The Self-Contained Italian Negation Test (SCIN)

Viola Gullace[1,2,3,†], David Kletz[1,4,*,†], Thierry Poibeau[1], Alessandro Lenci[2] and Pascal Amsili[1]

[1]Lattice, CNRS & ENS-PSl & U. Sorbonne-Nouvelle, 1 rue Maurice Arnoux F-92120 Montrouge, France
[2]CoLing Lab, Dipartimento di Filologia, Letteratura e Linguistica, Università di Pisa, Via Santa Maria, Pisa, 56126, Italy
[3]Scuola Normale Superiore, Piazza dei Cavalieri 7, Pisa, 56126, Italy
[4]LLF, CNRS & Université Paris Cité, 8 Rue Albert Einstein 75013 Paris, France

Abstract

Recent research has focused extensively on state-of-the-art pretrained language models, particularly those based on Transformer architectures, and how well they account for negation and other linguistic phenomena in various tasks. This study aims to evaluate the understanding of negation in Italian bert- and robert-based models, contrasting the predominant English-focused prior research. We develop the SCIN Set, an Italian dataset designed to model the influence of polarity constraints on models in a masked predictions task. Applying the SCIN Set reveals that these models do not adjust their behaviour based on sentences polarity, even when the resulting sentence is contradictory. We conclude that the tested models lack a clear understanding of how negation alters sentence meaning.

Keywords

negation, Italian PLMs, testing, self-contained

1. Introduction

Compositionality is a fundamental feature of human language, based on the principle that the meaning of a complex expression derives from its parts and their respective arrangements.

One notable compositional phenomenon is negation, formally defined as a semantic operator (or function) that reverses the truth-value of a sentence [1].

Given its importance, understanding how well pretrained language models (PLMs) can grasp and apply this principle is crucial.

These models achieve impressive performance across a wide array of language modeling tasks. Nonetheless, they often reveal to rely on shallow heuristics or exhibit other issues in handling specific aspects of language.

A prominent bias in the body of research is that the vast majority of research on language models has predominantly concentrated on English. This focus raises concerns about the generalizability of findings to other languages which may be structurally different from English. Conducting similar experiments in other languages could provide valuable context and material for comparison, potentially highlighting language-specific effects or revealing new generalization. Therefore, we decide to undertake a new experiment focusing on Italian negation.

Thus, in this article, we aim to explore whether the behavior of PLMs accurately models the polarity of sentences. We will investigate how the addition of negation to a sentence can alter its overall meaning (demonstrating the models' capability to handle shifts in meaning due to structural changes).

Given the limitations explained above, our work has deliberately chosen to concentrate on Italian. This choice not only addresses the need to explore how these models perform with languages other than English but also serves as a critical test for PLMs dedicated to Italian. We suspect that these models may not be as advanced or effective as their English counterparts, highlighting the need for further developments outside English.

We adapt the test set developed for English by Kletz et al. [2] to Italian, creating the *Self-Contained Italian Neg Set* (SCIN Set). Using the dataset to evaluate bert- and roberta-based models for Italian, we find that these models are unable to adjust their prediction in response to constraints posed by negation, often generating contradictory text.

The article will be structured as follows. The rest of Section 1 will introduce compositional phenomena and Italian negation in particular. Section 2 will briefly review related work. Section 3 will detail the composition of the SCIN Set. Section 4 will present the tests conducted on several bert-based Italian models using the SCIN Set; in particular, we tested the following bert-base-cased models:

- bert-base for Italian, both in its basic and its XXL versions (bert-base-italian-cased,

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
† These authors contributed equally.
viola.gullace@sns.it (V. Gullace);
david.kletz@sorbonne-nouvelle.fr (D. Kletz);
thierry.poibeau@ens.psl.eu (T. Poibeau); alessandro.lenci@unipi.it (A. Lenci); Pascal.Amsili@ens.fr (P. Amsili)
https://people.unipi.it/alessandro_lenci/ (A. Lenci);
https://lattice.cnrs.fr/amsili/ (P. Amsili)
0000-0003-3669-4051 (T. Poibeau); 0000-0001-5790-4308 (A. Lenci); 0000-0002-5901-5050 (P. Amsili)

bert-base-italian-xxl-cased)[1] [3],
- m-bert (multilingual bert)[2] [4],
- alb3rt0[3] [5], and
- UmBERTo[4][6].

Section 5 will discuss the results, followed by a final section containing our general conclusions and ideas for further research.

2. Related work

Although negation plays an essential role in human communication, it appears to present challenges for PLMs. In recent years, much research has focused on this topic.

2.1. Effect of negation on the model's prediction

Kassner and Schütze [7] and Ettinger [8] analyzed to what extent Transformer-based language models' predictions are sensitive to the presence or absence of negation in sentences involving factual knowledge, such as (1-a-b):

(1) a. Birds can [MASK].
b. Birds cannot [MASK].

They found that in such pairs the top-1 predictions are unchanged most of the time: models do not seem to take into account the polarity of the environment (presence or absence of a negation in the surrounding sentence) to adapt their predictions. They concluded that models do not deal correctly with negation.

Gubelmann and Handschuh [9] criticized such studies, noting in particular that the pragmatic component was overlooked in Ettinger's experiments. They noted that a statement containing a negation stating a false fact (for example, *Birds cannot fly*) can be more plausible than a formally true but unusual statement (say, *Birds cannot breastfeed*). In fact, a vast number of words could potentially fit the negative statement, making it true, many of them with little association with the rest of the sentence. This makes it challenging for any single word to become the top prediction in the negative case.

Gubelmann and Handschuh [9] developed a more pragmatically informed test set, in which each instance is (in [2]'s terms) *self-contained*. This means that each item in the set includes some context information, allowing direct evaluation of the model's completion. Building on this work, [2] developed the *Self-Contained Neg Test*, which aimed to address some issues in the test set from [9] and more accurately determine the model's handling of negation without interference of world knowledge.

[1] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased
[2] https://huggingface.co/bert-base-multilingual-cased
[3] https://github.com/marcopoli/AlBERTo-it
[4] https://github.com/musixmatchresearch/umberto

2.2. The *Self-Contained Neg Test*

The *Self-Contained Neg Test*, developed by Kletz et al. [2], is a set of pairs of sentences consisting of a context (C) and a target (T) sentence, either positive (p) or negative (n). The target sentence contains a masked position, syntactically constrained to be filled by a verb (2).

(2) Jessica is an architect who likes to dance. She isn't happy to [MASK].

The instances are designed in such a way that a model that predicts (in the masked position of T) the last verb of C will produce a semantically well-formed paragraph only if C and T have the same polarity. For instance, in (2), the context is positive (Cp), the target is negative (Tn), and as a consequence a model predicting *dance* in the masked position produces an ill-formed paragraph:

(3) #Jessica is an architect who likes to dance. She isn't happy to dance.

In contrast, a CnTn version of (3) would accept the verb *dance* in the same position:

(4) Jessica is an architect who doesn't like to dance. She isn't happy to dance.

To produce the sentences of the set, the pattern (5) is taken as a starting point, where NAME and PRON are substituted with a proper noun and a compatible third person pronoun, PRO is substituted with a profession name, and ACT is substituted with an action verb.

(5) NAME is a PROF who likes/doesn't like to ACT. PRON is/isn't happy to [MASK].

A large number of triplets (NAME, PRO, ACT) are tested with each model, and the ones that are retained are the ones such that the model's top one prediction is the ACT verb itself when C and T are both positive (CpTp). Here for instance, assuming that (6) are a model's predictions, the triplet (Jessica, architect, dance) would be retained while the triplet (Luke, janitor, swim) would not.

(6) a. Jessica is an architect who likes to dance. She is happy to dance.
b. Luke is a janitor who likes to swim. He is happy to ski.

Once triplets have been selected (the set of all triplets such that the ACT verb is repeated in CpTp instances), CpTn and CnTp instances can be formed, and the expectation is that a model that "understands" negation should not predict the ACT verb in those cases since it would lead to contradictory instances. As a control, two additional confirgurations are considered: CnTn where it is expected that the repetition of ACT is possible (though

not required), and CpTv in which an adverb (*very*) is inserted in the positive target, which should not change the preferred prediction of ACT since both sentences are positive. The different configurations are illustrated below.

(7)
CpTp	Jessica is an architect who likes to dance. She is happy to [MASK].
CpTn	Jessica is an architect who likes to dance. She isn't happy to [MASK].
CnTp	Jessica is an architect who doesn't like to dance. She is happy to [MASK].
CnTn	Jessica is an architect who doesn't like to dance. She isn't happy to [MASK].
CpTv	Jessica is an architect who likes to dance. She is very happy to [MASK].

3. SCIN construction

In Italian, negation is most commonly expressed by the negative invariable proclitic *non* (not) [10].

It is this expression of negation that we use for the Italian adaptation of the *Self-Contained Neg Test* that we present in this section: the SCIN set.

3.1. Italian patterns

Following the preparation of the *Self-Contained Neg Test*, we collect a list of Italian verbs, professions and names that will be used to create the triplets to be tested. The verbs are taken from the *Dizionario Italiano Sabatini Coletti* 2022 (online version); only the intransitive (3138 verbs) are retained; among these, for each of the tested models we further exclude the verbs that are not tokenized as a single token. The selected names are the 100 most popular in Italy in 2024[5]. Lastly, the professions are taken from a site specializing in job searches in Italy[6]; of those present on the site, only those consisting of a single word have been selected.

The patterns cannot simply be a direct translation of English patterns into Italian. In fact, for the test to be adequate for evaluating models, we need the masked position to be syntactically constrained to be a verb. This would not be the case if we used a direct translation of the original sentences: for example, the sequence (8) can be completed with the token "questo" (= *PRON is happy to do this*).

(8) NAME è un PROF che ama ACT. È felice di MASK.
NAME is a PROF who loves to ACT. (PRON) is happy to MASK.

[5]https://www.nostrofiglio.it/gravidanza/nomi-per-bambini/i-100-nomi-per-bambini-piu-amati-dai-genitori-di-nostrofiglio-it
[6]https://www.wecanjob.it/pagina9_elenco-professioni.html

We choose instead to rely on the pair (9), involving a semantic inference relation.

(9) ha l'abitudine di / molto spesso
is used to / very often

The final form of the SCIN set is available in table 1. The shape of the contexts is given in row 1, that of the targets in row 2, and the test target Tv is added in row 3.

Our assumption is that, if the model repeats the ACT token in the CpTp configuration, it is proof that the model has resolved the *ha l'abitudine di / molto spesso* inference. When confronted with the CpTn or CnTp configuration, the model should have the addition of the negation as the only element that can explain the modification of its predictions. Finally, the CpTv control allows us to check the extent to which the addition of a different, non-negative adverb in the sequence modifies the model's predictions; we can assume that any modification of greater magnitude than that associated to CpTv are due to the influence of negation.

The complete list of new patterns is available in Table 1.

3.2. Pattern selection

The triplets (*name*, *profession*, *verb*) used for testing are selected by testing them on the CpTp configuration: only triplets leading to a repetition of the ACT token are retained (see Table 2). This ensures that only patterns for which the model is already biased towards repetition are tested, and the model has to understand the influence of negation on sentence semantics to reverse this tendency.

All available triplets are tested, i.e. all configurations between verbs monotokenized by the model, first names and occupations selected in subsection 3.1. As tokenization is model-dependent, the number of verbs tested is not the same for each model: details are available in the first row of table 3.

The results of this test are available in table 3. The results are highly model-dependent: while the bert-base-italian-cased model predicts the ACT token in almost 25% of cases, this is the case in only 0.03% of cases for alb3rt0.

4. Testing

4.1. Setup

Tests are performed as in Kletz et al. [11]. Contexts (C) and targets (T) are combined to create two test patterns CpTn, CnTp; in addition to these two, the test includes two control patterns CnTn and CpTv where the repetition of the ACT verb is not contradictory.

All selected triplets are then used to saturate the patterns, and the resulting patterns are provided as inputs to

	pol.	C(ontext)	T(arget)
1	p	NAME è un(a) PROF che ha l'abitudine di ACT. *NAME is a PROF who is used to ACT-ing.*	PRON [MASK] molto spesso. *PRON [MASK] often.*
2	n	NAME è un(a) PROF che non ha l'abitudine di ACT. *NAME is a PROF who is not used to ACT-ing.*	PRON non [MASK] molto spesso. *PRON doesn't [MASK] often.*
3	v	-	PRON [MASK] davvero molto spesso. *PRON [MASK] really often.*

Table 1
Complete list of contexts and targets used to build masked sequences in the SCIN dataset. Masks are always in the target. Contexts and targets can be either positive or negative, and the target can also have an adverb added which is not a negation cue. Patterns are made up of a context and a target, i.e. 5 possible patterns.

Instantiated NAME/PROF: *Jessica / Ballerina (Dancer)*		
Tested verb: *Fumare (To smoke)*		
Tested example: *Jessica è una ballerina che ha l'abitudine di fumare. Lei [MASK] spesso.*		
Model	Top 1 pred.	Retained?
b-b-italian-c	*fuma*	✓
b-b-italian-xxl-c	*fuma*	✓
m-bert	*balla*	no
alb3rt0	*parla*	no

Table 2
An example of selecting a triplet for testing. A NAME/PROF/VERB triplet is used to saturate the CpTp pattern of SCIN. The sequence contains a mask and is used as input to a PLM. If the model prediction is the ACT token, the triplet is retained (indicated by the ✓ symbol). In the name of the models given as examples, "b-b" means bert-base, "it" stands for italian and "c" for cased.

the models. Predictions at masked positions are collected.

We use *drop* as a measure of the models' performance: for each pattern, given the rate t_r of repetitions of the Act Token in the predictions, the drop is defined as $100 - t_r$. The higher the drop for the CpTn and CnTp patterns and the lower for the CnTn and CpTv controls, the better the model has understood the negation.

4.2. Results and Discussion

Results are shown in table 4.

In contrast with the observations made by [8] and [7], the models are not insensitive to the presence of negation in a sentence: all the models show a drop in both configurations CpTn and CnTp, showing an adaptation of their predictions to the presence of a negation cue. This observation is confirmed by the fact that the drops in the CpTv control are always lower than those observed in CpTn or CnTp.

This shows that simply adding an adverb is not sufficient to change the model's predictions. While we cannot definitively attribute this to its logical function, the negation marker does exert a distinct influence.

Nevertheless, it is important to emphasize the very clear limitations of these results. Firstly, the drops never exceed 25%, meaning that 75% of the times the model predicts a semantically prohibited token. On the other hand, with the exception of m-bert, all the models have a highe drop for the CnTn control than for the CnTp configuration, thus indicating that even though the models have acquired a certain understanding of negation, this remains superficial and does not, for example, clearly include an understanding of the positive value of a double negation.

A broader examination of the results reveals that while the drops in CpTn and CnTp configurations increase together, the CnTn controls also show a corresponding increase.

Finally, the training corpus of the models seems to have an influence on their performance. For example, note that the alb3rt0 model is the model obtaining the results least in line with our expectations, while bert-base-italian-xxl-cased and bert-base-italian-cased had better drop values, with the former performing better than the latter. However, these three models have identical numbers of layers, attention heads and hidden sizes, the difference between them only consisting in their training data. The alb3rt0 model was trained exclusively on tweets, which likely limits the diversity of its data, particularly with respect negation. In contrast, bert-base-italian-cased and bert-base-italian-xxl-cased models were trained on more varied corpora, with the latter featuring a larger dataset.

In the future, this should lead us to study the correlation between the performance of the models and the fine-grained distribution of negative and affirmative contexts in their training corpus.

5. Comparison with English

In this section we compare the results obtained with the SCIN Set with those observed by [2] in English.

Model	b-b-it-c	b-b-it-xxl	m-bert	alb3rt0	UmBERTo
# tested contexts	5880000	5880000	780000	18800000	280000
Repetitions	1498456	1236899	141609	5464	93284
%	25.48	21.03	18.16	0.03	33.31
# retained contexts	20000	20000	19973	2088	20000

Table 3
Details of the verb sets created for each model. The first line shows the number of triples available per model, the second the number of these triples which, in a CpTp configuration, led to a repetition (prediction by the ACT token model), and line 3 the percentage of triples this represents.) The last line shows how many of the triplets leading to a repeat were retained, the maximum for one model being 20,000. In the column titles, "b-b" means bert-base, "it" stands for italian and "c" for cased

Pattern	b-b-it-c	b-b-it-xxl	m-bert	alb3rt0	UmBERTo
CpTn	16.5	**22.1**	23.0	9.7	9.9
CnTp	11.0	14.5	**19.7**	4.4	11.9
CnTn	11.6	14.6	18.6	**9.3**	20.6
CpTv	1.3	14.3	1.0	**0.2**	1.7

Table 4
Drops of Italian pretrained language models on the SCIN Set, for each pattern type. In the two first rows, a high number is expected — the higher number of each row in bold face; in the two last rows, a lower number is expected. In the column titles "b-b" means bert-base, "it" stands for italian and "c" for cased

The scale of the drops in the two articles is notably very different: the maximum drop observed in Italian is 23% (CpTn m-bert), while in English it's 82.8%. Similarly, the CpTv drops of Italian-speaking models hardly exceed 15%, while those of English-speaking models are never less than 25%.

On the other hand, model architecture and type of training do not seem to have a major influence: Umberto has the same architecture as roberta-base, but while the latter is the best performing model in [2], the former's drops are the lowest for all configurations of the SCIN Set. Conversely, the other Italian models are built with the same architecture as bert-base-cased, i.e. the worst performing model for English; however, even the worst performing Italian model, namely alb3rt0, features higher drops than bert-base-cased. This confirms the observation from the previous section, that while architecture is indeed a limiting criterion, training data probably plays a significant role.

In general, we note that none of these models, neither for Italian nor for English, shows definitive drops compatible with a full understanding of the semantic constraints of negation.

6. Conclusion

In this paper, we investigated the ability of several Italian PLMs to take negation into account in their predictions. To do this, we adapted to Italian the *Self-Contained Neg Test* proposed by Kletz et al. [2], which is based on minimal pairs of aligned sentences.

Applying this test to six models enabled us to show that negation modifies their predictions, but that this does not happen consistently or in a way that is always coherent with the semantic effect that we expect negation to have on sentences. These results suggest a strong need to adapt these models to make them more sensitive to negation and its semantic consequences.

Nevertheless, we also noted a fairly marked difference in performance from one model to another, correlated with the different corpora used to train them. We thus suggest that a lexical and statistical study of these corpora could shed further light on the behavior of the models.

Lastly, it would be interesting to compare these results with the performance of generative models, in order to study the relative importance of the number of model parameters in relation to their architecture.

Acknowledgments

We would like to express our gratitude to Marie Candito for her valuable assistance and guidance throughout the course of this study.

This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA0001 (PRAIRIE 3IA Institute). This research was also partially funded by the Labex EFL (ANR-10-LABX-0083) and by PNRR–M4C2–Investimento 1.3, Partenariato Esteso PE00000013––"FAIR—Future Artificial Intelligence Research"–Spoke 1 "Human-centered AI," funded by the European Commission under the NextGeneration EU programme.

References

[1] L. R. Horn, H. Wansing, Negation, in: E. N. Zalta, U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy, Winter 2022 ed., Metaphysics Research Lab, Stanford University, 2022.

[2] D. Kletz, P. Amsili, M. Candito, The self-contained negation test set, in: Y. Belinkov, S. Hao, J. Jumelet, N. Kim, A. McCarthy, H. Mohebbi (Eds.), Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Singapore, 2023, pp. 212–221. URL: https://aclanthology.org/2023.blackboxnlp-1.16. doi:10.18653/v1/2023.blackboxnlp-1.16.

[3] S. Schweter, Italian bert and electra models, 2020. URL: https://doi.org/10.5281/zenodo.4263142. doi:10.5281/zenodo.4263142.

[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[5] M. Polignano, V. Basile, P. Basile, M. de Gemmis, G. Semeraro, AlBERTo: Modeling italian social media language with bert, IJCoL 25 (1984) 11–31. URL: https://doi.org/10.4000/ijcol.472.

[6] L. Parisi, S. Francia, P. Magnani, Umberto: an italian language model trained with whole word masking, https://github.com/musixmatchresearch/umberto, 2020.

[7] N. Kassner, H. Schütze, Negated and misprimed probes for pretrained language models: Birds can talk, but can+not fly (2020). URL: https://aclanthology.org/2020.acl-main.698.

[8] A. Ettinger, What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models, Transactions of the Association for Computational Linguistics 8 (2019) 34–48. URL: https://doi.org/10.1162/tacl_a_00298.

[9] R. Gubelmann, S. Handschuh, Context matters: A pragmatic study of PLMs' negation understanding, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, p. 4602–4621. URL: https://aclanthology.org/2022.acl-long.315.

[10] L. Renzi, L. G. Salvi, A. Cardinaletti, Grande grammatica italiana di consultazione, volume 2, Il Mulino, Bologna, 2001.

[11] D. Kletz, M. Candito, P. Amsili, Probing structural constraints of negation in pretrained language models, in: The 24rd Nordic Conference on Computational Linguistics, 2023. URL: https://openreview.net/forum?id=_7VPETQwnPX.

A. Verb statistics by PLM

Details of the number of monotokenised intransitive verbs available for each PLM tested are available in table 5.

model	monotokenized verbs
bert-base-italian-cased	294
bert-base-italian-xxl-cased	294
m-bert	39
alb3rt0	940
UmBERTo	14

Table 5
Detail of the number of Italian intransitive verbs tokenised as a single token for each of the Italian models tested.