Exhuming a Swedish Temporal Relation Dataset from the Past

Richard Johansson
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
richard.johansson@gu.se

Abstract

I resurrected a dataset containing annotated event-to-event temporal relations in Swedish traffic accident reports, and then benchmarked how well contemporary large language models classify these relations. Can the current generation of LLMs outperform our results from two decades ago?

1 Introduction

My initiation as researcher in NLP took place in 2003 at Lund University in the context of Pierre's Vinnova-funded project *Development of a Text-to-Scene Converter for Vehicle Accident Reports*, where we developed an automatic system that converted traffic accident reports written in Swedish into animated sequences. In retrospect, the project had goals that seem absurdly ambitious for its time; only very recently, I encountered a paper describing an implementation of essentially the same idea, building on 2020s technology (Elmaaroufi et al., 2024). In the end, we were only able to scratch the surface of this immense task, given the state of technology at the time. However, the great variety of fundamental technical, linguistic, and philosophical challenges that we encountered in that project has influenced my research interests profoundly in the ensuing years.

The system we developed – *Carsim* (Johansson et al., 2005) – consisted of three subsystems: an NLP system, which converted the raw text into a formal representation describing the objects and events mentioned in the narrative; a scenario planner that applied physical reasoning to determine a plausible realization in the physical world of the formal representation; and finally a 3D visualizer presenting the output of the planner as an animated video. The NLP part of the Carsim system consisted of a chain of modules that carried out linguistic analysis at different levels. Except for the part-of-speech tagger, all of them were in-house implementations.

While some of these linguistic modules were rule-based domain-specific heuristics, others were based on supervised machine learning solutions where the training data was annotated by members in the group, including thesis students supervised by Pierre. For instance, Carsim's noun phrase coreference solver came out of a thesis project by Danielsson (2005), where he annotated a substantial number of coreference chains and then implemented the decision tree-based approach by Soon et al. (2001). I also annotated some of these training datasets; most importantly for the research in the later part of my PhD period, Carsim had a semantic role labeler based on a domain-specific adaptation of a small set of FrameNet frames. Pierre lent me his printed copy of a recent *Computational Linguistics* issue and pointed me to the article by Gildea and Jurafsky (2002), and we reimplemented their statistical semantic role classifier and trained it on a dataset I annotated.

For this Festschrift, I thought that it would be interesting to revisit one or more of these tasks and datasets, and see how our solutions from that time compare to what can be achieved with modern NLP techniques. For reasons I will describe below, I decided to carry out a set of experiments with a dataset annotated by Anders Berglund for his Master's thesis project (Berglund, 2004), where he considered the problem of determining *temporal relations* between events mentioned in a text. This is not only an exercise in nostalgia but also an interesting benchmarking experiment for modern large language models (LLMs), in which we can investigate their capability of reasoning about complex narratives written in Swedish.

2 Recreating the Temporal Relation Dataset

I ran into the practical difficulty that the Carsim implementation has not been published as a repository and appeared to be lost in the mists of time. However, I found a solution. When I was about to leave Lund in early 2009, Lars Nilsson was generous

From Electrophoresis to Wikidata: Festschrift in honor of Pierre Nugues, pages 11–15.
Published by University of Tartu Library, May 2025.

enough to let me keep the hard drive from my desktop computer. For some reason, I did not discard this disk and it eventually ended up among some old junk in a box in my apartment, where I recently rediscovered it. I had a SCSI-to-USB adapter and the disk had not broken down in the years since it was last used. Eventually, I was able to locate several interesting files from the past including the full directory of the Carsim implementation, with all source code and data still present.

After exploring our old files, I concluded that Berglund's temporal relation dataset was the easiest one to work with as well as the most interesting on an intellectual level. From a practical point of view, it was convenient that Berglund had used a comparatively modern stand-off annotation format, which made it possible for me to extract his data (after some mild annoyances and manual correction of token offsets). The noun phrase coreference dataset by Danielsson (2005) was also neatly formatted, but it seemed to me that analyzing temporal relations between events would be a more challenging and interesting task for a modern NLP evaluation.

3 Modeling and Annotating Temporal Relations between Events

To exemplify the annotation of temporal relations between events, consider the example below. In the example, the five events (e_1–e_5) mentioned in the text have been underlined.

> Two people died$_{e_1}$ late yesterday evening when a car drove off$_{e_2}$ the road and crashed$_{e_3}$ into a tree. The car was overtaking$_{e_4}$ another car when the driver lost control$_{e_5}$ of it.

It is worth stressing here that the events are not presented in a chronological order, and the writer first expresses the most salient information: that there were fatalities. To understand how the events are positioned in time in relation to each other, the reader has to form a mental model of what happened in the described scenario. To understand the temporal structure, the reader can also consider time expressions (*late yesterday evening*), temporal connectives (*when*) and the interplay of tenses and aspects of the verbs (*crashed*, *was overtaking*).

There are multiple conceptual frameworks for modeling events and their relations. One famous framework is by Allen (1984), who defined 13 relation types. If we apply Allen's framework to annotate the event-to-event relations in the example above, we get the structure presented in Figure 1.

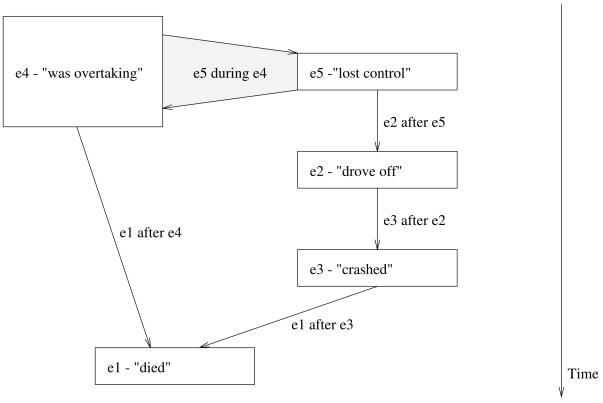

Figure 1: Allen-style relations between the events in the example text. Figure by Berglund (2004).

TimeML (Pustejovsky et al., 2003a) is one of the most widely known annotation models for annotating events and their temporal relations, and this model has been used to annotate the TimeBank corpus (Pustejovsky et al., 2003b) in English, as well as several related projects for other languages. Berglund's thesis project used a simplified TimeML annotation scheme (Berglund et al., 2006a): our project only required us to annotate relations between events that took place within the narrative described in the text. The full TimeML model can also represent e.g. hypothetical scenarios.

4 Experimental Setup

4.1 Data

After reverse-engineering Berglund's format (§2), I ended up with a dataset consisting of 27 news reports taken from a larger collection gathered in the Carsim project from various Swedish news sources, primarily *Sydsvenskan*. This corpus includes 904 annotated temporal relation instances, after discarding a few instances where the annotator was uncertain. In the experiments, I only included the 904 relations directly annotated in the dataset, and I did not expand the set of instances by applying temporal reasoning (e.g. transitivity and symmetry).

The annotation uses five temporal relation types: *after*, *before*, *includes*, *is_included*, and *simultaneous*. The most common annotated relation type (385 instances) is *before*. This reflects the fact that relations are annotated from the first event to the second, and that events are often presented somewhat chronologically in the narrative.

4.2 Selected Models

In the experiments, I evaluated a set of LLMs from the following families. They were accessed through their respective APIs without any fine-tuning.

- OpenAI's GPT models: *GPT-4o*, *GPT-4.1*,[1] and the *o3-mini* reasoning model.
- Meta's Llama models: *Llama 3 70B* (Llama Team, 2024), *Llama 3.1 405B*, and *Llama 4 Maverick*, via the Replicate API.
- DeepSeek's *V3* (chat) and *R1* (reasoning) models (DeepSeek-AI, 2025).

Before considering whether how well these models are capable of reasoning about complex scenarios, it is worth asking whether these models have any Swedish-language capabilities at all. In previous benchmarking experiments for Swedish, they have been found to perform quite well, including in a sense disambiguation exercise I carried out previously (Johansson, 2024).

4.3 Prompt Design and Model Execution

The prompts consisted of the following parts:

1. a preamble describing the temporal relation classification task and defining the five relation labels;
2. a demonstration of the classification task;
3. the full text, where the two events under consideration are surrounded by XML tags `<event1>` and `<event2>`, respectively.

The models were applied in two different settings:

- *direct* prediction: the model is given the prompt and has to output the predicted relation label directly;
- *chain-of-thought* prediction (Wei et al., 2022): the model is given the prompt and is asked to provide a textual explanation before predicting the label.

The two reasoning models (OpenAI's `o3-mini` and DeepSeek's V1) use an internal chain-of-thought process so the direct prediction approach is not applicable for those models.

Since I am primarily interested in how well the LLMs model the temporal structure of the narrative, temporal relations were predicted for individual event pairs and I did not enforce global consistency of the temporal graph (e.g. by breaking cycles).

[1]This model was released as on the day I was finishing this paper, so I had to carry out some last-minute experiments.

Model	D	CoT
`o3-mini`		0.691
`deepseek-reasoner`		0.684
`gpt-4.1`	0.486	0.640
`deepseek-chat`	0.311	0.633
`llama-4-maverick`	0.243	0.620
`llama-3.1-410b-instruct`	0.480	0.553
`gpt-4o`	0.335	0.508
`llama-3-70b-instruct`	0.327	0.362
Baseline	0.426	

Table 1: Accuracies for all models with direct prediction (D) or chain-of-thought prediction (CoT).

5 Results

5.1 Can LLMs Classify Temporal Relations?

Table 1 shows the classification accuracies for all models. This comparison also includes a trivial baseline that assumes that events are presented linearly in a chronological order: that is, it consistently predicts that the first event happens *before* the second event. This is also the majority-class baseline.

A few observations can be made about these results. First, it is clear that it is difficult for all these LLMs to classify the temporal relations *directly*: while all evaluated models outperform a uniform random baseline (0.20), only a couple of them reach the accuracy of the trivial baseline assuming a chronological order. This shows that temporal relation classification is a comparatively difficult task for LLMs; one has to be careful in drawing "cognitive" conclusions from experiments like this, but these results do not suggest that the current generation of LLMs form an internal representation of the events described in the narrative, or at least not one that easy for the model to access.

On the other hand, there are consistent improvements in the quality of predictions when the models are prompted to present their reasoning. This improvement is most notable for a couple of the more recent models (Llama 4 and DeepSeek chat), which saw low accuracies in the direct prediction setting but much higher in the chain-of-thought setting. The strongest models are the two reasoning models, which include mechanisms to improve chain-of-thought reasoning at training and inference time.

Are these results comparable to the capability of

Model	U	D
o3-mini	0.704	0.900
deepseek-reasoner	0.682	0.811
gpt-4.1	0.440	0.367
llama-3.1-410b-instruct	0.833	0.401
gpt-4o	0.610	0.181
llama-3-70b-instruct	0.714	0.103

Table 2: Consistency of predicted undirected (U) and directed (D) temporal relations.

human annotators? It is difficult to compare these predictive accuracies to human-to-human agreement levels, since Berglund (2004) did not carry out an inter-annotator agreement study. The accuracy of the best model compared to the human annotation corresponds to a Cohen's κ of 0.56, which is lower than the κ of 0.71 for relation type annotation reported for TimeBank,[2] but this comparison must of course be taken with a grain of salt.

5.2 How Consistent are the Predicted Relations?

If LLMs form some sort of representation of the events in the narrative, one would expect the predicted relations to be structurally consistent. For instance, if the model predicts that the police arrived *after* a traffic accident, it should also predict that the accident happened *before* police arrived.

To investigate the consistency of predicted relations, I switched the order of the <event1> and <event2> tags in the prompt and compared the predictions. The evaluation considers one *undirected* relation type (*simultaneous*), where labels should not change, and four *directed* types where labels should change to their inverses (e.g. *after*→*before*).

Table 2 shows the result of the consistency evaluation. The results again show that non-reasoning LLMs have rather poor temporal processing capabilities, but reasoning models are much better in this respect. In particular, the predicted directed relations are much more consistent for these models.

5.3 Can the Best Models Outperform a Decision Tree?

Berglund's thesis project also included the development of a decision tree-based classifier that predicts the type of temporal relation holding between two given events. (This work was later published as a separate EACL paper (Berglund et al., 2006b).) This classifier used a large feature set based on linguistic features of the two event mentions as well as various structural features (e.g. distances). Do current LLMs perform better than our twenty-year-old decision tree classifier?

Model	Accuracy
Carsim	0.728
o3-mini	0.728
deepseek-reasoner	0.717

Table 3: Results on the Carsim-annotated subset.

In the short time I had to write this paper, I was unable to run the Carsim implementation or to disentangle the decision tree classifiers from the rest of the code. However, I found a file where Carsim had computed temporal relation labels for event pairs in 10 of the texts in the corpus, which I could compare to Berglund's manual annotations. I then looked at the LLM predictions for the same subset.

Table 3 shows the result of this evaluation. The best LLM (o3-mini) correctly classified 134 event pairs (an accuracy of 0.728), which is exactly the same as the number of pairs correctly labeled by Carsim's classifier. I suppose this means that we can tentatively conclude that 2025 was the year when the field of NLP caught up with the work we did in Pierre's group at LTH two decades earlier.

6 Final Words

As I mentioned in the introduction, my interests as a researcher have been shaped by the research problems we encountered in this first project I worked on as a PhD student under Pierre's supervision. My current research is a bit different, but I have often thought it would be interesting to return to some of the research questions we worked on at that time, as in the little investigation carried out for this paper. I am grateful to Pierre for approaching me out of the blue after I took his course in Natural Language Processing and Computational Linguistics to suggest that I apply for his PhD position, and then for being a patient PhD supervisor in the five years after that. Those were fun years when I could be free and irresponsible and work on fascinating research problems, and along the way have interesting conversations with a free-thinking and erudite supervisor on all sorts of topics!

[2] https://timeml.github.io/site/timebank/documentation-1.2.html#iaa

References

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

Anders Berglund. 2004. Extracting temporal information and ordering events for Swedish. Master's thesis, Lund University.

Anders Berglund, Richard Johansson, and Pierre Nugues. 2006a. Extraction of temporal information from texts in Swedish. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Anders Berglund, Richard Johansson, and Pierre Nugues. 2006b. A machine learning approach to extract temporal information from texts in Swedish and generate animated 3D scenes. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 385–392, Trento, Italy. Association for Computational Linguistics.

Magnus Danielsson. 2005. Maskininlärningsbaserad koreferensbestämning för nominalfraser applicerat på svenska texter. Master's thesis, Lund University.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Karim Elmaaroufi, Devan Shanker, Ana Cismaru, Marcell Vazquez-Chanlatte, Alberto Sangiovanni-Vincentelli, Matei Zaharia, and Sanjit A. Seshia. 2024. ScenicNL: Generating probabilistic scenario programs from natural language. In *Proceedings of the Conference on Language Modeling (COLM)*, Philadelphia, United States.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Richard Johansson. 2024. How well do large language models disambiguate Swedish words? In *Proceedings of the Swedish Language Technology Conference (SLTC)*, Linköping, Sweden.

Richard Johansson, Anders Berglund, Magnus Danielsson, and Pierre Nugues. 2005. Automatic text-to-scene conversion in the traffic accident domain. In *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, pages 1073–1078, Edinburgh, United Kingdom.

AI @ Meta Llama Team. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir Radev. 2003a. TimeML: Robust specification of event and temporal expressions in text. In *AAAI Spring Symposium on New Directions in Question-Answering (Working Papers)*, pages 28–34.

James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, Daniel Day, Lisa Ferro, and Marcia Lazo. 2003b. The TIMEBANK corpus. In *Corpus Linguistics*, pages 647–656.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.