# MLLabs-LIG at TempoWiC 2022: A Generative Approach for Examining Temporal Meaning Shift

**Chenyang Lyu**[†*] **Yongxin Zhou**[‡*] **Tianbo Ji**[†]
[†] School of Computing, Dublin City University, Dublin, Ireland
[‡] LIG, Univ. Grenoble Alpes, Grenoble, France
{chenyang.lyu2, tianbo.ji2}@mail.dcu.ie, yongxin.zhou@univ-grenoble-alpes.fr

## Abstract

In this paper, we present our system for the EvoNLP 2022 shared task Temporal Meaning Shift (TempoWiC). Different from the typically used discriminative model, we propose a generative approach based on pre-trained generation models. The basic architecture of our system is a seq2seq model where the input sequence consists of two documents followed by a question asking whether the meaning of target word changed or not, the target output sequence is a declarative sentence describing the meaning of target word changed or not. The experimental results on TempoWiC test set show that our best system (with time information) obtained an accuracy and Macro-F1 score of 68.09% and 62.59% respectively, which ranked 12th among all submitted systems. The results have shown the plausibility of using generation model for WiC tasks, meanwhile also indicate there's still room for further improvement.

## 1 Introduction

The EvoNLP Shared Task Temporal Meaning Shift (TempoWiC) (Loureiro et al., 2022) aims to judge whether the meaning of a target word in a pair of sentences (two tweets in this case) change or not. Different from original WiC (Word-in-Context) task, TempoWiC takes into account the temporal information in the text, as tweets in each pair are with the date when they are posted. Therefore, that brings new challenges into this task - *how to make use of the temporal information in the tweets?*

Unlike Conventional approaches for WiC tasks that typically adopt discriminative models (such as BERT or RoBERTa) with an input consisting of a pair of sentences (Devlin et al., 2019; Liu et al., 2019; Loureiro et al., 2021), here we propose a generative approach. In our approach we treat the temporal information in tweets as normal words without complicated designation for task schema, aiming at exploiting the natural language understanding (NLU) capability of the pre-trained generation model (Radford et al., 2019; Lewis et al., 2020). Moreover, that could potentially inspire further developments for such tasks based on large language models (LLMs) such as GPT-3 (Brown et al., 2020) with Prompt Learning which has been the focus of the community recently due to its superior zero-shot performance (Liu et al., 2021).

Specifically, for the input sequence we concatenate two tweets which are followed by a question asking whether the meaning of the target word change or not. The output sequence is a declarative sentence stating whether the meaning of the target word changed or not in the two tweets. In our approach, TempoWiC is framed as a generative QA task and the construction of the data follows the manner of template-based Question Generation (Lewis and Fan, 2019; Heilman and Smith, 2009; Fabbri et al., 2020; Lyu et al., 2021). With the training data constructed in such format, we fine-tune the pre-trained generation models as a seq2seq model using a vanilla autoregressive generation objective (Sutskever et al., 2014; Johnson et al., 2017; Yang et al., 2019).

We submitted two systems (one with time information and the other one without time information) to TempoWiC for evaluation. Results show that our best best system (with time information) on the TempoWiC test set obtained an accuracy and Macro-F1 score of 68.09% and 62.59% respectively, which slightly outperforms the other system without time information with an accuracy and Macro-F1 score of 68.02% and 61.14%. Based on the evaluation results on TempoWiC test set, we found that pre-trained generation models are capable of capturing the meaning shift of target word in context. Besides, results also show that time information (date of each tweet) can provide further improvement for performance, showing the

[*]These authors contributed equally to this work.



*Proceedings of the The First Workshop on Ever Evolving NLP (EvoNLP)*, pages 1 - 6

importance of temporal information. In the rest of this paper, we will introduce the architecture of our system and give more detailed experimental results.

## 2 Methodology

In this section, we briefly introduce how TempoWiC task is formulated in this paper and how to train our system.

### 2.1 Task formulation

We frame the TempoWiC task as a seq2seq generation task where the source sequence consists of two tweets followed by a question, the target sequence is a declarative sentence. Moreover, in order to avoid the generation of naive output (e.g., all same output), we have some specific designation for the input and output sequence: the question in the source sequence must be an interrogative sentence specific to the target word, also the target sequence has to include the target word. As a generative approach, the format of input and output sequence of our system are as follows:

- **Input**: *Tweet1 - Tweet2 - Question: Does the meaning of word X change?*
- **Output**: *Is the meaning of X different in the last two tweets?*

Note *X* is the target word that we wish to examine whether its meaning changed or not in the two tweets. For instance, a concrete example in such format is shown as follows:

**Input:** *Tweet-1: The book in 19th century is fantastic..... Date: 2018-03. Tweet-2: Need help to book the next-day flight..... Date: 2019-03. Question: Is the meaning of **book** different in the last two tweets?*

**Output:** *Answer: No, the meaning of **book** is not the same.*

where we highlight the target word ***book***, of which the presence in the input and output sequence is essential for PLMs to learn how to measure the meaning shift of the target word.

### 2.2 Training objective

Our training objective is to minimize the Negative Log Likelihood Loss with respect to the parameters $\theta$ of our autoregressive generation systems:

$$J(\theta) = -logP(q|c,a) = \sum_i logP(a_i|a_{<i}, c, q) \tag{1}$$

where $a$ is our target sequence *answer*, $c$ is the *context* (two tweets) and $q$ is the *question*.

## 3 Experiment

| Model | Accuracy | Macro F-1 |
|---|---|---|
| BART-base | 65.91 | 63.33 |
| BART-large | 69.19 | 65.72 |

Table 1: Performance of BART-base and BART-large on TempoWiC validation set.

### 3.1 Data

The training, validation and test set of TempoWiC data contain 1428, 396 and 10000 examples respectively. We show the average length of the two tweets in TempoWiC dataset in Table 2, where we found that the average length of the first tweet is typically longer than the second tweet in all splits especially in validation set.

### 3.2 Training Setup

We employ BART (Lewis et al., 2020) as our seq2seq model, which is Pre-trained Language Models (PLMs) that have been shown to be effective in various natural language generation tasks (Lai et al., 2021; Lewis et al., 2021; Zhou et al., 2022). Our implementation is based on BART-base and BART-large (Lewis et al., 2020) from Huggingface (Wolf et al., 2020). We train our system with a learning rate of $3 \times 10^{-5}$ for 10 epochs, the batch size is set to 4. We use a maximum source sequence length of 512 and a maximum target sequence length of 64.

### 3.3 Results

We report the results on validation set of BART-base and BART-large in Table 1, the results show that BART-large outperforms BART-base by 3.28 accuracy and 4.39 Macro F-1 score. Therefore, our two submitted systems are based on BART-large. Table 3 shows the results of the official TempoWiC Competition Leaderboard (29 September 2022), our best system (with time information) ranked 12th with an accuracy and Macro F-1 score of 68.09% and 62.59% respectively, meanwhile our

| Split | Number of Examples | Avg. Len. of Tweet-1 | Avg. Len. of Tweet-2 | Number of Target Words |
|---|---|---|---|---|
| Train | 1428 | 31.79 | 30.68 | 15 |
| Validation | 396 | 28.20 | 24.29 | 4 |
| Test | 10000 | 26.74 | 25.70 | 15 |

Table 2: Average length for the tweets in the train, validation and test set as well as the number of target words of TempoWiC dataset.

| Rank | User/Baseline | Accuracy | Macro-F1 |
|---|---|---|---|
| 1 | dma | 78.34% | 77.05% |
| 2 | macd | 77.53% | 76.60% |
| 3 | zackchen | 75.49% | 74.87% |
| 4 | dmi | 74.13% | 73.37% |
| 5 | wangkangxu | 73.46% | 73.08% |
| 6 | vol | 73.66% | 72.54% |
| - | TimeLMs - SIM | 74.07% | 70.33% |
| 7 | lisatukhtina | 70.94% | 70.09% |
| 8 | subhamkumar | 70.47% | 69.79% |
| 9 | mahhars | 70.47% | 69.79% |
| 10 | eternalfeather | 69.18% | 68.49% |
| - | RoBERTa-L - SIM | 72.98% | 67.09% |
| 11 | nst | 72.51% | 63.75% |
| 12 | **MLLabs-LIG** | **68.09%** | **62.59%** |
| 13 | yashamz | 63.00% | 61.97% |
| 14 | pgatti | 66.67% | 59.38% |
| - | RoBERTa-L - FT | 66.49% | 59.10% |
| 15 | adityakane | 67.41% | 57.94% |
| - | TimeLMs - FT | 66.46% | 57.70% |
| 16 | PaulTrust | 62.66% | 55.38% |
| 17 | virk | 55.13% | 54.25% |
| 18 | daminglu123 | 50.78% | 50.13% |
| - | Random | 50.00% | 50.00% |
| - | All True | 36.59% | 26.79% |

Table 3: Official TempoWiC Codalab Competition Leaderboard (29 September 2022). Baselines on the TempoWiC are marked in lightgray with their results: **TimeLMs - SIM** is Logistic Regression based on Similarity of Contextual Embeddings from TimeLMs-2019-90M (pooled following LMMS-SP); **RoBERTa-L - SIM** is Logistic Regression based on Similarity of Contextual Embeddings from RoBERTa-Large (pooled following LMMS-SP); **RoBERTa-L - FT** is Fine-tuned RoBERTa-Large (following configuration used in SuperGLUE) and **TimeLMs - FT** is TimeLMs - FT Fine-tuned TimeLMs-2019-90M (following configuration used in SuperGLUE). Besides, **Random** means predictions are randomly assigned T/F and **All True** refers that All instances assigned T (Loureiro et al., 2022).

system without time information obtained an accuracy and Macro F-1 score of 68.02% and 61.14%. Our generative approach outperforms the baselines including *RoBERTa-L-FT*, *TimeLMs-FT*, *Random* and *All True*, whereas underperforms compared to the baselines *TimeLMs-SIM* and *RoBERTa-L-SIM*.

From the results shown in Table 1 and Table 3, we have three main findings:

- The use of a generative approach for the WiC task is plausible as the results show that our system outperforms several competitive baseline models based on BERT
- Temporal information has positive effect in predicting whether the meaning of the target word changed or not
- Larger PLMs is more likely to produce higher performance, indicating further improvement can be achieved with increasing the size of PLMs

| Tweet-1 | Tweet-2 | Target Word | Label | Prediction |
|---|---|---|---|---|
| *I can't believe impostor syndrome just tried to equate the tattoo I'm about to get that symbolizes something I studied for 3.5 years + graduated with a 1.7 in to trying a sport once and getting a tattoo about it, just because I'm sort of mediocre at the actual subject (2019-09)* | *'damn I was in a game of among us and I was SURE red did it like I saw red kill pink but hen red wasn't the impostor??? wtf my mother did suspect I am mildly colorblind' (2020-09)* | *impostor* | 0 | 0 |
| *Today run bts was epic as the Christmas one or the one in lotte world, the boys competitiveness is (2019-09)* | *so we've got namjoon's live on youtube, cns 1 year anniversary, the lotte online concert, tiktok ot7, and the new album announcement today (2020-09)* | *lotte* | 1 | 0 |
| *i watched that lotte world run episode again and now i want to go back and ride french revolution it was so much fun :( (2019-09)* | *Dude I still haven't recovered from lotte family I'm half asleep how do I even react to be? (2020-09)* | *lotte* | 1 | 0 |
| *COLTS/TEXANS TRIVIA for 4 primo seats to the showdown this Thursday night + $5k to help toward travel: Name the Colts player who recovered his own onside kick vs. Houston AND the final score!! (Note, the $ will be received at the game). Alyssa's hat pick! (2019-11)* | *'Somali guys used to be very good tukicheza futa either primo ama zile matches za estate. I wonder why à good number of them never end up in professional football clubs or the national team.' (2020-1)* | *primo* | 0 | 0 |
| *In 2016 Sheila Dixon wanted a recount, Pugh had no integrity. She talked greasy about Sheila. Karma is a real bitch huh Pugh? (2019-11)* | *Delaying the transition only affects the suffering of Americans during Covid. All thank to Trump's false claims. Can't wait to see #Georgia recount show Trump losing. #TrumpConcede' (2020-11)* | *recount* | 1 | 1 |

Table 4: Examples from TempoWiC validation set with corresponding predictions from our system, where 1 represents the meaning of the target word has not changed whereas 0 represents meaning of the target word has changed.

### 3.4 Error analysis

We show some examples with corresponding predictions in Table 4. From Table 4, we found that our proposed generative approach is capable of detecting most meaning shift for the target word (for example *impostor*, *recount* and *primo*). However, it still has difficulties in some cases, such as *lotte*, as shown in the two examples in Table 4 where our system predicts that the meaning of *lotte* has not changed - which is not true. Experimental results as well as error analysis show that pre-trained generative models still have difficulties recognizing some language variation, which needs to be addressed in future developments. We think the fast evolving and changing meanings of words on the web, especially on social media, make this task even more challenging.

## 4 Conclusion

In this paper, we present a generative approach based on Pre-trained Language Models for the TempoWiC task. Experimental results show the plausibility of using pre-trained generative model for the TempoWiC task, which could potentially inspire further developments based on more pre-trained models.

## Limitations

While pre-trained generative models exhibit strong capabilities for natural language understanding with prompts (Liu et al., 2021), there are still issues to be addressed such as explainability, controllability of outputs as these systems fully reply on the generative ability of pre-trained models. Moreover,

how to correctly understand instructions in context/prompt (such as questions) is still challenging for pre-trained models (Min et al., 2022; Jang et al., 2022).

## References


Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. Template-based question generation from retrieved sentences for improved unsupervised question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4508–4513, Online. Association for Computational Linguistics.

Michael Heilman and Noah A Smith. 2009. Question generation via overgenerating transformations and ranking. Technical report, Carnegie-Mellon University.

Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2022. Can large language models truly understand prompts? a case study with negated prompts.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021. Thank you BART! rewarding pre-trained models improves formality style transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 484–494, Online. Association for Computational Linguistics.

Mike Lewis and Angela Fan. 2019. Generative question answering: Learning to answer the whole question. In *International Conference on Learning Representations*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. PAQ: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586.

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. *CoRR*, abs/1907.11692.

Daniel Loureiro, Aminette D'Souza, Areej Nasser Muhajab, Isabella A. White, Gabriel Wong, Luis Espinosa Anke, Leonardo Neves, Francesco Barbieri, and Jose Camacho-Collados. 2022. Tempowic: An evaluation benchmark for detecting meaning shift in social media.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and Evaluation of Language Models for Word Sense Disambiguation. *Computational Linguistics*, 47(2):387–443.

Chenyang Lyu, Lifeng Shang, Yvette Graham, Jennifer Foster, Xin Jiang, and Qun Liu. 2021. Improving unsupervised question answering via summarization-informed question generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4134–4148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yongxin Zhou, François Portet, and Fabien Ringeval. 2022. Effectiveness of French language models on abstractive dialogue summarization task. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3571–3581, Marseille, France. European Language Resources Association.