University of Trento
Sociology Department
Master in Data Science

Sony CSL

Automatic Text Simplification for Italian Poor Readers and Comprehenders

Francesca Padovani ♠ ♢ Daniele Nardi ♡ ♣ ♥ ♢ Martina Galletti ♢ ♡ ♥
Computer Science Laboratories Sony Paris ♢ Università di Trento ♠ University "La Sapienza" of Rome♡ Centro di Studi e Ricerche Enrico Fermi ♥ CINI-AIIS, Italy ♣
♢francesca.padovani98@gmail.com ♢nardi@diag.uniroma1.com ♢martina.galletti@sony.com

Introduction

Automatic Text Simplification (ATS) is the process of modifying a text to reduce its overall linguistic complexity[5]. It is not particularly explored for Italian, because of data scarcity and poor data quality [4], [3]. The output of this work is three-fold:

1. Built a new enriched corpus [1];
2. Fine-tuned a transformer-based encoder-decoder model ;
3. Parameterise grammatical text features to control simplifications [2];

***Keywords*— NLP, Automatic Text Simplification, Sentence-to-Sentence Simplification, DSA**

1 Put up the Dataset

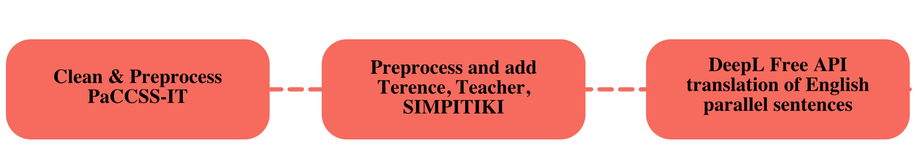

Figure 1: The steps we took in order to construct the Augmented Dataset

	COMPLEX	SIMPLE
PaCCSS-IT	Quale sarebbe allora la soluzione giusta?	È questa la soluzione giusta?
Teacher	I bei tempi finirono nel maggio 1940; prima la guerra, la capitolazione, l'invasione tedesca, poi cominciarono le sventure per noi ebrei.	I tempi felici finiscono nel maggio 1940; dopo la guerra, la sconfitta, e l'arrivo dei soldati tedeschi, cominciano i problemi per noi ebrei.
Terence	Ernesta Sparalesta è una bambina alta, più o meno, un metro e una noce.	Ernesta Sparalesta è una bambina alta poco più di un metro.
SIMPITIKI	Said spiega che questo processo è stato reso possibile attraverso una conoscenza superficiale di ciò che è in effetti l'Oriente.	Said spiega che questo processo si è realizzato mediante una conoscenza superficiale di ciò che è in effetti l'Oriente.
Translated English Sentences	L'orso bruno dell'Himalaya, noto anche come orso rosso dell'Himalaya, orso isabellino o Dzu-Teh, è una sottospecie dell'orso bruno.	L'orso bruno himalayano è una sottospecie dell'orso bruno.

Figure 2: Composition of the Augmented Dataset

2 Experimental Set-Up

We used a pretrained **transformer based encoder-decoder model** and we finetuned it on the simplification task using our dataset.
We made use of a pretrained model's checkpoints. Its id is *bert-base-italian-xxl-cased*. You can find it in the HuggingFace Hub repository.

3 Results of the first training on PaCCSS-IT

Complex Sentence	La ringrazio per l' aiuto.
Simplified Sentence	La ringrazio per il suo prezioso aiuto.
Reference Sentence	La ringrazio tantissimo per il suo aiuto!

Table 1: Comparison between senttences.

Normal	Simple
La ringrazio dell' attenzione e la saluto cordialmente	La ringrazio per l'attenzione prestatami e la saluto cordialmente.
Questa sarebbe una cosa positiva " .	Questa è una cosa gravissima
Grazie Si rivolga alle università degli studi .	Deve rivolger - si alle università degli studi
E a posteriori è anche facile	E a lungo è stato tutto più facile.
Ma facciamo un passo più avanti .	Ma facciamo un lungo passo indietro
Così non si va da nessuna parte !	Non potete andare da nessuna parte!

Figure 3: Result of training on PaCCSS-IT.

The valuation Metrics' Values:
SARI = 57.41
BLEU = 49.65

4 Second Training on Augmented Dataset

Normal	Simple1
Nel mio caso si tratta di questo	In questo caso si tratta di un ulteriore caso
Ma c' è un altro nodo da sciogliere	Il vero nodo da sciogliere era un'altro
Aidan stava sudando perché era caldo, ed era fuori allenamento	Aidan stava sudando perché era molto caldo, ed era felice
Si , ma è un processo lento	è comunque un processo lento
Ma non è una cosa così scontata	Ma non deve essere così semplice
Anche questo è un segno dei tempi	è il segno dei tempi

Figure 4: Result of training on the Augmented Dataset.

The valuation Metrics' Values:
SARI = 51.51
BLEU = 47.40

5 Third training with the addition of the Adaptive Component

semplifica: 0.20 0.20 0.90 0.90 1 Mi sembra una cosa molto buona	Mi sembra una cosa buona
semplifica: 0.20 0.20 0.90 0.90 1 Poi vi sono le situazioni particolari	Vi sono situazioni particolari
semplifica: 0.20 0.20 0.90 0.90 1 Dopo il comma 1 , inserire il seguente comma : 2	Dopo il comma 1 inserire il comma 2
semplifica: 0.20 0.20 0.90 0.90 1 La risposta è decisamente " no	La risposta è no

Figure 5: Result of training on the Adaptive Dataset.

The valuation Metrics' Values:
SARI = 47.76
BLEU = 29.00

6 More information at:

Figure 6: Scan the QR Code

References

[1] Gianni Barlacchi and Sara Tonelli. Ernesta: A sentence simplification tool for children's stories in italian. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 476–487. Springer, 2013.

[2] Louis Martin, Benoît Sagot, Eric de la Clergerie, and Antoine Bordes. Controllable sentence simplification. *arXiv preprint arXiv:1910.02677*, 2019.

[3] Angelo Luigi Megna, Daniele Schicchi, Giosué Lo Bosco, and Giovanni Pilato. A controllable text simplification system for the italian language. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 191–194. IEEE, 2021.

[4] Alessio Palmero Aprosio, Sara Tonelli, Marco Turchi, Matteo Negri, and A Di Gangi Mattia. Neural text simplification in low-resource conditions using weak supervision. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation (NeuralGen)*, pages 37–44. Association for Computational Linguistics (ACL), 2019.

[5] Kim Cheng Sheang and Horacio Saggion. Controllable sentence simplification with a unified text-to-text transfer transformer. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 341–352, Aberdeen, Scotland, UK, August 2021. Association for Computational Linguistics.