodologie utilizzate sono ancora attuali, a causa della maggiore disponibilità di vasti archivi storici di testi in formati digitali con errori di OCR. Si è trattato del primo progetto di digitalizzazione dell'intero archivio storico di un quotidiano italiano, e uno dei primi progetti internazionali di annotazione automatica di un intero archivio. Nel 2008 il New York Times aveva rilasciato un corpus annotato contenente circa 1,8 milioni di articoli dal 1987 al 2007 (New York Times Annotated Corpus, 2008), in cui erano state annotate manualmente persone, organizzazioni, luoghi e altre informazioni rilevanti utilizzando vocabolari controllati.

L'Archivio Storico de La Stampa comprende complessivamente 1.761.000 pagine digitalizzate, per un totale di oltre 12 milioni di articoli, di diverse pubblicazioni (La Stampa, Stampa Sera, Tuttolibri, Tuttoscienze, ecc.), dal 1867 al 2005. Il riconoscimento automatico di entità si è limitato agli articoli della testata La Stampa successivi al 1910, identificati come tali dalla presenza di un titolo, cioè a circa 4.800.000 documenti.

L'annotazione delle menzioni negli articoli consente di effettuare analisi sulla co-occorrenza tra entità e altri dati linguistici, sui loro andamenti temporali, e la generazione di infografiche, che non possiamo approfondire in questo articolo. Nella figura 1 mostriamo solamente come esempio il grafico delle persone più citate negli articoli del giornale nel corso dei decenni.

Nel resto dell'articolo presentiamo brevemente una analisi degli errori di OCR presenti nelle trascrizioni, prima di descrivere le procedure adottate per il riconoscimento automatico delle menzioni e i risultati ottenuti.

2 Analisi degli errori di OCR

Le tecniche di OCR (Optical Character Recognition) per il riconoscimento e la trascrizione auto-

[1] Come si legge nel sito web dell'Archivio Storico (www.archiviolastampa.it), "Il progetto di digitalizzazione dell'Archivio Storico La Stampa è stato realizzato dal Comitato per la Biblioteca dell'Informazione Giornalistica (CBDIG) promosso dalla Regione Piemonte, la Compagnia di San Paolo, la Fondazione CRT e l'editrice La Stampa, con l'obiettivo di creare una banca dati online destinata alla consultazione pubblica e accessibile gratuitamente."

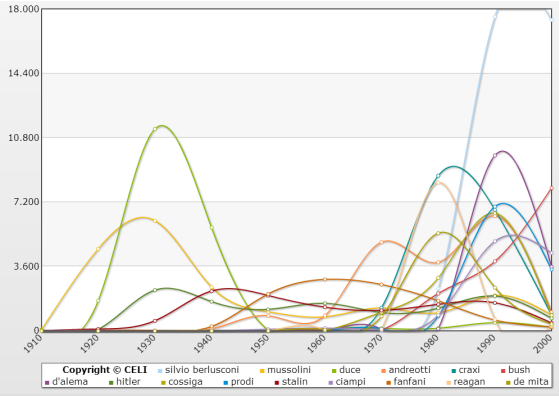

Figure 1: Grafico delle persone più citate

matica del testo comportano di per sé un margine di errore ”fisiologico”. Quando il documento cartaceo è di ottima qualità (buona carta e ottima qualità di stampa), le tecniche di OCR più moderne possono arrivare a percentuali di accuratezza molto alte. Nel caso di un archivio storico, dove la qualità dei documenti originali è ben lungi dall'essere ottimale, la quantità di errori di OCR aumenta notevolmente, soprattutto per i documenti più vecchi o in peggiore stato di conservazione, come attestato ad es. in (Holley, 2009) e nelle pubblicazioni del progetto europeo (IMPACT, 2010)).

Gli errori di OCR incidono sulle possibilità e sulla qualità delle elaborazioni successive sul testo. Una ricerca ”full-text” sui testi degli articoli non sarà ad esempio in grado di trovare le parole che contengono errori, oppure il testo sarà talvolta travisato, quando l'errore dà origine a una parola esistente ma diversa da quella originale. Allo stesso modo, il riconoscimento automatico delle menzioni di entità dovrà confrontarsi con questa situazione problematica.

2.1 Annotazione manuale e tipi di errore di OCR

Una misura affidabile dell'accuratezza dell'OCR si basa sul confronto tra il risultato dell' OCR e la trascrizione manuale corretta. Abbiamo quindi svolto un lavoro di annotazione manuale durante il quale dei linguisti hanno esaminato un campione dell'archivio (”corpus di valutazione”), individuando e classificando le anomalie di riconoscimento del testo.

Forniamo qui una sintesi della tipologia degli errori:

- Segmentazione degli articoli (”segmentation”): l'errata interpretazione degli elementi grafici della pagina (linee, titoli, cambiamento del corpo del carattere) può portare ad una errata segmentazione degli articoli, con diversi effetti possibili: un articolo risulta diviso in più parti, oppure diversi articoli vengono interpretati come uno solo, oppure un articolo risulta composto da porzioni di testo provenienti in realtà da articoli diversi.

- Segmentazione delle parole (”wordSep”): errori nell'interpretazione delle spaziature tra caratteri, parole o righe, che danno origine ad errori di segmentazione, ad esempio ”parlamenta re” (parlamentare), ”doc um e nto” (documento)

- Sillabazione (”hyphenation”): parole che vanno a capo nel testo originale (normalmente con trattino) vengono interpretate come parole separate, o come un'unica parola con trattino infisso. Ad esempio ”disprezzare” , ”prin-cipio”, ”re- latore”. Nei casi in cui neppure il trattino viene interpretato correttamente, la parola viene spezzata in due parti, es. ”seces sionista”.

- Riconoscimento dei caratteri alfabetici (”charError”): difetti nella qualità di stampa, macchie sulla carta, pieghe, graffi sul microfilm, ecc. possono portare ad una errata interpretazione dei caratteri. Ad esempio ” Kffégati” al posto di ”delegati”, ”coiitr'amnvirag'iao” anziché ”contrammiraglio”, ”cattchlico” per ”cattolico”. Un caso particolare è rappresentato dalla confusione tra lettere e numeri, ad esempio ”c0n” invece di ”con”, ecc.

- Sequenza delle parole (”wordSequence”): talvolta l'individuazione delle righe di testo operata dall'OCR può commettere errori dando origine a un testo dove le righe sono frammentate e mescolate impropriamente.

- Interpretazione di elementi grafici (”graphics”): linee, disegni, immagini possono essere interpretate dall'OCR come testo, dando origine a sequenze di caratteri errate.

- Punteggiatura (”punct”): l'errata interpretazione della punteggiatura può portare all'introduzione di segni di punteggiatura inesistenti, o all'assenza di segni necessari.

Spesso accade che punti, virgole, apici appaiono in posti sbagliati, ad es. ”sconqua.ssate””

Altri errori di OCR rilevanti per l'analisi automatica del testo riguardano l'interpretazione di maiuscole e minuscole, (ad es. ”DOSti” anziché ”posti”) e il significato delle parole: gli errori nel riconoscimento dei caratteri alfabetici possono dare origine a parole di senso compiuto che possono essere corrette soltanto considerando il contesto in cui occorre la parola (ad esempio ”casa” per ”cosa”, ”Baciale” al posto di ”sociale”).

2.2 Risultati dell'analisi degli errori di OCR

L'annotazione manuale degli errori di OCR è stata effettuata utilizzando una piattaforma web sviluppata ad hoc per consentire una annotazione collaborativa e veloce. Oltre ad annotare l'errore, il linguista annotava anche la possibile correzione. Sono stati annotati a mano 894 articoli di prima pagina del periodo 1930-2005, secondo le modalità descritte nel paragrafo precedente. Gli errori annotati sono complessivamente 16.842. I più frequenti sono gli errori di tipo ”charError”, cioè errori nell'interpretazione dei caratteri di una parola; seguiti dagli errori di tipo ”hyphenationSeparate”, cioè casi in cui una parola che andava a capo è stata interpretata come due parole distinte, con o senza trattino infisso.

A titolo di esempio elenchiamo alcuni degli errori più frequenti nelle edizioni dei decenni '90 e 2000, rilevanti per la NER: l'articolo ”una” è trascritto come ”ima”; la sequenza di caratteri ”li” viene riconosciuta come ”h” (ad es. l'articolo ”gli” é spesso scritto: ”gh”, ”pohtica” = ”politica”, ”poh”=”poli”); ”o” si trova scritto come ”0”; la lettera ”c” è interpretata come ”e” (es: ”dc” diventa ”de”, ”pci” diventa ”pei”).

Una analisi sistematica degli errori indotti da OCR direttamente sui nomi propri, abbastanza frequenti e variegati, sarebbe sicuramente interessante e non banale. Tra le annotazioni automatiche di persone, ad es., sono emerse le menzioni ”dustin hoffmann”, ”dustin hoflman”, ”dustin hoftman”, ”dustin holfman”, ”dustin hollman”, ”dustin hotfman”, ”dustin hotlman”, che potrebbero riferirsi all'attore americano.

Il post-processing dei documenti consentirebbe la correzione di alcuni degli errori risultati dall'OCR, utilizzando diverse tecniche, tra le quali:

- utilizzo di risorse linguistiche e semantiche, come dizionari ad alta copertura, risorse del semantic web come DBpedia, pattern sintattici;
- utilizzo di modelli statistici creati con apprendimento automatico;
- correzione manuale da parte degli utenti in modalità crowdsourcing, realizzata ad es. nel British Newspaper Archive [2] e nell'archivio dei Digitised newspapers della National Library of Australia [3]

3 Il riconoscimento delle menzioni di entità

All'analisi degli errori di OCR è seguito l'arricchimento semantico dei documenti tramite il riconoscimento automatico delle entità nominate (o ”Named Entity Recognition”, NER), cioè le persone, i luoghi e le organizzazioni menzionate negli articoli. Oltre alle persone citate nei testi, abbiamo annotato automaticamente gli autori degli articoli, per aggiungere un metadato utile ma, inaspettatamente, non banale da riconoscere.

Per effettuare il riconoscimento delle entità abbiamo utilizzato un metodo misto di apprendimento automatico e regole linguistiche, cioè abbiamo applicato in cascata un classificatore automatico SVM (Support Vector Machine) e un annotatore a regole (pattern linguistici). L'apprendimento automatico ha ovviamente richiesto una fase di annotazione manuale per creare il training set e il test set, utilizzato per valutare l'accuratezza.

3.1 Annotazione manuale e automatica

La vastità dell'archivio, sia come numero di documenti che come copertura temporale (5 milioni di articoli dal 1910 al 2005) e la varietà dei documenti (tutti gli articoli del giornale, dalla politica allo sport, dalla cultura alla cronaca, da inizio novecento al 2005) hanno posto problemi di scelta del corpus di articoli da annotare a mano per creare il data set di sviluppo.[4]

Abbiamo effettuato l'annotazione manuale delle menzioni di entità su un corpus di circa

[2] http://www.britishnewspaperarchive.co.uk

[3] http://trove.nla.gov.au

[4] La selezione del corpus di sviluppo è stato un processo molto articolato che non possiamo descrivere in dettaglio nel presente articolo.

1800 articoli, selezionati prevalentemente dalle prime pagine, dal 1910 al 2005 (per un totale di circa 582.000 token). Nell'annotazione manuale abbiamo seguito, per quanto possibile, le linee guida dell'I-CAB, Italian Content Annotation Bank, utilizzato a partire da Evalita 2007 nel task di Named Entity Recognition su articoli di giornale in italiano [5], contenente circa 525 articoli (I-CAB Evalita, 2007) del giornale L'Adige dell'anno 2004.

Il riconoscimento automatico di entità in testi storici che presentano errori di OCR è un tema affrontato in letteratura, ad es. in (Packer, 2010) e più recentemente in (Rodriquez, 2012), che riceverà probabilmente maggiore attenzione nei prossimi anni, grazie alla maggiore diffusione di archivi storici in formati digitali. Non disponendo di una soluzione sicura per questo problema, né di studi specifici per l'italiano, abbiamo deciso di utilizzare una metodologia di NER affidabile ed efficiente, cioè quella descritta in (Pianta, 2007), e utilizzata nel sistema che aveva dato i risultati migliori in Evalita 2007. Nelle edizioni successive di Evalita (2009 e 2011) le percentuali di accuratezza non sono migliorate in modo significativo e sono, viceversa, peggiorate nel task di NER da trascrizioni di notizie radiofoniche (Evalita, 2011), che contengono errori di trascrizione.

L'analisi linguistica di pre-processing del testo è stata effettuata con la pipeline UIMA (Apache UIMA, 2009) di CELI (annotatori UIMA in cascata per tokenizzazione, sentence splitting, analisi morfologica, disambiguazione, uso di gazetteers, ecc).

Il componente SVM utilizzato per il training e la creazione del modello è YamCha, Yet Another Multipurpose CHunk Annotator ((Kudo, 2001)). Per l'analisi automatica dei 5 milioni di testi, abbiamo integrato nella pipeline UIMA un componente di classificazione delle NE (cioè un annotatore di NE) che utilizzava il modello SVM creato. Dopo l'annotazione automatica con SVM, veniva applicato un componente a regole (che usava pattern linguistici), indispensabile per migliorare la correttezza delle annotazioni sia in casi particolari, come il riconoscimento degli autori, sia in altri casi rilevanti che non erano stati inclusi nel corpus di training.

[5] http://www.evalita.it/2007/tasks/ner

3.2 Risultati della NER

Forniamo qui sinteticamente alcuni dati sui risultati ottenuti dall'annotazione automatica dell'Archivio. Nella tabella seguente mostriamo il numero di named entities estratte da 4.800.000 articoli (solo le named entities che occorrono più di 10 volte) :

Tipo di NE	Num di NE	Num di documenti
PER	113.397	1.586.089
GPE	10.276	1.693.496
ORG	6.535	1.203.345
Autori	1.027	350.732

Nella tabella seguente mostriamo le misure di accuratezza (precision e recall), ottenute sul corpus di testing di 500 documenti.

Tipo di NE	Precision %	Recall %
PER	80.19	78.61
GPE	87.82	82.54
ORG	75.47	50.49
Autori	91.87	47.58

Tra le entità "standard", quelle di tipo ORG si sono dimostrate le più difficili da annotare automaticamente, come era prevedibile. Sorprendentemente invece è stato difficile il riconoscimento automatico degli autori, a causa degli errori di segmentazione degli articoli, dell'uso di sigle, della posizione variabile a inizio articolo o alla fine, e della mancanza, a volte, di punteggiatura nelle porzioni di testo rilevanti.

Conclusioni

In questo breve articolo abbiamo accennato alcune delle metodologie e delle problematiche del progetto di annotazione automatica di 5 milioni di articoli dell'Archivio Storico de La Stampa. Abbiamo segnalato alcune difficoltà legate alla presenza considerevole di errori di OCR e alla vastità e varietà dell'archivio (l'intero archivo va dal 1867 dal 2005). Queste problematiche potrebbero essere affrontate positivamente utilizzando informazioni e metodologie che non abbiamo potuto sperimentare in questo progetto, come ad es. il crowdsourcing.

Acknowledgments

Ringraziamo Francesco Cerchio, Vittorio Di Tomaso, Dario Gonella, Gianna Cernuschi, Roberto Franchini e gli altri colleghi che hanno contribuito alla realizzazione del progetto.

References

Anderson, N., A. Conteh and N. Fitzgerald. 2010. *IMPACT: Building Capacity in Mass Digitisation*. Presentation at the IS&T Archiving conference (1-4 June, The Hague, The Netherlands)

The Apache Software Foundation. 2009. *Apache UIMA Specifications (Unstructured Information Management Architecture)* http://uima.apache.org/uima-specification.html The Apache Software Foundation

Rose Holley. 2009. *How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs* D-Lib Magazine.

Taku Kudo e Yuji Matsumoto. 2001. *Chunking with Support Vector Machines*. Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies Pages 1-8

Valentina Bartalesi Lenzi, Manuela Speranza, Rachele Sprugnoli 2013. *Named Entity Recognition on Transcribed Broadcast News at EVALITA 2011*. In Evaluation of Natural Language and Speech Tools for Italian, International Workshop, EVALITA 2011, Rome, Italy, January 24-25, 2012. Springer 2013 Lecture Notes in Computer Science

Packer, T. L., J. F. Lutes, A. P. Stewart, D. W. Embley, E. K. Ringger, K. D. Seppi, and L. S. Jensen. 2010. *Extracting person names from diverse and noisy OCR text*. Proceedings of the fourth workshop on Analytics for noisy unstructured text data. Toronto, ON, Canada: ACM.

Pianta, E., Zanoli, R. 2007. *Exploiting SVM for Italian Named Entity Recognition*. In Intelligenza Artificiale, Special Issue on NLP Tools for Italian, IV-2.

K.J. Rodriquez and Michael Bryant and T. Blanke and M. Luszczynska 2012. *Comparison of Named Entity Recognition tools for raw OCR text*. KONVENS Proceedings 2012 (LThist 2012 workshop).

Evan Sandhaus. 2008. *The New York Times Annotated Corpus*. Linguistic Data Consortium, Philadelphia.

Manuela Speranza 2007. *EVALITA 2007: The Named Entity Recognition Task*. In Proceedings of EVALITA 2007, Workshop held in conjunction with AI*IA, Rome, 10 September 2007. Intelligenza artificiale, 4-2.