Semplificazione automatica e personalizzata di documenti amministrativi

Sara Tonelli, Alessio Palmero Aprosio, Marco Pistore
Fondazione Bruno Kessler
{satonelli,aprosio,pistore}@fbk.eu

Abstract

Questo contributo descrive le attività di semplificazione automatica di testi italiani svolta all'interno del progetto europeo H2020 SIMPATICO (SIMplifying the interaction with Public Administration Through Information technology for Citizens and Companies). In particolare, presentiamo alcuni esperimenti di profilazione della complessità del testo, attualmente integrati in un *authoring tool* utilizzato al Comune di Trento, e di semplificazione lessicale adattiva, che tiene conto della lingua madre dell'utente.

1 Introduzione

L'Agenzia per l'Italia digitale (AgID) ha presentato nel 2018 il Libro Bianco sull'Intelligenza Artificiale al servizio del cittadino[1], curato dalla task force promossa dalla stessa autorità, al fine di studiare le opportunità offerte dall'IA nel miglioramento dei servizi pubblici e del rapporto tra pubblica amministrazione e cittadini. Tra le sfide dell'IA al servizio del cittadino identificate nel documento, quella *tecnologica* ricopre particolare rilevanza, e in questo contesto il problema della semplificazione della comunicazione tra pubblica amministrazione (PA) e cittadino viene evidenziata come uno degli ambiti in cui l'IA potrebbe essere utilizzata efficacemente. Un'altra delle sfide elencate nel Libro Bianco, la *prevenzione delle diseguaglianze*, è altrettanto rilevante, poiché l'adozione di tecnologie di IA possono contribuire a ridurre divari socio-economici tra cittadini, tra cui il gap linguistico che coinvolge non solo la fetta di popolazione coinvolta nel flusso migratorio ma anche le fasce a bassa scolarità.

In risposta a queste sfide, sono state svolte diverse attività nell'ambito del progetto H2020 SIMPATICO (SIMplifying the interaction with Public Administration)[2], coordinato dalla Fondazione Bruno Kessler, che si è occupato dell'applicazione di soluzioni IA all'interno degli sportelli online delle amministrazioni comunali. Nello specifico, FBK ha sviluppato delle tecnologie di profilazione della complessità del testo, integrate in un *authoring tool* attualmente in uso al Comune di Trento, e un sistema adattivo di semplificazione automatica del testo che seleziona termini complessi in base alla lingua madre dell'utente. Queste due tecnologie saranno dettagliate nelle sezioni seguenti.

2 Profilazione automatica del testo

La profilazione automatica del testo è un processo che identifica, in un testo di input, le componenti linguistiche più difficili da comprendere per l'utente, fornendo una misurazione della complessità lessicale, sintattica e semantica attraverso alcune metriche di leggibilità. Alcuni lavori sono stati presentati in passato relativi alla profilazione dell'italiano [Tonelli *et al.*, 2012], ma nessuna è stata utilizzata per lo sviluppo di un sistema a supporto della scrittura di documenti, un *authoring tool*. L'unico sistema attualmente disponibile per la profilazione di testi italiani è READ-IT [Dell'Orletta *et al.*, 2011], il quale però è accessibile soltanto tramite interfaccia online, mentre il relativo software non è disponibile per il download. Per questo, abbiamo sviluppato il modulo di analisi della leggibilità integrato nella suite open source TINT per l'analisi di testi italiani [Aprosio e Moretti, 2018], disponibile anche nella repository github del progetto SIMPATICO[3]. Il modulo fornisce una serie di analitiche che includono statistiche sulla lunghezza media del testo in frasi, parole, lemmi e parole semanticamente piene e sulla presenza di nomi, aggettivi, verbi e avverbi. Inoltre, comprende alcune metriche di leggibilità che aiutano a identificare eventuali elementi complessi nel testo, fornendo non soltanto l'indice Gulpease [Lucisano e Piemontese, 1988], considerato la metrica standard per l'italiano, ma anche altre metriche relative alla variabilità lessicale (es. rapporto tipo-unità, incidenza di parole presenti nel lessico dell'italiano di base di Tullio de Mauro, densità lessicale), e alla complessità sintattica (es. numero medio di frasi per periodo, profondità dell'albero sintattico). L'authoring tool che integra queste funzionalità è stato sviluppato in due versioni: una più estesa[4], per mostrare i risultati del progetto SIMPATICO, include anche funzionalità per altre lingue, ma è stato giudicato troppo complesso e dalla bassa usabilità dal personale del Comune di Trento coinvolto nella sperimentazione del tool. Per questo motivo, è stato adottato un approccio di co-design per riprogettare l'interfaccia grafica e selezionare

[1] https://ia.italia.it/assets/librobianco.pdf

[2] https://www.simpatico-project.eu/

[3] https://github.com/SIMPATICOProject/simpatico-adaptation-engines/wiki/SAT

[4] Disponibile qui: http://simpatico.fbk.eu/demo2

soltanto metriche utili e interpretabili ai fini della redazione di documenti. Per questo motivo, è stata implementata una seconda versione della piattaforma[5] che, anche attraverso l'uso del colore e di raccomandazioni specifiche relative a aspetti linguistici delle singole frasi, supporta il Comune di Trento nella redazione di documentazione per garantirne una elevata leggibilità. Il tool consente di copiare un testo in italiano e, in tempo reale, fornisce le analisi e le raccomandazioni necessarie per modificare il testo se necessario e ridurne la complessità.

3 Semplificazione adattiva del testo

La seconda applicazione sviluppata nel progetto SIMPATICO consente di effettuare una semplificazione lessicale adattiva, basata sulla lingua madre dell'utente. In particolare, mentre sistemi tradizionali di semplificazione automatica selezionano le parole considerate difficili utilizzando repertori di frequenza o il lessico di base di Tullio de Mauro, noi abbiamo implementato un algoritmo rivolto a cittadini con bassa conoscenza dell'italiano che, in base alla lingua madre dell'utente, individua le parole complesse da semplificare. A questo fine, abbiamo fatto training di un classificatore binario che, per diverse coppie di lingue, identifica falsi amici (parole simili in due lingue diverse con significati divergenti, per es. *attualmente* e *actually*) e cognati (cioè parole che sono simili in due lingue perché imparentate, per es. *colore* e *colour*). Le coppie considerate sono Italiano – Spagnolo, Italiano – Francese, Italiano – Tedesco e Italiano – Inglese. Il tool è stato allenato su liste di cognati e falsi amici create manualmente, descritti nella Tabella 1.

Lingue	Cognati	Falsi amici
Ita-En	960	1,144
Ita-Fr	940	591
Ita-De	466	170
Ita-Sp	523	384

Tabella 1: Statistiche sui dataset utilizzati per il training.

In via preliminare, abbiamo creato un uno spazio di embedding multilingue, allineando i due spazi monolingue creati per ognuna delle coppie di lingue di interesse secondo l'approccio descritto in [Smith *et al.*, 2017]. Successivamente, abbiamo addestrato un classificatore basato su SVM utilizzando unicamente feature derivate dal coseno di similitudine nello spazio di embedding multilingue per ogni coppia di parole da classificare. I dettagli dell'algoritmo sono riportati in [Aprosio *et al.*, 2018], in cui però presentiamo soltanto una valutazione per la coppia italiano-francese. Per le quattro coppie di lingue considerate, la performance del classificatore è riportata nella Tabella 2.

Alla luce dei risultati della valutazione, considerati soddisfacenti, il classificatore è stato integrato nella versione più completa dell'authoring tool, consentendo un'individuazione adattiva delle parole semplificate in base alla lingua madre indicata dall'utente.

[5]https://simpatico.smartcommunitylab.it/simp-engines/tae/webdemo/index-tn.html

	Precision	Recall	F1	Accuracy (%)
Inglese	0.86	0.86	0.86	84.45
Francese	0.88	0.76	0.81	84.85
Spgnolo	0.88	0.91	0.89	89.47
Tedesco	0.86	0.68	0.76	87.23

Tabella 2: Risultati della classificazione (10-fold cross-validation).

4 Conclusioni

Nel presente lavoro abbiamo presentato due applicazioni sviluppate all'interno del progetto SIMPATICO a supporto dei dipendenti del Comune di Trento nella redazione di documenti. La prima è un authoring tool che integra alcune metriche di leggibilità e raccomandazioni per redigere documenti, mentre l'altra consente di identificare le parole complesse in un testo in base alla lingua madre dell'utente.

Riferimenti bibliografici

[Aprosio *et al.*, 2018] Alessio Palmero Aprosio, Stefano Menini, Sara Tonelli, Luca Ducceschi, e Leonardo Herzog. Towards personalised simplification based on L2 learners' native language. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, 2018.

[Aprosio e Moretti, 2018] Alessio Palmero Aprosio e Giovanni Moretti. Tint 2.0: an all-inclusive suite for NLP in italian. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018.*, 2018.

[Dell'Orletta *et al.*, 2011] Felice Dell'Orletta, Simonetta Montemagni, e Giulia Venturi. Read–it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, Edinburgh, Scotland, UK, July 2011. Association for Computational Linguistics.

[Lucisano e Piemontese, 1988] Pietro Lucisano e Maria Emanuela Piemontese. Gulpease. Una formula per la predizione della difficoltà dei testi in lingua italiana. *Scuola e Città*, 3:57–68, 1988.

[Smith *et al.*, 2017] Samuel L. Smith, David H.P. Turban, Steven Hamblin, e Nils Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *ICLR*, 2017.

[Tonelli *et al.*, 2012] Sara Tonelli, Ke Tran Manh, e Emanuele Pianta. Making readability indices readable. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations (PITR@NAACL-HLT 2012)*, pages 40–48, 2012.