



CENTER FOR
DIGITAL SOCIETY

FONDAZIONE
BRUNO KESSLER

Towards automatic spoken grammatical error correction of L2 learners of English

Stefano Bannò, Michela Rais, Marco Matassoni



UNIVERSITY
OF TRENTO



SPEECHTEK
SPEECH TECHNOLOGY LAB

Ital-IA 2023 - Pisa 29-31 Maggio 2023

{matasso,sbanno}@fbk.eu

Our research topics



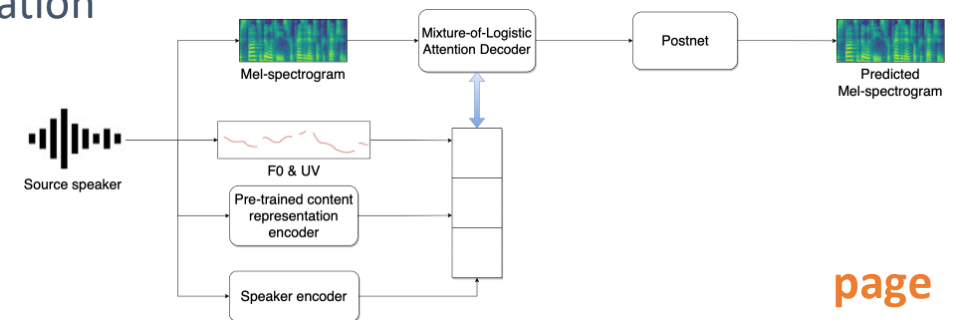
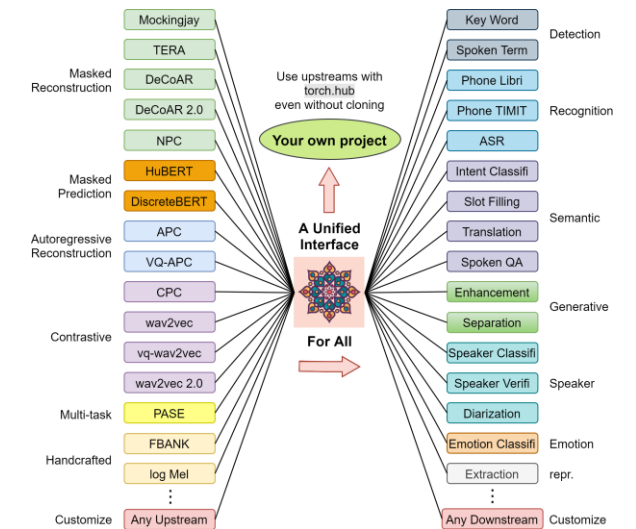
automatic speech recognition and understanding

AI topical research directions to the speech context, such as: continual learning, large-scale models, self-supervised adaptation, edge processing



speech processing

speech enhancement and separation, speaker identification/verification, voice conversion/anonymization



spoken language proficiency



automated approaches to evaluate proficiency of L2 learners of English in Computer-Assisted Language Learning (CALL)

language is used to communicate meaning and requires distinct competences:

- linguistic: pronunciation, vocabulary, grammar
- sociolinguistic: politeness, socio-pragmatics
- pragmatic: fluency, coherence and cohesion, turn-taking

goals

improve the assessment of L2 learners' grammar with spoken grammatical error correction (GEC)



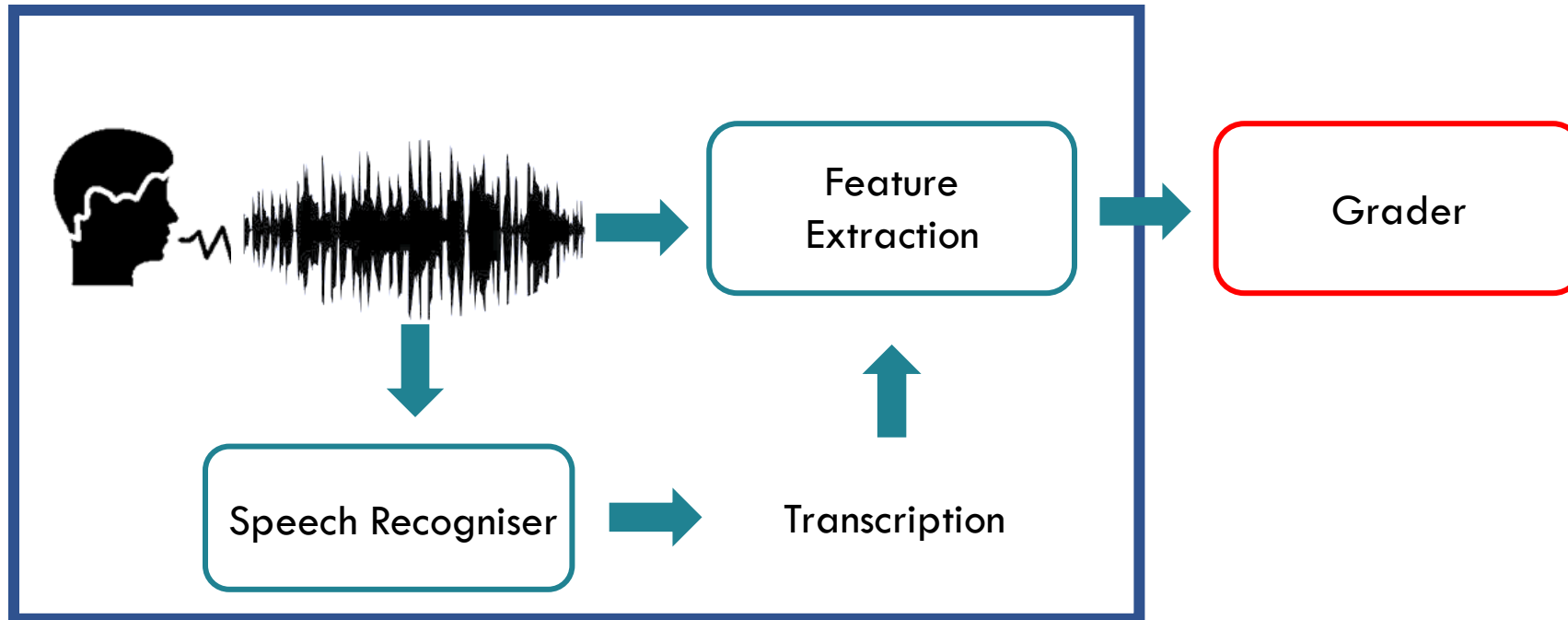
spoken GEC still relatively new with limited availability of specifically designed and annotated corpora for the assessment of spoken language proficiency

Language scoring and assessment

Written



Spoken



Challenges



L2 learner speech is often hard to transcribe even for humans



Lack of publicly available data specifically designed and annotated for assessment purposes



ASR might introduce errors



Human-annotated grades often suffer from inconsistency and incoherence

Written vs Spoken

L2 speech is often characterized by **disfluencies** which appear in transcriptions.

Written GEC

Original: He **see** the thief is **catched** by policeman the last night

Corrected: He **saw** the thief **caught** by a policeman last night

Spoken GEC

Original: **uhm** he **see the** the thief is **catched** by policeman the **la-** last night

Corrected: He **saw** the thief **caught** by a policeman last night

Trentino Language Testing



TLT-school: a Corpus of Non Native Children Speech

Roberto Gretter, Marco Matassoni, Stefano Bannò, Daniele Falavigna

This paper describes "TLT-school" a corpus of speech utterances collected in schools of northern Italy for assessing the performance of students learning both English and German. The corpus was recorded in the years 2017 and 2018 from students aged between nine and sixteen years, attending primary, middle and high school. All utterances have been scored, in terms of some predefined proficiency indicators, by human experts. In addition, most of utterances recorded in 2017 have been manually transcribed carefully. Guidelines and procedures used for manual transcriptions of utterances will be described in detail, as well as results achieved by means of an automatic speech recognition system developed by us. Part of the corpus is going to be freely distributed to scientific community particularly interested both in non-native speech recognition and automatic assessment of second language proficiency.





The TLT GEC corpus

TLT-GEC used as **test set**

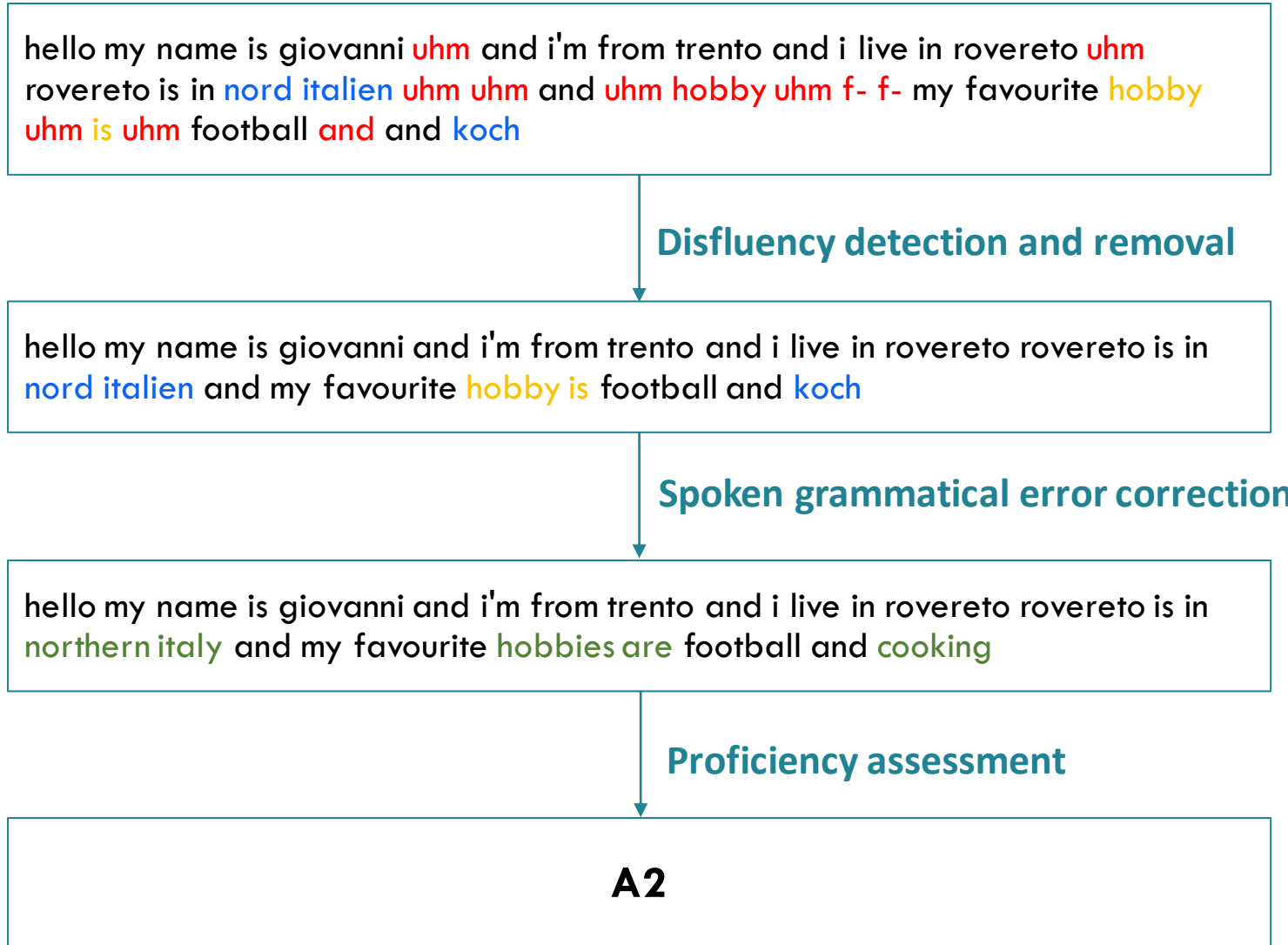
- Section of corpus of spontaneous speech utterances of young Italian learners of English (CEFR levels ranging A2 and B1) recorded in schools in Trentino between 2017-18
- ASR transcriptions (5h of recordings), manually segmented into sentences (1127) and annotated with disfluencies + two sets of grammatical error corrections by two different human annotators.

Spoken example

hello my name is giovanni uhm and i'm from trento
and i live in rovereto uhm rovereto is in nord italien
uhm uhm and uhm hobby uhm f- f- my favourite
hobby uhm is uhm football and and koch

-  **Disfluencies** (hesitations, repetitions, false starts, etc.)
-  **Code-switched words** (from L1 and L3)
-  **Grammatical errors**
-  **Named-entities**

A possible pipeline



Disfluencies Detector Training

DD - a binary sequence tagging task using a BERT-based token classifier; binary tag indicates whether each word is fluent or disfluent.

BERT-based model consists of a BERT layer in the version provided by the HuggingFace Transformer Library (*bert-base-uncased*), a dropout layer, a dense layer of 768 nodes, a dropout layer, another dense layer of 128 nodes, and finally the output layer.

Model is trained on NICT-JLE and KIT Speaking Test Corpus

The National Institute of Information and Communications Technology Japanese Learner English corpus is a collection of manual transcriptions of approximately 300 hours of oral interviews of Japanese learners of English

The Kyoto Institute of Technology Speaking Test Corpus consists of manual transcriptions of approximately 4,448 hours of interviews of 574 Japanese undergraduate students

Grammar Error Correction Training

GEC - T5 model initialized from the version provided by the HuggingFace Transformer Library (*t5-base*)

Model is trained on EFCAMDAT and BEA-2019

EFCAMDAT consists of 1,180,310 scripts written by 174,743 L2 learners, annotated with POS tags and information on grammatical dependencies, and are partially error-tagged by human experts

BEA-2019 text-based corpora (Cambridge Learner Corpus First Certificate English, Write & Improve, LOCNESS, Lang-8, National University of Singapore Corpus of Learner English) tagged with GEC annotations

Metrics: MaxMatch (M^2) and General Language Evaluation Understanding (GLEU) – higher is better

Results

		GLEU \uparrow	M^2 \uparrow
CLC-FCE test	Our model	70.05	57.86
	[4]	-	56.60
TLT-GEC test (manual transcriptions)	Agreement	80.32	79.86
	dsf	35.73	49.11
	flt	66.44	65.81
	autoflt	58.89	57.65
TLT-GEC test (ASR transcriptions)	dsf	33.85	39.23
	autoflt	38.35	40.45

dsf: original transcriptions with disfluencies

flt: transcriptions in which disfluencies have been manually removed

autoflt: transcriptions with automatically removed disfluencies through DD

[4] – model proposed in (Lu, Bannò, Gales, 2022)

Findings and open questions

- Spoken GEC performance can be improved when disfluencies are removed from the transcriptions
- DD positively impacts GEC performance both on the manual and ASR transcriptions of the TLT-GEC corpus
- improvements can be achieved using only publicly available data for training

Open problems:

- GEC model struggles with common use of Italian named entities as well as code-switched words
- Relevance of the answer:

Q: What country would you like to visit in the future? Why?

A: I like to visit turkey because i like speaking the language

Conclusions

- Speech technologies applied to language proficiency assessment: problems and solutions
- Integration of ASR and NLP neural models
- Interesting directions for language learning using model feedback (grammar)
- Need for a multi-disciplinary approach: inject pragmatic aspects into models, interpretability



Thank you.

Questions?