How About Time?
Probing a Multilingual Language Model for Temporal Relations

Tommaso Caselli[♥], **Irene Dini**[♣], **Felice Dell'Orletta**[♣]
[♥]CLCG, University of Groningen [♣]ItaliaNLP Lab, ILC-CNR "Antonio Zampolli" Pisa
t.caselli@rug.nl, {irene.dini, felice.dellorletta}@ilc.cnr.it

Abstract

This paper presents a comprehensive set of probing experiments using a multilingual language model, XLM-R, for temporal relation classification between events in four languages. Results show an advantage of contextualized embeddings over static ones and a detrimental role of sentence level embeddings. While obtaining competitive results against state-of-the-art systems, our probes indicate a lack of suitable encoded information to properly address this task.

1 Introduction

Time is a pervasive element of human life with no counterpart in any other cognitive domains (Bonomi and Zucchi, 2001). Such pervasiveness is mirrored in natural languages through sets of devices that allows speakers to refer to time, to reason about time and things that unfold in time. Reasoning about time is one of the central components of common sense knowledge (Pianesi and Varzi, 1996; Boyd, 2010; Geva et al., 2021) and its modeling has been at the core of many early approaches in Computational Linguistics and Artificial Intelligence (Schank and Abelson, 1975; McDermott, 1982; Allen, 1984; Passonneau, 1988; Moens and Steedman, 1988). More recently, specific Natural Language Understanding (NLU) tasks related to time have been developed, ranging from the identification of temporal expressions (Mani et al., 2001; Mazur and Dale, 2010), to measuring the duration of events (Pan et al., 2006b,a; Zhou et al., 2019), and the ability to order them chronologically(Mani et al., 2003; UzZaman and Allen, 2010; Ning et al., 2018; Wen et al., 2021). More complex tasks have challenged models to extract storylines (Chambers and Jurafsky, 2008; Minard et al., 2015; Caselli and Inel, 2018), understand narratives (Mostafazadeh et al., 2017, 2020; Lal et al., 2021), and answer temporally related questions (Llorens et al., 2015; Ning et al., 2020). Recent work has focused on recasting temporal relation classification as a Natural Language Inference (NLI) task where fine-tuned pre-trained language models (PTLMs) have achieved good results (Vashishtha et al., 2020).

Embedding representations, both static and contextual, have shown to play a key role to improve systems' results on different time-related benchmarks, especially for the classification of temporal relations between pairs of events (Mirza and Tonelli, 2016; Cheng et al., 2020). When it comes to contextualized embeddings the probing of such models for temporal knowledge has not been properly investigated yet. If we embrace the vision of PTLMs as large repositories of linguistic knowledge (Derby et al., 2021; Mosbach et al., 2020; Miaschi et al., 2020), it is a natural question to probe these models for their knowledge about events and time. We present an extensive study on temporal relation probing of PTLMs using five temporally annotated corpora in four languages (i.e., English, French, Spanish, and Italian). Although the selected languages all belong to Indo-European family, they present differences for the tense-mood-aspect (TMA) system while showing similarities at the lexico-pragmatic level. Our probing tasks focus on *temporal ordering of pairs of events* **(E–E)**, either in the same sentence or in different ones.

Our contributions Our work has three contributions: (i) it is the first work to probe a multilingual PTLM, XLM-R base, for temporal knowledge between event pairs; (ii) we study the impact of multilingual contextualized representations against monolingual counterparts based on static word embeddings; (iii) we compare zero-shot PTLM against fine-tuned models for temporal reasoning to investigate whether the models have acquired real temporal knowledge. Code and data are available.[1]

[1] https://github.com/irenedini/tlink_probing

Proceedings of the 29th International Conference on Computational Linguistics, pages 3197–3209
October 12–17, 2022.

2 Data Overview

We have selected five corpora annotated with language specific adaptations of ISO-TimeML (Pustejovsky et al., 2010). ISO-TimeML is, at the same time, an annotation meta-model for marking events, temporal expressions, and relations between them, and a full-fledged annotation language. ISO-TimeML has 13 values used to classify temporal relations, based on Allen's interval temporal logic (Allen, 1983) where each value expresses how an event chronologically relates to another event or a temporal expression. In the following paragraphs we present a short overview of the five corpora we have used. For our experiments, we have extracted all temporal relations between event pairs, either occurring in the same sentence or in difference sentences. Table 1 presents a summary of the temporal relations between events for each corpus.

EN-TimeBank The English TimeBank (Pustejovsky et al., 2003) is a corpus of 183 documents manually annotated following the TimeML annotation guidelines (Saurı et al., 2006). The whole corpus has gone through a curation phase for the SemEval 2013 TempEval-3 task (UzZaman et al., 2013), where an extra test set of 20 documents has been annotated with the same guidelines. In our experiments, we follow the TempEval-3 split for training and test distributions, excluding the automatically annotated data (i.e., silver data distribution). EN-TimeBank uses the full 13 temporal values from ISO-TimeML to classify event-event relations.

IT-TimeBank The Italian TimeBank (Caselli et al., 2011) has 254 documents, comparable in size and annotation to the EN-TimeBank. We have followed the official split into train and test from the EVALITA 2014 EVENTI task (Caselli et al., 2014). Similarly to EN-TimeBank, the 13 fine-grained temporal values have been used to classify temporal relations.

FR-TimeBank French TimeBank (Bittar et al., 2011) is a corpus of 107 documents in French annotated following an adaptation to French of TimeML. The corpus does not present an official split into train and test. To obviate to this, we have first extracted all temporally annotated event pairs and then created a train and test distribution following a 75-25 split. FR-TimeBank also uses the full 13 temporal values for classifying temporal relations

EN-TB-Dense The TimeBank-Dense corpus (Cassidy et al., 2014) contains only 36 documents from the training portion of the EN-TimeBank. EN-TB-Dense approximates a complete graph of all possible temporal relations over events and temporal expressions by labeling all pairs locally, i.e., same sentence and adjacent sentence pairs. EN-TB-Dense simplifies the set of possible temporal relation values by reducing it to five and introducing a new value, VAGUE, for all relations that do not carry a clear semantics.

ES-TimeBank The Spanish TimeBank (Saurı and Badia, 2012) contains 210 documents in Spanish. We have followed the official release into train and test splits. Similarly to the EN-TimeBank-Dense, the authors have simplified the set of possible temporal relations to five plus VAGUE. However, the overlap is limited only to BEFORE AFTER, with the other three being new.

3 Temporal Probing

Our probing task investigates the capabilities of PTLMs to encode information about events and their temporal ordering. To probe such information across multiple languages, we use `XLM-R base` (Conneau et al., 2020),[2] a large multilingual model that has achieved state-of-the-art results on many NLU tasks. Following previous work (Tenney et al., 2019; Jawahar et al., 2019; Vulić et al., 2020; de Vries et al., 2020, *inter alia*), we extract embedding representations from each layer and use them to train a linear SVM whose objective is to predict the value of a temporal relation between a given pair of events. By default, we feed the SVM with four concatenated embeddings: the embeddings of the sentence containing each event in the pair and those of each event. In case the event pair occurs in the same sentence, we duplicate the sentence representation. Sentences are represented by averaging the embeddings of the tokens excluding `XLM-R base`'s special tokens.

We compare the default settings with three variations: (i) we use only the embeddings of the events in the pair; (ii) we use the embeddings from two `XLM-R base` models previously fine-tuned with the EN-TimeBank and EN-TB-Dense

[2] https://huggingface.co/xlm-roberta-base

Temporal Relation	EN-TimeBank Train	EN-TimeBank Test	IT-TimeBank Train	IT-TimeBank Test	FR-TimeBank Train	FR-TimeBank Test	EN-TB-Dense Train	EN-TB-Dense Test	ES-TimeBank Train	ES-TimeBank Test
BEFORE	180	83	167	38	107	32	884	378	834	62
AFTER	184	90	79	15	36	10	729	275	499	47
INCLUDES	67	15	47	19	2	3	207	57	–	–
IS_INCLUDED	64	29	46	28	2	1	265	52	–	–
DURING	4	0	0	0	125	30	–	–	–	–
SIMULTANEOUS	132	45	131	46	26	9	72	22	–	–
IMM_BEFORE	11	1	2	1	6	1	–	–	–	–
IMM_AFTER	5	1	3	0	0	1	–	–	–	–
BEGINS	11	0	0	0	15	5	–	–	–	–
BEGUN_BY	11	0	1	0	5	1	–	–	–	–
ENDS	4	1	0	1	3	2	–	–	–	–
ENDED_BY	21	0	2	0	7	1	–	–	–	–
IDENTITY	140	15	217	50	78	42	–	–	–	–
OVERLAP	–	–	–	–	–	–	–	–	4,478	307
BEFORE_OVERLAP	–	–	–	–	–	–	–	–	907	74
OVERLAP_AFTER	–	–	–	–	–	–	–	–	336	26
VAGUE	–	–	–	–	–	–	1,995	634	29	5

Table 1: Summary of the distribution of the temporal relations between pairs of events in all five corpora. For the EN-TB-Dense, the values for the training are obtained by merging together the training and the development sets.

corpora recasted in forms of Natural Language Inference pairs for temporal reasoning as described in Vashishtha et al. (2020); (iii) we use monolingual static word embeddings obtained with the `word2vec skip-gram` (w2v) model (Mikolov et al., 2013) (see Appendix A for details). Lastly, all probing variations are compared with a dummy classifier predicting the majority class in each corpus.

4 Results

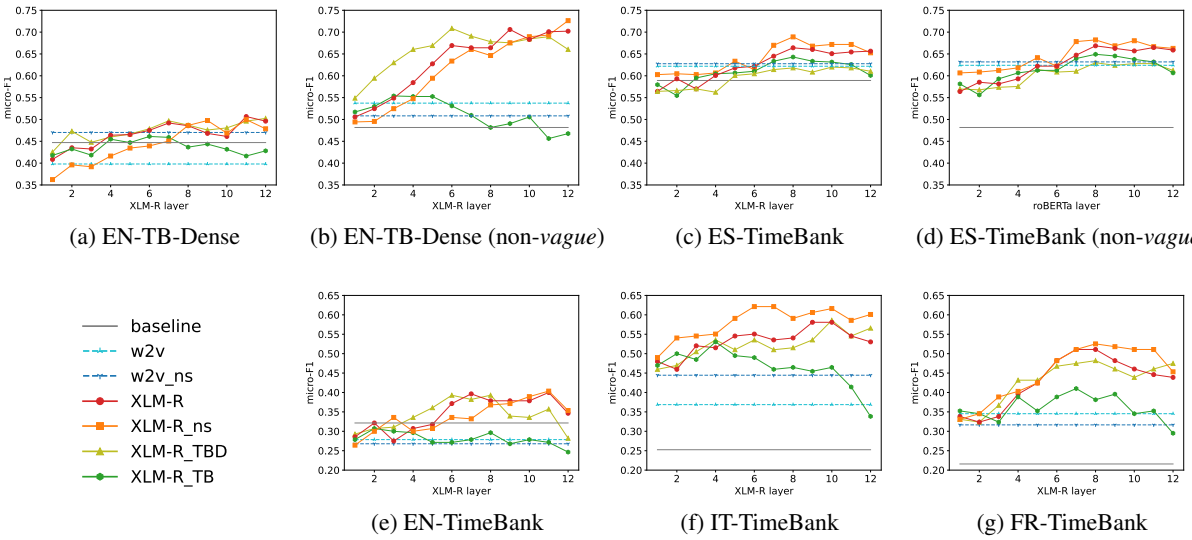
Figure 1 summarizes the results for all corpora and settings. Details per corpus are in Appendix B.

Although all probing models outperform their respective baselines, our results further confirm that temporal relation classification is a challenging task. `XLM-R` embeddings consistently obtain the best results across all languages and granularities of the temporal relations, improving the performance of static embeddings. With the exclusion of EN-TB-Dense, the presence of sentence embeddings is detrimental, confirming previous findings (Miaschi and Dell'Orletta, 2020). Although temporal relations are a discourse phenomenon at the interface of the semantics and pragmatics dimensions, it appears that the event only embeddings from `XLM-R` already store sufficient semantic information to perform this task.

Regardless of the granularity of the temporal relations, it clearly emerges from all the plots that the best results for `XLM-R` are obtained between layers 6 and 8. Performances are consistently sub-optimal for early layers, especially 1–4. For higher layers, i.e., 10–12, results are disappointing, with the exception for English whose best probe is at layer 11.Given the task and previous findings on the encoding of linguistic knowledge in PTLMs (Tenney et al., 2019; Jawahar et al., 2019), this is not fully expected. Ideally, if PTLMs tend to encoded more semantic features in the top layers, performances for this task should not degrade on the top layers, as we see for Spanish, Italian, and French, or, at least, they should remain on a plateau.

A further finding concerns the role of fine-tuned models for temporal reasoning, namely `XLM-R_tbd` and `XLM-R_tb`. The models have been fined-tuned using the English corpora EN-TB-Dense and EN-TimeBank recasted for temporal reasoning. We expected the embeddings from these models to be more competitive than basic `XLM-R`, but this is not the case. In general, we observe a better performance for `XLM-R_tbd` than `XLM-R_tb`, in-line with the results reported by Vashishtha et al. (2020). The better results of `XLM-R_tbd` hold also in cross-lingual settings, regardless of the granularity of the temporal values used in the specific corpus.

When comparing results across corpora, two dimensions are at play: the first is the granularity of the temporal values; the second is the number of training examples. A general pattern we observe is the following: the less temporal values are to be learned, the better the results of a trained model, provided that the annotated data are consis-



(a) EN-TB-Dense (b) EN-TB-Dense (non-*vague*) (c) ES-TimeBank (d) ES-TimeBank (non-*vague*)

(e) EN-TimeBank (f) IT-TimeBank (g) FR-TimeBank

Figure 1: Overview of the probing results across all corpora. For each plot, the *x*-axis reports the layer id of `XLM-R base`, the *y*-axis reports the micro-F1. In the legend on the bottom left side, `ns` stands for "no sentence"; `tbd` and `tb` refer to the fine-tuned `XLM-R base` models for temporal reasoning: `tbd` stands for EN-TB-Dense, `tb` stands for EN-TimeBank.

tent. To better illustrate this, we focus our analysis on the EN-TB-Dense and ES-TimeBank first. Both corpora adopt coarse grained temporal values and have the largest number of annotated data. Nevertheless, the way the value VAGUE is used in the two corpora is not the same. In EN-TB-Dense, given the specific annotation framework, VAGUE is used both in case of an existing temporal relation with an unclear semantics but also for event pairs with no temporal relation. This is not the case in Spanish. Such a difference is mirrored in the results: as soon as we remove VAGUE, scores in the EN-TB-Dense improve while they remain the same in ES-TimeBank. When considering the corpora with fine-grained temporal values, we observe that the F1-score in EN-TimeBank is the lowest (in absolute terms), while results are much better for IT-TimeBank and FR-TimeBank. The differences in this case can only be due to a more consistent application of the annotation guidelines in Italian and French than in English. Support to this claim can be found in the fact that Italian and French have a lower number of sentences in training, 695 and 412 respectively, than English, namely 834. To gain insights, we have analyzed the overlapping events between Train and Test splits, i.e., how many times the same event appears in Train and Test, even if coupled with a different event and with a different temporal value. While FR-TimeBank has the largest overlap (58%), IT-TimeBank has the lowest (29%) and EN-TimeBank is in the middle (35%). If it was just a matter of data, we would expect the EN-TimeBank to obtain better F1-scores than IT-TimeBank.

Comparison with state of the art is limited to EN-TB-Dense and IT-TimeBank. No previous work for this task is available FR-TimeBank and ES-TimeBank, and for EN-TimeBank we only have access to systems which classified temporal relations from raw text. Concerning the EN-TB-Dense, the best system, SECT (Cheng et al., 2020), adopts a multi-task learning approach using a GRU architecture. On the **E–E** classification it achieves an F1-score 0.650, gaining 0.098 points with respect to our best training layer. A more similar architecture, CATENA (Mirza and Tonelli, 2016) a linear SVM combining pre-trained word embeddings and additional features, obtains an F1-score of 0.519, only 0.012 points above us. As for Italian, the best system, FBK-HTL-time (Mirza and Minard, 2014), a feature-based linear SVM, achieves an F1-score of 0.688, beating our approach of 0.062 points.

Statistical significance testing We also performed statistical significance tests across all the probing systems using the McNemar's test. We ran the significant tests using two different settings: first by considering the last embedding layer of the PTLMs and subsequently the embedding layer that gave the best results for each probing. Details can found in Appendix C.

All probing experiments are consistently significant when compared to their respective baselines, with the exclusion of the EN-TimeBank corpus.

When focusing on the differences between the PTLM embeddings and the static ones, the results are more scattered, with different behaviors across each dataset. We observe significant differences in the majority of cases when sentence representations are excluded from the static embeddings. Two datasets, ES-TimeBank and IT-TimeBank, present peculiar behaviors when compared to the others. For the ES-TimeBank, probes with PTLM embeddings tend to be not statistically significant with respect to the static ones. The opposite trend, on the contrary, can be observed for the IT-TimeBank.

Finally, across all the PTLM probes, a clear tendency that emerges is that significant differences can be observed only when using `XML-R_tb` embeddings, while only in few cases the significant difference can be observed when using the `XML-R_tbd` embeddings.

5 Conclusion

This paper investigates the knowledge encoded in a large multilingual PTLM, `XLM-R base`, for temporal relation classification between pairs of events in four languages and five corpora with varying granularities of temporal values. Our results point out that temporal relation classification between events is very challenging and the linguistic knowledge in `XLM-R` is limited to properly address it. While contextual embeddings are more "powerful" than static ones, current fine-tuned models for temporal reasoning (Vashishtha et al., 2020) are not helpful as one would expect. Our probes indicate that adding more information, i.e., sentence representations, to lexical entities is detrimental, meaning that "global" semantic information is already encoded at the lexical level. Finally, our models are competitive with state-of-the-art systems, indicating that improvements are due either to specific architectures or extra features capturing additional knowledge not available in the contextual embeddings.

Acknowledgements

The authors acknowledge the project "Human in Neural Language Models" (*IsC93_HiNLM*), funded by CINECA[3] under the ISCRA initiative and the Peregrine High Performance Cluster of the University of Groningen for the availability of computing resources and support.

[3]https://www.cineca.it/en

References

James F. Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

James F Allen. 1984. Towards a general theory of action and time. *Artificial intelligence*, 23(2):123–154.

André Bittar, Pascal Amsili, Pascal Denis, and Laurence Danlos. 2011. French TimeBank: An ISO-TimeML annotated reference corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 130–134, Portland, Oregon, USA. Association for Computational Linguistics.

Andrea Bonomi and Alessandro Zucchi. 2001. *Tempo e linguaggio: introduzione alla semantica del tempo e dell'aspetto verbale*. Pearson Italia Spa.

Brian Boyd. 2010. *On the origin of stories: Evolution, cognition, and fiction*. Harvard University Press.

Tommaso Caselli, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta, and Irina Prodanof. 2011. Annotating events, temporal expressions and relations in Italian: the it-timeml experience for the ita-TimeBank. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 143–151, Portland, Oregon, USA. Association for Computational Linguistics.

Tommaso Caselli and Oana Inel. 2018. Crowdsourcing StoryLines: Harnessing the crowd for causal relation annotation. In *Proceedings of the Workshop Events and Stories in the News 2018*, pages 44–54, Santa Fe, New Mexico, U.S.A. Association for Computational Linguistics.

Tommaso Caselli, Rachele Sprugnoli, Manuela Speranza, and Monica Monachini. 2014. Eventi: Evaluation of events and temporal information at evalita 2014. *EVENTI: EValuation of Events and Temporal INformation at Evalita 2014*, pages 27–34.

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506, Baltimore, Maryland. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Fei Cheng, Masayuki Asahara, Ichiro Kobayashi, and Sadao Kurohashi. 2020. Dynamically updating event

representations for temporal relation classification with multi-category learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1352–1357, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Wietse de Vries, Andreas van Cranenburgh, and Malvina Nissim. 2020. What's so special about BERT's layers? a closer look at the NLP pipeline in monolingual and multilingual models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4339–4350, Online. Association for Computational Linguistics.

Steven Derby, Paul Miller, and Barry Devereux. 2021. Representation and pre-activation of lexical-semantic knowledge in neural language models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 211–221, Online. Association for Computational Linguistics.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Yash Kumar Lal, Nathanael Chambers, Raymond Mooney, and Niranjan Balasubramanian. 2021. TellMeWhy: A dataset for answering why-questions in narratives. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 596–610, Online. Association for Computational Linguistics.

Hector Llorens, Nathanael Chambers, Naushad UzZaman, Nasrin Mostafazadeh, James Allen, and James Pustejovsky. 2015. SemEval-2015 task 5: QA TempEval - evaluating temporal information understanding with question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 792–800, Denver, Colorado. Association for Computational Linguistics.

Inderjeet Mani, Barry Schiffman, and Jianping Zhang. 2003. Inferring temporal ordering of events in news. In *Companion Volume of the Proceedings of HLT-NAACL 2003 - Short Papers*, pages 55–57.

Inderjeet Mani, George Wilson, Lisa Ferro, and Beth Sundheim. 2001. Guidelines for annotating temporal information. In *Proceedings of the First International Conference on Human Language Technology Research*.

Pawel Mazur and Robert Dale. 2010. WikiWars: A new corpus for research on temporal expressions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 913–922, Cambridge, MA. Association for Computational Linguistics.

Drew McDermott. 1982. A temporal logic for reasoning about processes and plans. *Cognitive science*, 6(2):101–155.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2020. Linguistic profiling of a neural language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 745–756, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alessio Miaschi and Felice Dell'Orletta. 2020. Contextual and non-contextual word embeddings: an in-depth linguistic investigation. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 110–119, Online. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Anne-Lyse Minard, Manuela Speranza, Eneko Agirre, Itziar Aldabe, Marieke van Erp, Bernardo Magnini, German Rigau, and Rubén Urizar. 2015. SemEval-2015 task 4: TimeLine: Cross-document event ordering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 778–786, Denver, Colorado. Association for Computational Linguistics.

Paramita Mirza and Anne-Lyse Minard. 2014. Fbk-hlt-time: a complete italian temporalprocessing system for eventi-evalita 2014. *FBK-HLT-time: a complete Italian TemporalProcessing system for EVENTI-Evalita 2014*, pages 44–49.

Paramita Mirza and Sara Tonelli. 2016. On the contribution of word embeddings to temporal relation classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2818–2828, Osaka, Japan. The COLING 2016 Organizing Committee.

Marc Moens and Mark Steedman. 1988. Temporal ontology and temporal reference. *Computational Linguistics*, 14(2):15–28.

Marius Mosbach, Stefania Degaetano-Ortlieb, Marie-Pauline Krielke, Badr M. Abdullah, and Dietrich Klakow. 2020. A closer look at linguistic knowledge in masked language models: The case of relative clauses in American English. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 771–787, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: GeneraLized and COntextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586, Online. Association for Computational Linguistics.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. LSDSem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51, Valencia, Spain. Association for Computational Linguistics.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172, Online. Association for Computational Linguistics.

Qiang Ning, Hao Wu, and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328, Melbourne, Australia. Association for Computational Linguistics.

Feng Pan, Rutu Mulkar, and Jerry R. Hobbs. 2006a. Extending TimeML with typical durations of events. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 38–45, Sydney, Australia. Association for Computational Linguistics.

Feng Pan, Rutu Mulkar, and Jerry R. Hobbs. 2006b. Learning event durations from event descriptions. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 393–400, Sydney, Australia. Association for Computational Linguistics.

Rebecca J. Passonneau. 1988. A computational model of the semantics of tense and aspect. *Computational Linguistics*, 14(2):44–60.

Fabio Pianesi and Achille C Varzi. 1996. Events, topology and temporal relations. *The monist*, 79(1):89–116.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An international standard for semantic annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Roser Saurı and Toni Badia. 2012. Spanish TimeBank 1.0 Corpus documentation.

Roser Saurı, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky. 2006. TimeML Annotation Guidelines version 1.2.1.

Roger C Schank and Robert P Abelson. 1975. Scripts, plans, and knowledge. In *IJCAI*, volume 75, pages 151–157.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Naushad UzZaman and James Allen. 2010. TRIPS and TRIOS system for TempEval-2: Extracting temporal information from text. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 276–283, Uppsala, Sweden. Association for Computational Linguistics.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA. Association for Computational Linguistics.

Siddharth Vashishtha, Adam Poliak, Yash Kumar Lal, Benjamin Van Durme, and Aaron Steven White. 2020. Temporal reasoning in natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4070–4078, Online. Association for Computational Linguistics.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Haoyang Wen, Yanru Qu, Heng Ji, Qiang Ning, Jiawei Han, Avi Sil, Hanghang Tong, and Dan Roth. 2021. Event time extraction and propagation via graph attention networks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 62–73, Online. Association for Computational Linguistics.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.

A Monolingual Static Embeddings

All monolingual static word embeddings have been taken from this repository: https://vectors.nlpl.eu/repository/ . We have used the versions from the 2017 CoNLL shared task.

B Probes Models: Detailed Results

The following Tables presents detail results for each corpus from our experiments. We have highlighted in green the best results for each corpus across all models. For each probe model, we have highlighted in bold the best results. Scores correspond to micro-F1.

Model	Layer score 1	2	3	4	5	6	7	8	9	10	11	12	Score
XLM-R	0.409	0.435	0.432	0.464	0.465	0.475	0.492	0.486	0.468	0.461	**0.507**	0.496	–
XLM-R_NS	0.362	0.396	0.392	0.416	0.434	0.439	0.451	0.486	0.498	0.469	**0.500**	0.479	–
XLM-R_TBD	0.425	0.473	0.448	0.46	0.467	0.479	0.497	0.487	0.476	0.481	0.496	**0.502**	–
XLM-R_TB	0.418	0.432	0.418	0.455	0.447	**0.461**	0.459	0.437	0.444	0.432	0.416	0.428	–
W2V						–							0.398
W2V_NS						–							0.470
BASELINE						–							0.447

Table B.1: Results on EN-TB-Dense

Model	Layer score 1	2	3	4	5	6	7	8	9	10	11	12	Score
XLM-R	0.564	0.593	0.57	0.601	0.618	0.622	0.645	**0.664**	0.66	0.651	0.655	0.656	–
XLM-R_NS	0.603	0.605	0.603	0.607	0.633	0.614	0.670	**0.689**	0.668	0.672	0.672	0.653	–
XLM-R_TBD	0.564	0.566	0.570	0.562	0.601	0.605	0.614	0.618	0.608	**0.620**	0.618	0.610	–
XLM-R_TB	0.580	0.555	0.595	0.605	0.607	0.610	0.633	**0.643**	0.633	0.631	0.626	0.601	–
W2V						–							0.622
W2V_NS						–							0.628
BASELINE						–							0.589

Table B.2: Results on ES-TimeBank

Model	Layer score 1	2	3	4	5	6	7	8	9	10	11	12	Score
XLM-R	0.286	0.321	0.275	0.307	0.318	0.371	0.396	0.379	0.379	0.379	**0.400**	0.346	–
XLM-R_NS	0.264	0.300	0.336	0.300	0.307	0.336	0.332	0.368	0.371	0.389	**0.404**	0.354	–
XLM-R_TBD	0.293	0.307	0.311	0.336	0.361	0.393	0.382	**0.393**	0.339	0.336	0.357	0.282	–
XLM-R_TB	0.279	**0.307**	0.300	0.296	0.271	0.271	0.279	0.296	0.268	0.279	0.271	0.246	–
W2V						–							0.279
W2V_NS						–							0.268
BASELINE						–							0.321

Table B.3: Results on EN-TimeBank

Model	Layer score 1	2	3	4	5	6	7	8	9	10	11	12	Score
XLM-R	0.480	0.460	0.520	0.515	0.545	0.551	0.535	0.540	**0.581**	**0.581**	0.545	0.530	–
XLM-R_NS	0.490	0.540	0.545	0.551	0.591	**0.621**	**0.621**	0.591	0.606	0.616	0.586	0.601	–
XLM-R_TBD	0.460	0.470	0.505	0.535	0.510	0.535	0.510	0.515	0.535	0.586	0.545	**0.566**	–
XLM-R_TB	0.470	0.500	0.485	0.530	**0.495**	0.490	0.460	0.465	0.455	0.465	0.414	0.338	–
W2V						–							0.369
W2V_NS						–							0.444
BASELINE						–							0.253

Table B.4: Results on IT-TimeBank

Model	Layer score 1	2	3	4	5	6	7	8	9	10	11	12	Score
XLM-R	0.338	0.324	0.338	0.396	0.424	0.482	**0.511**	**0.511**	0.482	0.460	0.446	0.439	–
XLM-R_NS	0.331	0.345	0.388	0.403	0.424	0.482	0.511	**0.525**	0.518	0.511	0.511	0.453	–
XLM-R_TBD	0.331	0.324	0.367	0.432	0.432	0.468	0.475	**0.482**	0.460	0.439	0.460	0.475	–
XLM-R_TB	0.353	0.345	0.324	0.388	0.353	0.388	**0.410**	0.381	0.396	0.345	0.353	0.295	–
W2V						–							0.345
W2V_NS						–							0.317
BASELINE						–							0.216

Table B.5: Results on FR-TimeBank

C Statistical Testing

The Tables from C.1 to C.10 illustrate the results of the McNemar's tests for each language and each probing model (including the baseline based on the most frequent class). The values in all the Tables correspond to p-values. The threshold of the α value for significance has been set to < 0.05.

C.1 PTLMs last layer

The Tables from C.1 to C.5 report the p-value scores for each language when using the last layer of each PTLM.

Model	XLM-R	XLM-R_NS	XLM-R_TBD	XLM-R_TB	W2V	W2V_NS	BASELINE
XLM-R	–	0.235	0.734	< 0.001	< 0.001	0.147	0.007
XLM-R_NS		–	0.171	0.006	< 0.001	0.652	0.085
XLM-R_TBD			–	< 0.001	< 0.001	0.055	< 0.001
XLM-R_TB				–	0.057	0.002	0.033
W2V					–	< 0.001	0.002
W2V_NS						–	0.069
BASELINE							–

Table C.1: Significance EN-TB-Dense - last layer.

Model	XLM-R	XLM-R_NS	XLM-R_TBD	XLM-R_TB	W2V	W2V_NS	BASELINE
XLM-R	–	0.911	0.037	0.014	0.136	0.203	0.003
XLM-R_NS		–	0.053	0.024	0.181	0.271	0.006
XLM-R_TBD			–	0.740	0.664	0.498	0.410
XLM-R_TB				–	0.242	0.065	0.031
W2V					–	0.736	0.057
W2V_NS						–	0.006
BASELINE							–

Table C.2: Significance ES-TimeBank - last layer.

Model	XLM-R	XLM-R_NS	XLM-R_TBD	XLM-R_TB	W2V	W2V_NS	BASELINE
XLM-R	–	0.888	0.038	0.004	0.053	0.028	0.555
XLM-R_NS		–	0.022	0.002	0.040	0.022	0.444
XLM-R_TBD			–	0.358	1.000	0.760	0.343
XLM-R_TB				–	0.386	0.624	0.033
W2V					–	0.780	0.290
W2V_NS						–	0.184
BASELINE							–

Table C.3: Significance EN-TimeBank - last layer.

Model	XLM-R	XLM-R_NS	XLM-R_TBD	XLM-R_TB	W2V	W2V_NS	BASELINE
XLM-R	–	0.029	0.435	< 0.001	0.024	0.556	< 0.001
XLM-R_NS		–	0.419	< 0.001	< 0.001	0.024	< 0.001
XLM-R_TBD			–	< 0.001	0.001	0.171	< 0.001
XLM-R_TB				–	0.047	< 0.001	0.009
W2V					–	0.065	< 0.001
W2V_NS						–	< 0.001
BASELINE							–

Table C.4: Significance IT-TimeBank - last layer.

Model	XLM-R	XLM-R_NS	XLM-R_TBD	XLM-R_TB	W2V	W2V_NS	BASELINE
XLM-R	–	0.832	0.359	0.003	0.099	0.019	< 0.001
XLM-R_NS		–	0.728	0.001	0.060	0.011	< 0.001
XLM-R_TBD			–	< 0.001	0.007	0.728	< 0.001
XLM-R_TB				–	0.222	0.766	0.080
W2V					–	0.327	0.005
W2V_NS						–	0.038
BASELINE							–

Table C.5: Significance FR-TimeBank - last layer.

C.2 PTLMs best layer

The tables from C.6 to C.10 report the *p*-value scores for each language when using the best layer of each PTLM.

Model	XLM-R	XLM-R_NS	XLM-R_TBD	XLM-R_TB	W2V	W2V_NS	BASELINE
XLM-R (11)	–	0.639	0.830	0.005	< 0.001	0.036	< 0.001
XLM-R_NS (11)		–	0.901	0.017	< 0.001	0.081	0.003
XLM-R_TBD (12)			–	0.013	< 0.001	0.055	< 0.001
XLM-R_TB (6)				–	< 0.001	0.608	0.430
W2V					–	< 0.001	0.002
W2V_NS						–	0.069
BASELINE							–

Table C.6: Significance EN-TB-Dense - best layer.

Model	XLM-R	XLM-R_NS	XLM-R_TBD	XLM-R_TB	W2V	W2V_NS	BASELINE
XLM-R (8)	–	0.198	0.037	0.152	0.067	0.124	0.002
XLM-R_NS (8)		–	< 0.001	0.005	0.001	0.004	< 0.001
XLM-R_TBD (8)			–	0.597	0.935	0.740	0.251
XLM-R_TB (10)				–	0.640	0.897	0.003
W2V					–	0.736	0.057
W2V_NS						–	0.006
BASELINE							–

Table C.7: Significance ES-TimeBank - best layer.

Model	XLM-R	XLM-R_NS	XLM-R_TBD	XLM-R_TB	W2V	W2V_NS	BASELINE
XLM-R (11)	–	1.000	0.043	0.002	< 0.001	< 0.001	0.045
XLM-R_NS (11)		–	0.066	0.003	< 0.001	< 0.001	0.035
XLM-R_TBD (9)			–	0.241	0.097	0.052	0.691
XLM-R_TB (8)				–	0.668	0.470	0.520
W2V					–	0.780	0.290
W2V_NS						–	0.184
BASELINE							–

Table C.8: Significance EN-TimeBank - best layer.

Model	XLM-R	XLM-R_NS	XLM-R_TBD	XLM-R_TB	W2V	W2V_NS	BASELINE
XLM-R (10)	–	0.256	1.000	0.212	< 0.001	0.072	< 0.001
XLM-R_NS (7)		–	0.382	0.015	< 0.001	0.003	< 0.001
XLM-R_TBD (10)			–	0.177	< 0.001	0.036	< 0.001
XLM-R_TB (4)				–	0.013	0.532	< 0.001
W2V					–	0.065	< 0.001
W2V_NS						–	< 0.001
BASELINE							–

Table C.9: Significance IT-TimeBank - best layer.

Model	XLM-R	XLM-R_NS	XLM-R_TBD	XLM-R_TB	W2V	W2V_NS	BASELINE
XLM-R (8)	–	0.804	0.503	< 0.001	< 0.001	< 0.001	< 0.001
XLM-R_NS (8)		–	0.286	0.003	< 0.001	< 0.001	< 0.001
XLM-R_TBD (8)			–	0.015	0.009	0.002	< 0.001
XLM-R_TB (7)				–	0.608	0.164	< 0.001
W2V					–	0.327	0.005
W2V_NS						–	0.038
BASELINE							–

Table C.10: Significance FR-TimeBank - best layer.