# Classifiers for Caused Motion Construction Identification

**João Coelho**
jmcoelho@andrew.cmu.edu

## Abstract

This study details efforts towards automatic identification of the caused-motion construction (CMC), more specifically its non-prototypical occurrences (e.g., “She sneezed the foam off the cappuccino.”). Building upon previous studies, we develop supervised classifiers for CMC identification, leveraging small transformer-based models and existing human-annotated data. We empirically demonstrate the successful identification of CMCs in open-source datasets, with the best model achieving 86% F1 score in 5-fold cross-validation, thereby providing means for large-scale CMC identification.

## 1 Introduction

Construction Grammar (CxG) is a theory of language that provides a robust framework for understanding how linguistic elements combine to form meaningful structures within language. By integrating both form (syntax) and meaning (semantics), CxG offers valuable insights into how sentences are constructed and interpreted. A particularly important aspect of CxG is the notion of Argument Structure Constructions (ASCs), which pair verbs with argument structures, delineating the number of arguments a verb takes along with their semantic and syntactic roles.

Understanding how state-of-the-art Language Models process and represent such linguistic structures could provide more insights into the inner workings of these natural language systems, which still show limited interpretability. Previous studies (Li et al., 2022) leveraged established psycho-linguistic methodologies to probe neural representations of sentences following a predefined set of ASCs. However, such studies are often conducted on small data samples, with augmentation through simple context-free grammars and/or rules, hindering the generalizability of the results.

Therefore, we aim to address this gap by building classifiers to identify the caused-motion construction. Caused-motion is particularly intriguing as it can manifest as both prototypical instances (e.g., "she kicked the ball into the net") and non-prototypical instances (e.g., "she sneezed the foam off the cappuccino"). While the former is common, the latter represents a tail occurrence where an intransitive verb is coerced into the caused-motion construction. Hence, we focus on building classifiers capable of performing well on both prototypical and non-prototypical instances of caused-motion constructions. For that, we will leverage data that has undergone a human/language model hybrid labeling pipeline, ensuring high-quality annotations. We will then evaluate the performance of multiple transformer-based encoder-only models to assess their ability to accurately identify the caused-motion construction.

More specifically, we analyze four model backbones, considering variations on their size, and also addressing the capacity of their frozen embeddings. The best model, which achieved 86% F1 score on 5-fold validation, was used to label a large corpus of web documents, extracting a set of 70 thousand potential caused-motion clauses. Manual verification hints that the models perform effectively on sentences with prototypical verbs, but have trouble with non-prototypical occurrences, which can be linked to the underlying distribution of the training data. Code and data are made publicly available here.

## 2 Related Work

Recently, there has been an increased interest in probing language models (LM) based on Construction Grammar (Weissweiler et al., 2023), as constructions are deemed essential for comprehensive language modeling. According to CxG, meaning is encoded in abstract configurations of linguistic units, hence the need for LMs to

assign meaning to various linguistic patterns. Recognizing and applying these patterns is crucial for modeling human language behavior and advancing LMs.

Previous studies on probing approaches for constructional information in LMs have employed various methodologies, e.g. focusing on sentence-level embeddings (Madabushi et al., 2020; Weissweiler et al., 2022), or token-level embeddings (Tseng et al., 2022), using them for tasks such as classification or clustering. Interpretation of construction evidence is then drawn for instance by comparing similarity between sentences following the same construction, also focusing on the context that is captured by the token corresponding to the main verb of the sentence.

For the specific case of argument structure constructions, previous work (Li et al., 2022) shows that sentences that share the same construction tend to be more similar than sentences that share the same verb, and the representation of a nonsensical verb within a sentence following a given ASC will be more similar to a prototypical verb if such verb is congruent with the construction, hinting that LMs associate construction with meaning.

However, the above studies are conducted on relatively small data samples, which are then augmented through rules or context-free grammars. Moreover, more robust probing techniques such as MDL (Voita and Titov, 2020) are unfeasible due to lack of training data. This motivates large-scale classification of the phenomenon so that statistically stronger studies can be conducted. For the caused-motion constructions, previous work has established that classification is possible, but the data and models were not made publicly available (Hwang and Palmer, 2015). In similar efforts, other authors have open-sourced a set of approximately 750 annotated caused-motion clauses (Weissweiler et al., 2024), obtained through a hybrid pipeline between human annotation and large language model filtering.

## 3 Methodology

In this study, we leveraged open-source data (Weissweiler et al., 2024) to train and evaluate transformer-based models for the identification of caused-motion constructions in English sentences, which was obtained through a multi-step process. First, dependency parsing

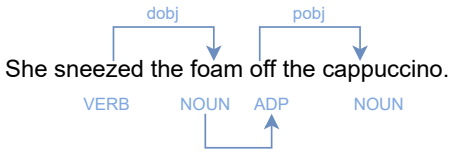


Figure 1: Dependency tree for the caused-motion construction, as used in (Weissweiler et al., 2024).

following the pattern depicted in Figure 1 was conducted. Then, the resulting sentences were filtered using Large Language Models, followed by human labeling. This resulted in a dataset containing 750 positive occurrences of the caused-motion construction, both prototypical and non-prototypical. Additionally, we randomly sampled 750 examples of false positives to use as negative examples for model training.

We consider four encoder-only transformer models, testing their performance and agreement. The models were trained to minimize the standard cross-entropy loss, using a 5-fold cross-validation approach (i.e., 80-20 splits). In terms of model size, both the models' base and base versions were used. Since the training data is small, we also compare full fine-tuning to a frozen-encoder scenario (i.e., only the classification head parameters are trainable). No calibration was conducted since the classes are balanced, hence an example is considered positive if the correspondent class probability is over 50%.

For large-scale identification, we consider the MS-MARCO (Nguyen et al., 2016) and Reddit-ToT (Bhargav et al., 2022) corpora. The documents in both corpora were split into individual sentences, followed by dependency parsing to minimize distributional shifts, and classification with the best-performing model.

## 4 Experiments

Under low data scenarios, freezing the parameters of the underlying language model and tuning only the classification head is a common technique (Devlin et al., 2019). In this work, we compare both versions, as it also provides insights regarding the expressiveness of pre-trained embeddings to detect the caused-motion construction. Table 1 shows that frozen encoders have near-random performance. This suggests that the frozen encoders lack the capacity to model the linguistic intricacies required for accurate

Table 1: Average accuracy and F1-score over 5-fold cross-validation, for base and base versions of the models, considering full fine-tuning and frozen base model.

| | base | | | | large | | | |
|---|---|---|---|---|---|---|---|---|
| | Full | | Frozen | | Full | | Frozen | |
| | Accuracy | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy | F1 |
| RoBERTa (Liu et al., 2019) | 0.8442 | 0.8527 | 0.5507 | 0.6648 | **0.8623** | **0.8676** | 0.5525 | 0.6645 |
| BERT (Devlin et al., 2019) | 0.8273 | 0.8327 | 0.6312 | 0.5273 | 0.8407 | 0.8453 | 0.5781 | 0.5518 |
| ELECTRA (Clark et al., 2020) | 0.8418 | 0.8527 | 0.6563 | 0.7108 | 0.8465 | 0.8506 | 0.6668 | 0.7099 |
| XLNet (Yang et al., 2019) | 0.8255 | 0.8400 | 0.5974 | 0.5769 | 0.6866 | 0.6820 | 0.5519 | 0.5519 |
| Ensemble | 0.8605 | 0.8629 | - | - | - | - | - | - |

CMC classification. This can be due to the semantic orientation of the models' pre-training tasks, and highlights the necessity of fine-tuning the entire model, including the underlying language representations. The full fine-tuned models show improved results in the full fine-tuning setting, with the best base-sized model achieving 84.4% accuracy and 85.2% F1 score. Moreover, we can see that increasing the model size has overall positive impacts on the results, with the RoBERTa-large modeling achieving the best results of 86.2% accuracy and 86.7% F1-score. However, frozen embeddings still achieve near-random performance. In terms of training latency, base models average 4 seconds per epoch, while large models take 15. In inference time, base models classify, on average, 728 sentences per second, while large models classify 204. For hyperparameter details, refer to Appendix A.1.

Despite the poor overall performance of the frozen setting, it is interesting to note that ELECTRA embeddings tend to perform better than the other encoders, for both base and base sized models. This may suggest that its adversarial pre-training task, which is conceptually different from the remaining base models, can impact the capture of linguistic nuances relevant to caused-motion constructions.

One question arises on how the predictions of the four models vary, given that they were trained on the same data. Considering only the base-sized models, given their better results to latency trade-off, Figure 2 shows a Cohen's $\kappa$ matrix for the four models, containing the average value of the 5 folds. While the agreement is substantial, it is not perfect. This motivated an ensemble of the four models through majority voting, which, as seen in the last group of Table 1, improves the results.

Finally, we look into the confusion matrices of

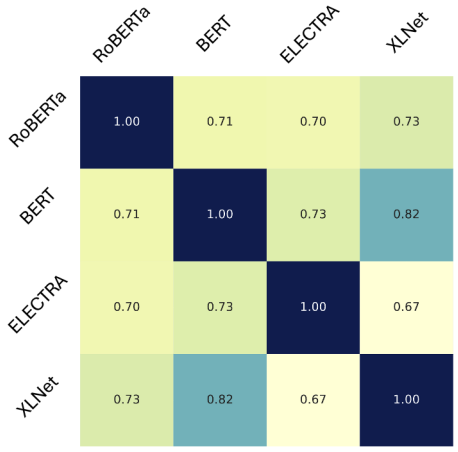


Figure 2: Inter-model agreement measured by Cohen's $\kappa$, averaged over 5-fold cross-validation.

the base-sized models in Figure 3. Overall, the models tend to have a larger rate of false positives when compared to false negatives. Using the majority voting technique provides more balance, but still shows a slight bias toward false positives. This can be attributed to the underlying data, where all the negatives were sampled from a pool of false positives labeled by a large language model.

## 5 Large Scale CMC identification

For large-scale identification, we resort to the fine-tuned RoBERTa-base model, which achieved the strongest performance within the smaller models. We start with approximately 37 million sentences extracted from two web collections, MS-MARCO (Nguyen et al., 2016) and Reddit-ToT (Bhargav et al., 2022). After deduplication and dependency parsing following the pattern on Figure 1, we are left with 3.6 million examples, roughly 10% of the original set. From those, RoBERTa-base identified approximately 150 thousand as caused-motion clauses, when using a

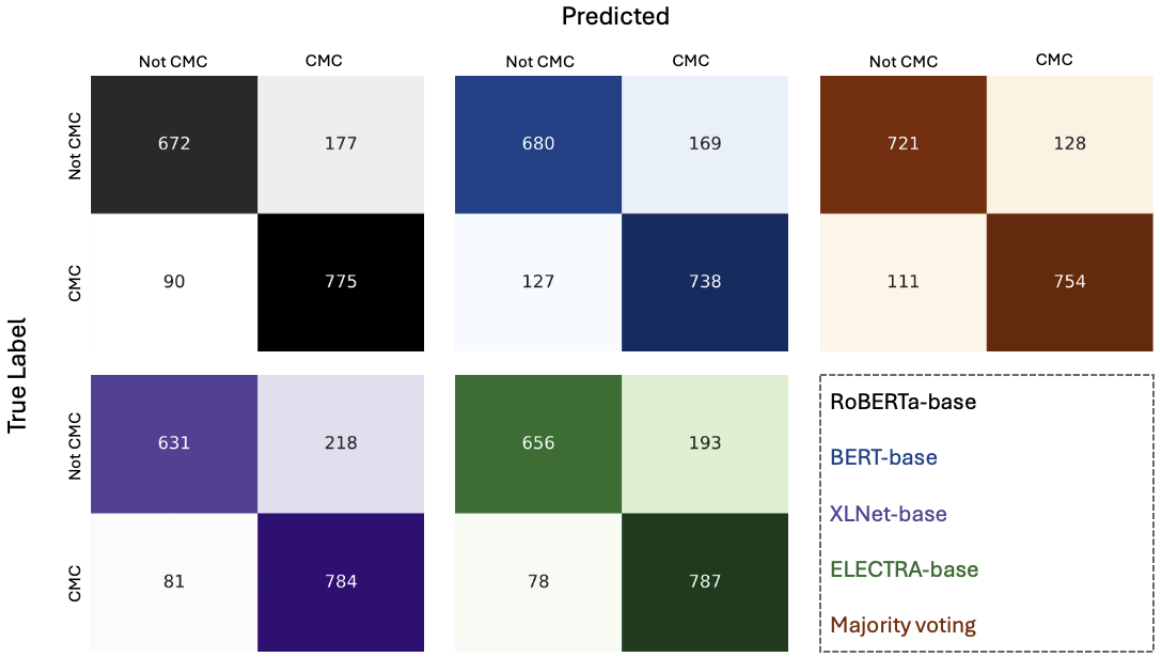


Figure 3: Merged confusion matrices for the models on 5-fold cross-validation, with ensemble results through majority voting.

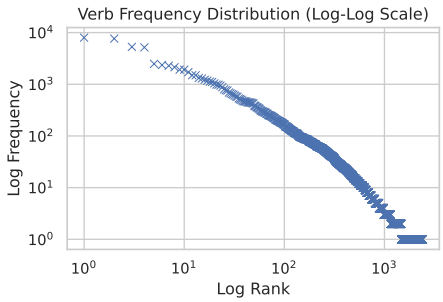


Figure 4: Verb frequency distribution on 70 thousand potential caused-motion clauses.

classification threshold of 50%. To minimize noise associated with the out-of-distribution nature of the initial set of sentences, we set the threshold to 99%, which yields 70 thousand potential caused-motion clauses.

Figure 4 shows the verb frequency distribution in the sentences identified as positives, which, as expected, follows a power-law. Head verbs are materializations of prototypical instances of the phenomena (e.g., "She then proceeds to *throw* the document in a well."), while tail verbs represent the non-prototypical occurrences (e.g., "She *crushed* sleeping pills into his mashed potato in July 2013.").

In order to estimate the precision of the classifier, we take two random samples of 100 sentences: $S_1$, meant to represent head occurrences, contains sentences in which the main verb has 200 or more frequency. In contrast, $S_2$ aims to represent tail occurrences, hence encompassing those where the main verb has 50 or less frequency.

Looking at $S_1$, the model achieves 78% precision. However, for $S_2$, this value drops down to 55%. While the precision of the classifier remains relatively high for sentences representing prototypical instances, it significantly decreases for non-prototypical occurrences. This drop suggests potential challenges in the distribution of training data between prototypical and non-prototypical instances of caused-motion clauses. Addressing this issue may require a more balanced training dataset that contain a diverse range of examples, ensuring the model learns to distinguish between both types of occurrences effectively.

## 6 Conclusions and Future Work

In this study, we developed supervised classifiers for identifying the caused-motion construction. We experimented with transformer-based models, comparing the performance of fine-tuned models versus frozen encoder settings. Our results indicate that fine-tuning the entire model, including the underlying language representations, significantly improves CMC identification when compared to frozen embeddings, suggesting limitations in linguistic understanding of pre-trained language

models. We observed that larger models generally perform better, with the RoBERTa-large model achieving the best results. The best model was used to label approximately 37 million sentences, identifying 70 thousand potential caused-motion clauses.

As for future work, it is worth noting that the models are particularly overconfident, which makes the class probabilities hard to use as thresholds. Hence, the adoption of techniques to mitigate this behavior, such as the generalized binary cross entropy loss (Petrov and MacDonald, 2023), can be useful, particularly when paired with larger sets of more diverse negative examples, which in our dataset were sampled from false positives from previous studies, potentially biasing the model.

Finally, we only evaluated the precision of the classifier on the large-scale corpus. Evaluating the recall should also be done, requiring a more careful sampling technique given the rareness of the phenomenon. Moreover, the data obtained from these manual evaluations can be used for model training to improve the accuracy of the classifiers.

## References


Samarth Bhargav, Georgios Sidiropoulos, and Evangelos Kanoulas. 2022. 'It's on the tip of my tongue' A new Dataset for Known-Item Retrieval. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 48–56.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

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 Conference of the North American Chapter of the Association for Computational Linguistics*.

Jena D. Hwang and Martha Palmer. 2015. Identification of caused motion construction. In *Fourth Joint Conference on Lexical and Computational Semantics*.

Bai Li, Zining Zhu, Guillaume Thomas, Frank Rudzicz, and Yang Xu. 2022. Neural reality of argument structure constructions. In *Annual Meeting of the Association for Computational Linguistics*.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*, abs/1907.11692.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. Cxgbert: BERT meets construction grammar. In *International Conference on Computational Linguistics*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches*.

Aleksandr Vladimirovich Petrov and Craig MacDonald. 2023. gSASRec: Reducing Overconfidence in Sequential Recommendation Trained with Negative Sampling. In *ACM Conference on Recommender Systems*.

Yu-Hsiang Tseng, Cing-Fang Shih, Pin-Er Chen, Hsin-Yu Chou, Mao-Chang Ku, and Shu-Kai Hsieh. 2022. Cxlm: A construction and context-aware language model. In *Language Resources and Evaluation Conference*.

Elena Voita and Ivan Titov. 2020. Information-Theoretic Probing with Minimum Description Length. In *Empirical Methods in Natural Language Processing*.

Leonie Weissweiler, Taiqi He, Naoki Otani, David R. Mortensen, Lori S. Levin, and Hinrich Schütze. 2023. Construction grammar provides unique insight into neural language models. *ArXiv*, abs/2302.02178.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. The better your syntax, the better your semantics? probing pretrained language models for the english comparative correlative. In *Conference on Empirical Methods in Natural Language Processing*.

Leonie Weissweiler, Abdullatif Köksal, and Hinrich Schütze. 2024. Hybrid Human-LLM Corpus Construction and LLM Evaluation for Rare Linguistic Phenomena. *ArXiv*, abs/2403.06965.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Annual Conference on Neural Information Processing Systems*.

# A Appendix

## A.1 Hyperparameters

Everything not stated is defaulted to an Huggingface Trainer. All models are trained on a single A100-40GB.

**All base and large models, frozen encoder:**

- learning rate: 2e-5
- batch size: 20
- epochs: 10

**All base models, full fine-tuning:**

- learning rate: 2e-5
- batch size: 20
- epochs: 10

**RoBERTa-large, full fine-tuning:**

- learning rate: 1e-5
- batch size: 20
- epochs: 10

**BERT-large, full fine-tuning:**

- learning rate: 5e-5
- batch size: 20
- epochs: 10

**ELECTRA-large, full fine-tuning:**

- learning rate: 2e-5
- batch size: 45
- epochs: 10

**XLNet-large, full fine-tuning:**

- learning rate: 8e-6
- batch size: 60
- epochs: 10

Note: XLNet did not achieve comparable results to other large models on any of the setups that were tested.