Tackling Italian University Assessment Tests with Transformer-Based Language Models

Daniele Puccinelli[1], **Silvia Demartini**[1], **Pier Luigi Ferrari**[2]
1. University of Applied Sciences and Arts of Southern Switzerland, Switzerland
2. University of Eastern Piedmont, Italy
{daniele.puccinelli, silvia.demartini}@supsi.ch,
pierluigi.ferrari@uniupo.it

Abstract

Cloze tests are a great tool to asses reading proficiency as well as analytical thinking, and are therefore employed in admission and assessment tests at various levels of the education system in multiple countries. In Italy, cloze tests are administered to incoming university students to ascertain their starting level. The goal of a cloze test is to determine several tokens that have been pre-deleted from a text; this is largely equivalent to the well-known NLP task of missing token prediction. In this paper, we show that cloze tests can be solved reasonably well with various Transformer-based pre-trained language models, whose performance often compares favorably to the one of incoming Italian university students.

1 Introduction

A cloze test is a reading comprehension assessment where participants are presented with a text in which selected tokens have been replaced with blanks. The goal is for the participant to choose tokens (often from a list) and use them to replace the blanks based on the overall context. Typically, one every 5-10 tokens is replaced with a blank.

Cloze tests are one of the most common linguistic tests in use for formative and summative purposes, along with written responses, multiple-choice tests, matching tests, ordering tests, summarizing tests etc. (Lugarini, 2010). Cloze tests were originally introduced in the United States in the 1950s to measure the readability of texts (Taylor, 1953) and involved the random and not predetermined deletion of words that appeared at predefined intervals. This method was too general for didactic and evaluation purposes, but it was quickly adapted and became very widespread as a teaching and testing technique (Radice, 1978). In education, cloze tests have become more targeted: words are deleted according to various criteria, depending on the specific testing goals. In general, cloze tests are designed to evaluate one of the following:

- field-specific knowledge acquisition, by asking to insert appropriate words about a topic or a discipline;
- text comprehension, by asking for information that can be inferred from the text (with no prior domain knowledge);
- linguistic aspects, typically with respect to L1, L2 and FL (foreign language) acquisition at different levels (i. e. vocabulary, specific parts of speech etc.).

If carefully designed, cloze tests can be a very effective tool at all educational levels; on the other hand, cloze tests may also show some limits and issues in assessing linguistic competence (Chiari, 2002), as they necessarily offer a partial and contextual view. However, the long tradition of study and use in the fields of educational linguistics and linguistic makes it very interesting to compare human and automatic performances in dealing with cloze tests.

2 Methodology

We tackle the cloze tests in our dataset with pre-trained language models based on the Transformer architecture (Vaswani et al., 2017). We employ both autoencoding and autoregressive models. Given the very small number of datapoints at our disposal, model fine-tuning is not a viable option; therefore, we use pre-trained versions of

such models, all of which are publicly available through Huggingface at the time of writing (summer 2021).

Dataset. Our dataset contains eleven cloze tests focusing on general linguistic competence that were administered to incoming first year students at the University of Eastern Piedmont in the cities of Alessandria and Vercelli in northwestern Italy between 2017 and 2019. Each cloze test was taken by a number of students in the low three digits, ranging from 130 to 390. As these are university-level tests, all students had at least a high school diploma. Most of the students were L1. The tests were offered on-site (in information technology classrooms) through the Moodle Learning Platform.

Our dataset contains two types of cloze tests: nine restricted tests where a list of options is provided for each blank to be filled, and two unrestricted tests where a global list of options is provided for all blanks with no token subgrouping (i.e., with no information about which tokens are supposed to go where). In the two unrestricted tests and three of the nine restricted ones, the list(s) contain single token options. In the other six restricted tests, the lists contain at least one multiple token option (e.g., *il quale* or *con l'utilizzo*). These cloze tests involved both function words as well as content words with both lexical and grammatical meanings

Autoencoding models. Our choices for autoencoding models are BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), DistilBERT (Sanh et al., 2019), and ELECTRA (Clark et al., 2020).

BERT is a natural choice because one of its two pre-training tasks is masked language modeling: a fraction of tokens in the pre-training data are masked so that BERT can be pre-trained to reconstruct them. Viewed as an NLP task, a cloze test is a special case of masked language modeling task where tokens are masked in an adversarial fashion: instead of choosing tokens to be masked uniformly at random, tokens are masked to challenge the test taker to reconstruct the meaning of the original text. Because a cloze test is functionally equivalent to a masked language modeling task, it is reasonable to use pre-trained BERT with no further task-specific fine-tuning.

RoBERTa improves on the original BERT by focusing on the aforementioned masked language modeling task and removing the other pre-training task (next sentence prediction). UmBERTo[1] is a RoBERTa-based model that contains some interesting optimization such as SentencePiece and Whole Word Masking. UmBERTo has been shown to perform very well compared to other BERT-based models (Tamburini, 2020).

DistilBERT (Sanh et al., 2019) is a more compact language model pre-trained with knowledge distillation(Hinton et al., 2015), a technique that uses the output of a larger teacher network to train a smaller student network. BERTino (Muffo and Bertino, 2020) is an Italian DistilBERT model that was recently proposed as a lightweight alternative to BERT specifically for the Italian language.

ELECTRA is pre-trained with replaced token detection: instead of being masked, tokens are replaced with plausible alternatives sampled from a generator network; the model is then pre-trained to discriminate whether each token was replaced by a generator sample or not. At the outset of this study, the authors posited that replaced token detection is enough to make ELECTRA reasonably ready to tackle cloze tests with no further task-specific fine-tuning; this is indeed the case, as confirmed by the results shown in Table 1.

To summarize, we employ the following autoencoding models (all cased, as the cloze tests in our dataset contain case-sensitive options):

- multilingual BERT-base[2] (BERT multi), which serves as a baseline for autoencoding models;
- the Bayerische Staatsbibliothek's Italian BERT model[3] (BERT it);
- a smaller version of multilingual BERT-base[4] (BERT it LWYN) based on the Load What You Need concept described in (Abdaoui et al., 2020);
- UmBERTo[5] as the representative of the RoBERTa family.
- BERTino[6] as the representative of the DistilBERT family;

[1] https://github.com/musixmatchresearch/UmBERTo
[2] bert-base-multilingual-cased
[3] dbmdz/bert-base-italian-xxl-cased
[4] Geotrend/bert-base-it-cased
[5] Musixmatch/UmBERTo-commoncrawl-cased-v1
[6] indigo-ai/BERTino

- the Bayerische Staatsbibliothek's Italian ELECTRA model[1].

Autoregressive models. The key limitation of masked language modeling as a proxy for cloze test is the focus on single token masking. Therefore, autoencoding models are not applicable to the six cloze tests in our dataset that feature at least one multiple token option. (In some cases, the multiple token options are consistently among the incorrect options; using our autoenconding models in such cases would therefore skew the results in the models' favor.) For these tests, we employ a simple strategy based on autoregressive models: we iterate over all possible substitutions given the options offered by a test and choose the one with the lowest perplexity as determined by each of our autoregressive language models, all of which are from the GPT-2(Radford et al., 2019) family and include the following:

- a standard GPT-2 model[2], which serves as a performance lower bound (Vanilla GPT-2);
- a *recycled* version of GPT-2[3] transferred to the Italian language(de Vries and Nissim, 2020) (Recycled GPT-2);
- GePpeTto[4](Mattei et al., 2020), the first generative language model for Italian, also built using the GPT-2 architecture.

3 Results

The results of our study are summarized in Table 1. We report the results obtained by the human test takers and the models for each of the eleven cloze tests in our dataset as well as aggregates (mean values) over the whole dataset. For each cloze test, we report the number of blanks to be filled (*Questions*, which varies from 4 to 6), the number of human test takers (*Human count*), as well as with the mean and the standard deviation of the scores. Each test is identified by the initial of its topic (S=Science, L=Legal, G=Geometry, R=Reasoning, E=Education, H=History, T=Technology) along with a numeral to disambiguate multiple tests on the same topic. As previously mentioned, two tests are unrestricted (all the provided options can go anywhere in the text) and the others are restricted (there are specific option lists for each blank to be filled). As previously explained, six tests (L2, G2, E, H1, H2, T) contain at least one multi-token option and are only tackled with autoregressive models. On average, we observe that:

- humans do better than the best model (Electra) by eight percentage points;
- Electra, UmBERTo, and GePpeTto are the top three performers;
- Vanilla GPT-2 aside, BERT it LWYN comes in last and underperforms BERT it multilingual.

Averages, however, hide the enormous gap between restricted and unrestricted tests. We illustrate this gap in Table 2, which compares these two categories of tests model by model and also shows averages across autoencoding and autoregressive models (computed over the best models for each category, i.e., without BERT-base-it LWYN and BERT-base-multi for autoencoding models and without Vanilla GPT-2 for autoregressive models). This leads us to the following observations:

- our best autoencoding models outperform the human average;
- as expected, our models perform much better in restricted tests (we see a gap of 30 percentage points for autoencoding model and 10 points for autoregressive models);
- autoregressive models outperform autoenconding models in unrestricted tests, while the converse holds in restricted tests;
- humans perform similarly on both our restricted and unrestricted tests (and so does our performance lower-bound, Vanilla GPT-2).

In our restricted tests, UmBERTo and Electra outperform the human average and emerge as the top performers among our models. Though far below the human average, GePpeTto and Recycled GPT-2 are the two top performers in unrestricted tests, where none of the autoencoding model reach the pass threshold of 0.6. Vanilla GPT-2 aside, BERT it LWYN comes in last and underperforms BERT it multilingual in restricted tests while matching its baseline performance in unrestricted tests.

[1] dbmdz/electra-base-italian-xxl-cased-generator
[2] `https://huggingface.co/gpt2`
[3] GroNLP/gpt2-medium-italian-embeddings
[4] LorenzoDeMattei/GePpeTto

	S1	L1	G1	R	S2	L2	G2	E	H1	H2	T	Ave.
Restricted	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	
Questions	6	6	6	4	4	6	6	6	5	5	6	
Human count	253	300	154	130	291	253	390	300	316	157	184	
Human mean	0.77	0.88	0.82	0.84	0.96	0.89	0.96	0.88	0.9	0.55	0.87	0.85
Humans std	0.21	0.16	0.15	0.25	0.1	0.14	0.1	0.14	0.14	0.25	0.14	
UmBERTo	0.34	0.68	0.76	1	1	-	-	-	-	-	-	0.76
BERTino	0.34	0.76	0.67	0.76	1	-	-	-	-	-	-	0.71
Electra	0.34	0.84	0.67	1	1	-	-	-	-	-	-	0.77
BERT it	0.34	0.84	0.67	0.76	1	-	-	-	-	-	-	0.72
BERT it LWYN	0	0.68	0.33	0.76	0.76	-	-	-	-	-	-	0.51
BERT multi	0	0.68	0.33	1	0.76	-	-	-	-	-	-	0.55
GePpeTto	0.34	1	0.5	0.76	0.76	0.67	0.84	1	1	0.4	1	0.75
Recycled GPT-2	0.34	0.92	0.5	0.76	0.76	0.83	0.66	0.84	0.8	0.6	0.84	0.71
Vanilla GPT-2	0.16	0.5	0.17	0.5	0.5	0.5	0.5	0.16	0.4	0	0.16	0.32

Table 1: Performance of various autoencoding and autoregressive language models on 11 different Italian-language cloze tests on various topics (S=Science, L=Legal, G=Geometry, R=Reasoning, E=Education, H=History, T=Technology) and comparison to human performance (the number of students who took each of the tests in reported as *Human count* along with the sample mean and the standard deviation of the scores).

	Unres.	Res.
Humans	0.83	0.85
UmBERTo	0.51	**0.92**
BERTino	0.55	0.81
Electra	0.59	**0.89**
BERT-base-it	0.59	0.81
BERT-base-it LWYN	0.34	0.62
BERT-base-multi	0.34	0.70
Autoencoding ave.	0.56	**0.86**
GePpeTto	0.67	0.77
Recycled GPT-2	0.63	0.73
Vanilla GPT-2	0.33	0.32
Autoregressive ave.	0.65	0.75

Table 2: Aggregate data for unrestricted and restricted cloze tests. The autoencoding average is shown without BERT-base-it LWYN and BERT-base-multi, while the autoregressive average is shown without Vanilla GPT-2.

4 Case Studies

In this section, we focus on two specific examples of cloze tests from our dataset that serve as case studies to shed further light on our results. Let us consider the following restricted cloze test (G1 in Table 1).

> Nelle frasi seguenti, tratte da un libro di geometria, inserite le parole opportune per mezzo dei menu a discesa.
>
> *Dati due punti distinti A e B esiste una e una sola retta r tale che A e B appartengono [1] r. Invece di "A appartiene a r" possiamo scrivere "A giace [2] r" oppure A è un punto [3] r. Due rette complanari hanno o un punto o nessun punto [4] comune. [5] una retta e un punto che non giace [6] medesima, può essere fatto passare uno e un solo piano.*

The replacements are reported in Table 3 and show that this specific cloze test focuses solely on function words.

UmBERTo offers the best performance. UmBERTo's only mistake is at blank 5, where *Tra* is chosen instead of *Per*. We note that this is a typical mistake made by the students who took this cloze

blank	replacement
1	a, su, di, in, per
2	su, a, di, in, per
3	di, a, da, in, per
4	in, a, di, su, per
5	Per, A, Sopra, In, Tra
6	sulla, alla, della, dalla, tra

Table 3: Replacements for example 1.

test. The correct answer, *Per*, ranks second among UmBERTo's top picks, with a probability of approximately $2.9\ 10^{-3}$ as opposed to $3.3\ 10^{-2}$ for *Tra*. The second best models, BERTino, BERT-base, and ELECTRA-base, make an additional mistake at blank 2.

Let us now consider the following unrestricted cloze test (L1 in Table 1).

> *Ai fini della sicurezza della circolazione e della tutela della vita umana la velocità [1] non può superare i 130 km/h per le autostrade, i 110 km/h per le strade extraurbane principali, i 90 km/h per le strade extraurbane secondarie e per le strade extraurbane locali, e i 50 km/h per le strade nei centri abitati, con la possibilità di [2] il limite fino a un massimo di 70 km/h per le strade urbane le cui caratteristiche costruttive e funzionali lo consentano, [3] installazione degli appositi segnali. Sulle autostrade a tre corsie più corsia di emergenza per ogni senso di marcia, dotate di apparecchiature [4] omologate per il calcolo della velocità media di percorrenza su tratti determinati, gli enti proprietari o concessionari possono elevare il limite massimo di velocítà fino a 150 km/h sulla base delle caratteristiche progettuali ed effettive del tracciato, previa installazione degli appositi segnali, [5] lo consentano l'intensità del traffico, le condizioni atmosferiche prevalenti e i dati di incidentalità dell'ultimo [6]. In caso di precipitazioni atmosferiche di qualsiasi natura, la velocità massima non può superare i 110 km/h per le autostrade e i 90 km/h per le strade extraurbane principali.*

The replacements are reported in Table 4 and show that this specific cloze test focuses primarily on content words.

blank	replacement
1	massima
2	elevare
3	previa
4	debitamente
5	purché
6	quinquennio
incorrect	indebitamente, ridurre, finché, secolo, compresa, sebbene, giorno, poiché, esclusa, velocemente, dimezzare, minima

Table 4: Replacements for example 2.

Autoregressive models ace this test. GePpeTto offers the best performance (no incorrect replacements). Recycled GPT-2 is second best, with only one incorrect replacement out of 6: *giorno* is chosen instead of the correct token *quinquennio*. This replacement requires a level of contextual understanding that cannot be realistically expected from a language model at this point in time; our conjecture is that, in this specific instance, GePpeTto's correct replacement is most likely fortuitous (its performance range across all of our tests seems to validate our conjecture). Autoenconding models fare substantially worse, even though ELECTRA and BERT-base are fairly close to the average human performance.

5 Conclusion

While these results are based on as few as eleven cloze tests (and only two unrestricted ones), the key takeaway is that **existing pre-trained Italian language models with no task-specific fine-tuning can successfully tackle (and pass) relatively sophisticated tests** designed for Italian students who have successfully completed their high school education. These results, though preliminary in nature, suggest various research questions, which could be answered based on a larger set of cloze tests. Such questions include whether there exists a pattern to the incorrect replacements made by the models, how the models fare with different parts of speech and with function words as opposed to content words, and how much their performance would improve with task-specific fine-tuning.

References

Amine Abdaoui, Camille Pradel, and Grégoire Sigel. 2020. Load what you need: Smaller versions of mutlilingual bert. In *SustaiNLP / EMNLP*.

Isabella Chiari. 2002. La procedura cloze, la ridondanza e la valutazione della competenza della lingua italiana. *ITALICA*, 79:466–481.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Wietse de Vries and Malvina Nissim. 2020. As good as new. how to successfully recycle english gpt-2 to make models for other languages.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Edoardo Lugarini. 2010. Franco Angeli.

Lorenzo De Mattei, Michele Cafagna, F. Dell'Orletta, M. Nissim, and Marco Guerini. 2020. Geppetto carves italian into a language model. *ArXiv*, abs/2004.14253.

Matteo Muffo and E. Bertino. 2020. Bertino: An italian distilbert model. In *CLiC-it*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

F. W. Radice. 1978. Using the cloze procedure as a teaching technique. *ELT Journal*, XXXII.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

F. Tamburini. 2020. How "bertology" changed the state-of-the-art also for italian nlp. In *CLiC-it*.

Wilson L. Taylor. 1953. 'cloze' procedure: A new tool for measuring readability. *Journalism Quarterly*, 30:415–433.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.