Problemi di accuratezza nel Sentence Splitting di testi letterari contemporanei: il caso dei romanzi di Wattpad Italia

Denise Atzori, Università di Pisa

Abstract

Questo articolo descrive un'analisi di valutazione delle performance di sei tools di sentence splitting applicati a un dataset di testi letterari della versione italiana di Wattpad; si concentra, nello specifico, sulle particolarità grammaticali e sintattiche di questo tipo di testi, mettendo in evidenza gli errori più comuni in cui incorrono i tools analizzati.

Keywords

sentence splitting, text segmentation, literary texts, Italian, Wattpad

1. Introduzione

Nel mondo del Natural Language Processing si ritiene che il Sentence Splitting sia un task risolto [4]. Per Sentence Splitting si intende la segmentazione di un testo in una o più frasi attraverso l'identificazione dei loro confini, ovvero (per le lingue occidentali) dei segni di punteggiatura che le delimitano. Non sempre, però, i segni di punteggiatura presenti nei testi identificano il confine di una frase ed è nella loro disambiguazione che consiste il task di Sentence Splitting.

Come analizzato da Radaelli e Sprugnoli nel 2024 [8], il task di Sentence Splitting è considerato un task di facile risoluzione, eppure ci sono ancora casi in cui la segmentazione automatica delle frasi presenta dei problemi: il caso dei romanzi ottocenteschi analizzati nel paper sopracitato è uno di questi.

L'estensione dell'applicazione di sistemi di NLP allo stato dell'arte a testi non contemplati in fase di addestramento viene detta Domain Adaptation e molte sono state e sono tutt'ora le campagne di

valutazione - anche italiane - che mirano a studiare l'accuratezza di questi sistemi quando testati su domini non standard[1].

Questa analisi prende in considerazione questo approccio, ampliando lo sguardo verso un dominio - quello dei testi letterari nati sul web - ancora poco esplorato, soprattutto in Italia. Oggetto dell'analisi è un corpus di romanzi estratti dalla versione italiana di Wattpad, piattaforma di lettura sociale in rapida ascesa. Lo scopo è valutare come si comportano i più popolari modelli pipeline nella segmentazione di testi letterari non standard che non appartengono al dominio di addestramento di questi tools.

2. Cos'è Wattpad e perché è interessante studiarne i testi

Negli ultimi anni si è visto un sensibile aumento nella pubblicazione cartacea di romanzi generati su Wattpad[2], piattaforma di lettura sociale che si è imposta su altre meno amate e frequentate. In Italia, è notevole il caso di romanzi come *My dilemma is you* di Cristina Chiperi (2016, Leggere Editore), primo di una saga, e *Fabbricante di lacrime* di Erin Doom (2021, Salani Editore), che hanno venduto rispettivamente oltre 150 mila copie[3] e 450 mila copie[4] anche e soprattutto grazie al passaparola su Tik Tok.

Se si vogliono studiare le narrazioni che coinvolgono persone giovani, Wattpad è un buon punto dal quale partire: la piattaforma raccoglie oltre 500 milioni di storie scritte da giovani e giovanissime/i in più di 50 lingue e lette da oltre 70 milioni di utenti - anch'essi giovani - al mese. Le analisi con strumenti di NLP su testi di Wattpad sono per ora poche, ma negli ultimi anni la piattaforma ha attirato l'attenzione di diversi gruppi di ricerca internazionali e italiani [3] [7].

Caratteristica principale di Wattpad è la natura democratica del suo catalogo: chiunque sulla piattaforma può scrivere e farsi leggere, senza dover passare attraverso una selezione; le storie sono quindi in buona parte prive di filtri editoriali, e ciò che viene scritto dalle e dagli utenti è attinto

[1] Si veda, ad esempio, la campagna di Evalita 2011 condotta proprio a questo fine: http://www.italianlp.it/resources/evalita-2011-domain-adaptation-for-dependency-parsing/

[2] Link alla versione italiana della piattaforma: https://www.wattpad.com/home

[3] Dato fornito da wired.it nel 2020.

[4] Dato fornito da illibraio.it nel 2023 per la sola prima edizione.

direttamente dal loro immaginario, spesso riversato spontaneamente tra le pagine digitali senza una reale conoscenza degli strumenti narratologici, e a volte elaborato in maniera collettiva[5].

Proprio questa "libertà narrativa" può costituire un'alleata per osservare ciò che giovani e giovanissime/i sognano e immaginano del mondo e delle relazioni, spesso (data la giovane età) ancor prima di vivere di persona le esperienze che raccontano. Questo rende Wattpad una fonte interessante di testi letterari contemporanei, da investigare con sguardo letterario e sociale attraverso lettura diretta supportata da strumenti automatici [7].

D'altrocanto, la mancanza di "sorveglianza editoriale" su questi testi e la loro natura amatoriale complica il loro trattamento automatico: spesso, infatti, i testi di Wattpad sono scritti senza rispettare - e conoscere - le norme editoriali sulla punteggiatura, presentano frequenti errori di grammatica e sintassi e si discostano notevolmente sia dagli standard letterari sia dalle convenzioni dei testi più formali utilizzati per l'addestramento dei più comuni strumenti di NLP[6].

3. Problemi di sentence splitting nei testi estratti da Wattpad

Dati gli assunti precedenti, non stupisce che durante l'analisi di testi estratti da Wattpad ci si imbatta in problemi in parte simili a quelli riportati da Radaelli e Sprugnoli [8]: strumenti di sentence splitting popolari come Stanza e spaCy, addestrati su testi di domini molto differenti - ad esempio testi legislativi, pagine Wikipedia o articoli di giornali - incorrono in diversi problemi quando si trovano a segmentare i testi estratti dalla piattaforma, nei quali la punteggiatura segue a tratti peculiarità stilistiche e creative proprie del mezzo narrativo, a tratti presenta vere e proprie infrazioni (per lo più involontarie) delle norme grammaticali tradizionali.

Si prenda come esempio questa frase, estratta da un Dataset creato a partire da storie raccolte da Wattpad Italia:

«Allie, vieni a darci una mano.» è mia madre che mi risveglia dai pensieri.

[5] Questo aspetto è riscontrabile, in molte storie, in quelli che vengono definiti "Spazi autrici", specchietti solitamente presenti al termine di un capitolo nel quale la persona che scrive si rivolge direttamente a lettrici e lettori raccontando il suo processo di scrittura e le difficoltà incontrate, e chiedendo aiuto a chi legge per decidere il futuro sviluppo della trama.

[6] Molti di questi problemi sono stati riscontrati anche dal gruppo di ricerca di Pianzola et all. nel 2020 [7].

nella quale, come si può vedere, viene infranta la regola della lettera maiuscola dopo il punto di chiusura del dialogo[7].

Ma se si guarda a quest'altra frase:

«Sono stata obbligata.» farfuglio lanciando un'occhiataccia a Trevor.

Si vede come lo stesso segno di punteggiatura (il punto a fine dialogo) viene utilizzato anche nel caso in cui al dialogo segua il verbo diegetico che lo regge.

Questo e altri fenomeni simili, molto frequenti sulla piattaforma, rendono la segmentazione automatica del testo complessa e influenzano al contempo gli stadi successivi di analisi (come ad esempio l'analisi del sentiment) che da essa dipendono strettamente [11].

Per queste ragioni, si è deciso di condurre una valutazione sulle performance di sentence splitting dei testi di Wattpad di sei tra gli strumenti di sentence splitting più utilizzati, per individuarne non solo l'accuratezza ma anche gli errori più frequenti.

3. Dataset e strumenti utilizzati

Il dataset di partenza è stato creato nell'ambito di un progetto di tirocinio e tesi triennale in linguistica computazionale e comprende le 250 storie più lette sulla piattaforma italiana al momento dell'estrazione[8]. Per questa analisi, dal dataset sono state estratte 744 frasi, corrispondenti ai primi 17 capitoli di una storia campione.

Le frasi sono state segmentate e tokenizzate a mano per creare il gold standard, tenendo conto delle scelte di punteggiatura non standard del testo analizzato e prendendo le seguenti decisioni di segmentazione:

1. Punteggiatura dentro la virgoletta seguita da verbo diegetico: non si spezza la frase
   es: *«Ora lo sai.» affermo.*
2. Punteggiatura dentro la virgoletta seguita da verbo non diegetico (con la minuscola): si spezza la frase.
   es: *«Okay ragazzi, vi devo lasciare.*
   Grazie per il braccialetto, buona giornata.»
   me ne vado di corsa per uscire da quello stato di imbarazzo che si era creato.
3. Virgolette di citazione dentro la frase (precedute o no dai due punti): non si spezza la frase

[7] Per questa e altre convenzioni di punteggiatura si è fatto riferimento a Mortara Garavelli, B. *Prontuario di punteggiatura*, Laterza, Bari, (2003).

[8] Per ragioni di copyright, il Dataset non è pubblico. I risultati delle analisi condotto sono invece disponibili al seguente link: https://github.com/DeniseAtzori/SA-Wattpad

es: *Guardo il telefono, c'è un messaggio di mio padre: "Tesoro, mi dispiace ma oggi tardo."*

4. Tre puntini di sospensione dentro la frase: non si spezza la frase (tranne nei casi in cui questi marcano la fine della frase stessa)

   es: *«Non si rivolge la parola a delle sfigate come me... possono farti perdere tempo.»*

5. Trattino di apertura e chiusura all'interno della frase: non si spezza la frase

   es: *Alla fine prendo un paio di jeans chiari, dei top di vari colori e delle magliette - alcune a tinta unita, altre con delle piccole stampe -.*

3.1 Tools di sentence splitting utilizzati

Facendo riferimento agli studi di Radaelli e Sprugnoli [8], si è scelto di utilizzare sei tools di sentence splitting per l'italiano sviluppati con approcci differenti: spaCy[9], SentenceSplitter[10], Stanza[11], UDPipe2[12], Wtp Split[13] e WtP Canine[14]. Tra questi, due sono basati su regole specifiche per ogni lingua e su liste di eccezioni e abbreviazioni (spaCy e SentenceSplitter) due sono sistemi supervisionati e dunque addestrati con frasi già correttamente segmentate (Stanza e UDPipe2) e gli ultimi due sono sistemi non supervisionati, e si basano dunque su features linguistiche come la lunghezza delle parole (WtP Split e WtP Canine).

I diversi approcci al task di segmentazione permettono di osservare il comportamento di questi tools su testi letterari non standard anche in funzione della loro diversa natura.

I tools sono stati eseguiti su Python seguendo le indicazioni del Colab notebook sviluppato da Radaelli e Sprugnoli nell'ambito della ricerca sui "Promessi sposi" e messo a disposizione su Github[15] [8].

3.2 Valutazione delle performance dei tools

Per valutare le performance dei tools, i file testuali segmentati automaticamente sono stati passati alla demo online della pipeline di UDPipe così da ottenere il formato standard CoNLL-U[16]. I CoNLL-U dei tools sono stati poi confrontati con il gold standard utilizzando l'API sviluppata nel

[9] spaCy: https://spacy.io/api/sentencizer
[10] Sentence Splitter: https://github.com/mediacloud/sentence-splitter
[11] Stanza: https://stanfordnlp.github.io/stanza/combined_models.html
[12] UDPipe2: https://ufal.mff.cuni.cz/udpipe
[13] WtP Split: https://github.com/segment-any-text/wtpsplit
[14] WtP Canine: https://github.com/segment-any-text/wtpsplit/blob/main/README_WTP.md
[15] Link al Notebook sviluppato da Radaelli e Sprugnoli: https://colab.research.google.com/drive/1j0RhdutBVAxfxgX4X3XAPf1d52_xB9xz?usp=sharing
[16] Link alla pipeline UDPipe: https://lindat.mff.cuni.cz/services/udpipe/info.php

2018 da Milan Straka e Martin Popel dell'Institute of Formal and Applied Linguistics (UFAL) della Charles University, in Repubblica Ceca [10], modificata da Giovanni Moretti nell'ambito della EvaLatin Campaigns del 2021[17].

L'API prende in input il file gold standard e il file segmentato da un tool, entrambi in formato CoNLL-U, e confronta i token, le frasi, i lemmi, le classi e altre metriche per misurare l'accuratezza della segmentazione automatica. Il risultato del confronto è un dizionario nel quale, per ciascuna metrica, vengono calcolati precision, recall ed F1 score, misure di accuratezza standard per questo tipo di analisi.

Per lo scopo di questa analisi è stata considerata solo la metrica delle frasi (Sentences) e si è fatto riferimento all'F1 score[18].

5. Risultati della valutazione

Come si può osservare dalla Tabella 1, i risultati dei tools sono molto variabili. Nessuno raggiunge performance sopra il 90 (tipiche di testi estratti dal web come i tweet) e uno di questi, spaCy sentencizer, si ferma sotto il 50.

Non sembra esserci una tipologia di modello che performa meglio delle altre e gli F1 score più alti si ottengono con SentenceSplitter (87.09, modello rule-based), WtP wtp-canine-s-121 (86.90, unsupervised) e UDPipe 2 VIT model (85.57, supervised).

Tipo	**Sistema**	**F1**
rule-based	spaCy sentencizer	44.67
rule-based	**SentenceSplitter**	**87.09**
supervised	Stanza combined	72.76
supervised	UDPipe 2 VIT model	85.57
unsupervised	Wtp Split	71.67

[17] L'API descritta è disponibile su Github in licenza libera al seguente link: https://github.com/CIRCSE/LT4HALA/blob/master/2022/data_and_doc/conll18_ud_eval_EvaLatin_2022_rev2.py

[18] I file utilizzati durante l'analisi sono disponibili su Github al seguente link: https://github.com/DeniseAtzori/Sentence-Splitting-Wattpad

unsupervised	WtP wtp-canine-s-121	86.90

Tabella 1: risultati dell'accuratezza nella metrica "Sentences" (misurata con l'F1 score) dei tools testati

Analizzando le segmentazioni automatiche effettuate dai tools, e confrontandole con il gold standard, si osservano alcuni errori comuni a quasi tutti i tools (alcuni esempi sono riportati nella Tabella 2).

Gold standard	**spaCy**	**WtP Split**
1) «Cazzo.» dico stiracchiandomi.	1) « 2) Cazzo.» 3) dico stiracchiandomi.	1) «Cazzo. 2) » dico stiracchiandomi.
1) «In realtà ieri ho comprato due magliettine e...» 2) «Non voglio obiezioni. 3) Tra dieci minuti ti voglio pronta in salone.»	1) « 2) In realtà ieri ho comprato due magliettine e...» «Non voglio obiezioni. 3) Tra dieci minuti ti voglio pronta in salone.»	1) «In realtà ieri ho comprato due magliettine e...» «Non voglio obiezioni. 2) Tra dieci minuti ti voglio pronta in salone.
1) «Guarda quello che fai, razza di squilibrato!» urlo togliendomi una cuffietta.	1) « 2) Guarda quello che fai, razza di squilibrato!» 3) urlo togliendomi una cuffietta.	1) «Guarda quello che fai, razza di squilibrato!» urlo togliendomi una cuffietta.

Tabella 2: esempi di errori dei tools, comparati con le frasi segmentate manualmente

Tra questi, il segno di punteggiatura ferma dentro il dialogo visto in precedenza sembra essere l'ostacolo più grande alla corretta segmentazione delle frasi. Anche la chiusura e riapertura dei caporali senza segno di interpunzione in mezzo "» «" crea problemi nella segmentazione automatica, e in generale la presenza dei caporali rende la segmentazione sporca e spesso errata.

Altre scelte di punteggiatura tipiche dei testi letterari come i tre puntini di sospensione dentro la frase o le virgolette di citazione non sembrano invece creare difficoltà neanche nei tools meno performanti come spaCy o WtP Split (che hanno F1 score rispettivamente di 44.67 e 71.67).

6. Conclusioni

Osservando le metriche di valutazione e gli errori più frequenti dei sei tools analizzati, appare chiaro che il sentence splitting automatico con questi strumenti, applicato a testi letterari italiani estratti dalla piattaforma Wattpad Italia, necessiti di un intervento manuale prima di poter condurre le fasi successive di analisi.

La presenza di forme di punteggiatura e di sintassi non standard e l'appartenenza dei testi a un dominio poco esplorato dagli strumenti di NLP sono fattori che complicano notevolmente quello che, in altri domini, è ormai considerato un task risolto.

Frasi mal segmentate come quelle viste nel corso di questa analisi possono condurre a errori nei successivi task, a partire da errori nel task di Dependency Parsing che spesso è la base per gli stadi successivi [1]. Diventa dunque importante, quando si lavora con testi letterari estratti da piattaforme come Wattpad, dotarsi di strumenti addestrati specificamente per questo tipo di testi, che permettano di ottenere un'accuratezza maggiore nel processo di sentence splitting. Il passo successivo a quest'analisi è dunque l'addestramento di uno degli strumenti utilizzati (tra quelli a più alta performance) attraverso un gold standard segmentato a mano, con un numero di frasi consistente che permetta di ottenere training set, test set e set di valutazione adatti all'addestramento.

È infine importante far presente che questa analisi è stata condotta su una storia campione estratta dalla piattaforma, scelta per la sua rappresentatività dei testi del Dataset ma non per questo in grado di esaurire tutti i casi di punteggiatura e sintassi non standard presenti nel corpus[19].

Sotto questo assunto, è chiaro che sia necessario condurre ulteriori analisi, che estendano i casi possibili e che si spingano anche nelle successive fasi del parsing dei testi, che possono presentare a loro volta varie problematiche generate dalla peculiare natura dei testi.

[19] A titolo di esempio, si segnala la forte presenza di errori di sintassi e grammatica nei testi analizzati, che può impattare sulla corretta creazione automatica delle strutture a dipendenze.

Bibliografia

[1] Bosco, C. Montemagni, S. Simi, M. et al. 2013. “Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank”. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, The Association for Computational Linguistics. http://medialab.di.unipi.it/downloads/ISDT/MIDT-STD2013_law.pdf

[2] Gildea, D. 2001. “Corpus variation and parser performance”. In *Proceedings of the 2001 conference on empirical methods in natural language processing*. https://aclanthology.org/W01-0521.pdf

[3] Kardiansyah, M. Yuseano. 2001. “Wattpad as a story-sharing website: Is it a field of literary production?”. In *English Language and Literature International Conference (ELLiC) Proceedings*. Vol. 3. https://jurnal.unimus.ac.id/index.php/ELLIC/article/view/4748

[4] Minixhofer, B., Pfeiffer, J. e Vulić, I. 2023. “Where’s the Point? Self-Supervised Multilingual Punctuation-Agnostic Sentence Segmentation”. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.398

[5] Mortara Garavelli, B. *Prontuario di punteggiatura*, Laterza, Bari, (2003).

[6] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher D. Manning. 2020. “Stanza: A Python Natural Language Processing Toolkit for Many Human Languages”. In *Association for Computational Linguistics (ACL) System Demonstrations*. https://doi.org/10.48550/arXiv.2003.07082

[7] Pianzola, F. Simone, R., Lauer, G. 2020. “Wattpad as a resource for literary studies. Quantitative and qualitative examples of the importance of digital social reading and readers’ comments in the margins”. *PloS one 15.1.* https://doi.org/10.1371/journal.pone.0226708

[8] Redaelli, A. Sprugnoli, R. 2024. “Is Sentence Splitting a Solved Task? Experiments to the Intersection Between NLP and Italian Linguistics”. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, Pisa, Italy. https://shorturl.at/DCahU

[9] Sprugnoli, R. Iurescia, F. Passarotti, M. 2024. “Overview of the EvaLatin 2024 evaluation campaign”. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*. https://aclanthology.org/2024.lt4hala-1.21/

[10] Straka, M. “UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task”. 2018. In: *Proceedings of CoNLL 2018: The SIGNLL Conference on Computational Natural Language Learning*, Association for Computational Linguistics, Stroudsburg, PA, USA. https://doi.org/10.18653/v1/K18-2020

[11] Wicks, R. Post, M. 2022. “Does sentence segmentation matter for machine translation?”. In *Proceedings of the Seventh Conference on Machine Translation*. https://aclanthology.org/2022.wmt-1.78/