Consistent Document-Level Relation Extraction via Counterfactuals

Ali Modarressi[1,2] **Abdullatif Köksal**[1,2] **Hinrich Schütze**[1,2]
[1]Center for Information and Language Processing, LMU Munich, Germany
[2]Munich Center for Machine Learning, Germany
amodaresi@cis.lmu.de

Abstract

Many datasets have been developed to train and evaluate document-level relation extraction (RE) models. Most of these are constructed using real-world data. It has been shown that RE models trained on real-world data suffer from factual biases. To evaluate and address this issue, we present CovEReD, **a counterfactual data generation approach for document-level relation extraction datasets using entity replacement**. We first demonstrate that models trained on factual data exhibit inconsistent behavior: while they accurately extract triples from factual data, they fail to extract the same triples after counterfactual modification. This inconsistency suggests that models trained on factual data rely on spurious signals such as specific entities and external knowledge – rather than on the input context – to extract triples. We show that by generating document-level counterfactual data with CovEReD and training models on them, consistency is maintained with minimal impact on RE performance. We release our CovEReD pipeline[1] as well as Re-DocRED-CF, a dataset of counterfactual RE documents, to assist in evaluating and addressing inconsistency in document-level RE.

Original Document:
The Eminem Show is the fourth studio album by American rapper **Eminem**, released on May 26, 2002 by **Aftermath Entertainment**, **Shady Records**, and Interscope Records. **The Eminem Show** includes the commercially successful singles "**Without Me**", "**Cleanin' Out My Closet**", "**Superman**", and "**Sing for the Moment**". **The Eminem Show** reached number one...

RE Model Prediction (TP): (Cleanin' Out My Closet, part of, The Eminem Show)

Counterfactual Document w. Entity Replacement:
London Calling is the fourth studio album by American rapper **Marilyn Manson**, released on May 26, 2002 by **Volcom Entertainment**, **Track Records**, and Interscope Records. **London Calling** includes the commercially successful singles "**Coco**", "**The Ultimate Collection**", "**Watching You**", and "**Déjà Vu**". **London Calling** reached number one...

Missed Prediction (FN): (The Ultimate Collection, part of, London Calling)

Figure 1: Document from Re-DocRED (Tan et al., 2022b) and counterfactual version generated with entity replacement. A model trained on factual data extracts the original triple, but fails on its counterfactual (CF) counterpart. Thus, the model is relying on spurious patterns such as entity biases. We address this by generating CF data and training RE models on them.

1 Introduction

Relation extraction (RE) extracts triples, semantic relations between two entities, from text. In document-level RE, triples can span multiple sentences (Yao et al., 2019; Tan et al., 2022b; Xiaoyan et al., 2023). RE datasets such as DocRED (Yao et al., 2019) and Re-DocRED (Tan et al., 2022b) consist of a factual corpus (Wikipedia) annotated with triples. Most recent DocRE models are based on pretrained language models (PLMs) (Tang et al., 2020; Zhou et al., 2021; Tan et al., 2022a) trained on these datasets. While PLMs perform strongly, they are susceptible to factual biases and other spurious correlations. To generate triples, instead of inferring from the input, they may use their parametric knowledge (McCoy et al., 2019; Kaushik et al., 2020; Paranjape et al., 2022). A common case is entity bias: the model relies on entities in its parametric knowledge to make a prediction (Longpre et al., 2021; Qian et al., 2021; Xu et al., 2022; Chen et al., 2023).

Wang et al. (2022) perform a counterfactual analysis (CoRE) for sentence-level RE. They remove the context and provide only the entity mentions. They then distil the biases and propose a debiasing method using a causal graph. ENTRE (Wang et al., 2023), a counterfactual modification of TACRED (Zhang et al., 2017), replaces entities to develop a robust sentence-level RE benchmark. They show

[1]https://github.com/amodaresi/CovEReD

Findings of the Association for Computational Linguistics: EACL 2024, pages 11501–11507
November 12-16, 2024 ©2024 Association for Computational Linguistics

that RE models rely on memorized facts instead of the sentence context. All of this work is focused on *sentence-level* RE.

This paper presents COVERED, a counterfactual (CF) data generation method for *document-level* RE. It replaces entities and thereby generates text containing triples with minimal factual alignment. We apply COVERED to Re-DocRED, creating RE-DOCRED-CF, a counterfactual document-level RE dataset. Since we apply replacements on the document level, our method handles multiple entity mentions and also multiple replacements at a time – unlike sentence-level methods. We achieve this by considering all triples that an entity is involved in and embeddings of its contexts. Evaluation on RE-DOCRED-CF allows us to measure how consistent a model is in RE. We show that models trained on factual documents lack robustness against nonfactual data (Figure 1). We then train an RE model on Re-DocRED and RE-DOCRED-CF and show that it has high consistency with only a negligible effect on accuracy on factual data. Our approach is novel in that it creates counterfactual datasets on the document level – the level at which RE is used in a real application – to analyze and improve DocRE models. Alongside COVERED, the data generation pipeline, we release RE-DOCRED-CF, a counterfactual dataset generated from Re-DocRED.

2 Counterfactual pipeline and dataset

To evaluate and address robustness against factuality bias, we need to generate documents from which such biases have been removed. Hence, in this section we describe COVERED, our mechanism for generating counterfactual (CF) documents from a document-level RE dataset. Our seed dataset is Re-DocRED (Tan et al., 2022b). For each Re-DocRED document d, we have a set of entities E and a set of relation triples R. For each entity node $e_i \in E$, the dataset provides the positioning of each mention of e_i and its type (ORG, TIME …). In a triple $r \in R$, we have the indices (i) of head and tail entities, the relation r_t and – if the triple comes from the original DocRED (Yao et al., 2019) – the IDs of the sentences that are the evidence for r.

To generate counterfactuals from Re-DocRED's documents, our pipeline COVERED proceeds with the following three steps (§2.1–§2.3).

2.1 Entity mention cleanup

If two entities e_i and e_j share a common (exactly matching) mention, then we merge them; this means that we merge the two sets of mentions and treat e_i and e_j as synonymous. Also, if two mentions overlap in a sentence, we discard the shorter one and only keep the longer one. Example: If "Great Britain" in a sentence is annotated with two mentions "Great Britain" and "Britain", then we only keep "Great Britain".

2.2 Gathering entity candidates

In sentence-level RE, entities are very rarely part of multiple triples, but in document-level RE this is common. If we want to generate consistent counterfactual documents, we have to replace all of an entity's occurrences. This makes it impractical to use simplistic replacement methods in document-level RE – such as relying on an entity bank for random replacement as in ENTRE (Wang et al., 2023).

Another challenge is that we need "plausible" counterfactual documents that do not obviously contradict general knowledge. For example, "Obama was born in Panthera leo" (where Panthera leo is a species, the lion) is too implausible to teach the model about correct RE. We therefore only replace e_i with e_j if they have (i) similar **relation maps** and (ii) similar **context snippets**. For this step, we use the set of entities over the entire seed dataset.

The **relation map** for an entity e_i is a set of pairs, each consisting of a relation and the position of e_i within that relation (head or tail). For example, the relation map of "United States" – occurring in triples such as $\langle$NBA, country, United States$\rangle$ – may contain the pair (country, tail). If two entities e_i and e_j have similar relation maps, then e_i is a good candidate for replacing e_j since they occur in similar triples.

The **context snippet** of a mention m includes up to 16 words on each side of m. For each context snippet, we compute its embedding (using Contriever (Izacard et al., 2022)). If two entities e_i and e_j have similar context embeddings (as measured by cosine similarity), then e_i is a good candidate for replacing e_j since they occur in similar contexts.

2.3 Generating counterfactual documents

Our general approach to generating counterfactual documents is to find suitable entity alternatives for each entity node and apply replacements.

In Algorithm 1, function GETALTS is responsible for finding suitable entity replacements. For each entity node, we compare its features (its type, relation map, mention and context snippet embeddings; see §2.2) with the *candidates* – the other entity nodes in the pool E gathered from the document collection. We deem a candidate a suitable alternative if it is similar and not from the same document. We do not want candidates's entity mentions (as measured by cosine similarity of the embeddings of the entity mention strings) to be too similar; e.g., "United States" vs "U.S.". The reason is that we of course want the candidate to be a different entity. GETALTS returns a list of sets, each containing, for a particular alternative entity e_i, possible mention strings for e_i. For instance, if the entity node we want to replace is "United States", an example mention set is {"United Kingdom", "UK", "Britain"}. For more details see Appendix A.

To generate counterfactual documents from the seed document d, we loop over all entity nodes in d. We attempt to apply replacements for each entity node. To achieve this, we first create an empty dictionary – denoted as $\mathbb{D}$ – for our newly generated documents (each created through replacements). After having replaced an entity node, we add the resulting counterfactual document to this dictionary. We use EditTuple to record which nodes have been replaced, preventing any node from being replaced more than once. We repeatedly loop over the dictionary to gather a large number of counterfactual documents. After the replacement process is completed, we select those counterfactual documents that are affected by the replacement more than a threshold τ_r. Thus, we require that a valid counterfactual document have at least τ_r percent of their triples altered (in either one or both entities).

Algorithm 1 Counterfactual Example Generator

Input:

d: Document with entity nodes and relation triples

τ_r: Affected relations threshold—An augmented document should have more than τ_r of its relations affected by the replacements

M_N: Maximum number of alternatives to sample from

Output:

$\mathcal{D}$: A set of documents with entity replacements applied on d

Auxiliary functions:

REPLACE(i, d, alt): A function that replaces node (i) and its mentions in the document (d) with a given alternative entity mentions set (alt).[2]

AFFECTR(EditTuple, d): This function returns the ratio of relation triples that are affected by the replacements specified in the EditTuple.

GETALTS(e_i,$\mathbb{E}$,$\tau_{\text{e[MAX]}}$,$\tau_{\text{e[MIN]}}$,τ_{c}): This function returns a list of sets of alternatives for the given entity node e_i. It requires the candidates pool $\mathbb{E}$, and other sets of hyperparameters (cf. Appendix A).

Initialize $\mathbb{D} \leftarrow \{\}$, EditTuple $\leftarrow ()$, $\mathcal{D} \leftarrow [\,]$
$\mathbb{D}$[EditTuple] $\leftarrow d$
for EditTuple, $\tilde{d}$ in $\mathbb{D}$ **do**
    for e_i in $\tilde{d}$[EntityNodes] **do**
        if $i \notin$ EditTuple **then**
            $alts \leftarrow$ GETALTS(e_i,$\mathbb{E}$,$\tau_{\text{e[MAX]}}$,$\tau_{\text{e[MIN]}}$,τ_{c})
            Sample alt from $alts[: M_N]$
            Add i to EditTuple
            $\mathbb{D}$[EditTuple] $\leftarrow$ REPLACE(i,$\tilde{d}$,alt)
        end if
    end for
end for
for EditTuple, $\tilde{d}$ in $\mathbb{D}$ **do**
    if AFFECTR(EditTuple, d) > τ_r **then**
        Add $\tilde{d}$ to $\mathcal{D}$
    end if
end for
return $\mathcal{D}$

3 Experiments

We generate RE-DOCRED-CF, our counterfactual dataset, from Re-DocRED using COVERED. We run COVERED five times on Re-DocRED train to produce RE-DOCRED-CF train (so it consists of five different counterfactual datasets). We run COVERED on Re-DocRED test once to generate RE-DOCRED-CF test. We set $\tau_{\text{e[MAX]}} = .8$, $\tau_{\text{e[MIN]}} = .2$, $\tau_{\text{c}} = .4$, $M_N = 3$ (to limit the search) and (to make sure at least 70% of triples are affected by the replacements) $\tau_r = .7$.

We first evaluate the hypothesis that models that are trained merely on factual data do not reliably use the context for the RE task. To test this, we measure how consistent these models are for documents that have undergone entity replacement. We use the KD-DocRE framework (Tan et al., 2022a)[3] to train DocRE models. The framework features axial attention modules, adaptive focal loss and knowledge distillation over the distant supervised examples. As we want to observe the effect of using counterfactual data in training, we do not use knowledge distillation and only do their first stage of training over the human-annotated data.

We follow Tan et al. (2022a)'s setup and hyperparameters in finetuning a RoBERTa-large model (Liu et al., 2019) for relation extraction. First,

[2]For each replacement the closest mention to the original mentions, in terms of embedding similarity would be selected.

[3]https://github.com/tonytan48/KD-DocRE

Training Data	PRC	REC	F1	CONS
Re-DocRED	**88.1**	**69.8**	**78.0**	68.6
CF Only	85.3	62.8	72.4	**89.5**
Re-DocRED + CF	*85.7*	*68.8*	*76.3*	*88.3*

Table 1: Evaluation results on factual (Re-DocRED), counterfactual (CF Only) and combined (Re-DocRED + CF) data. Our measures are Precision (PRC), Recall (REC) and F1 score on Re-DocRED's test set. Using a counterfactual counterpart of the test set, we report consistency (CONS) results of each approach. (All reported numbers are the median over 5 runs with different random seeds.)

with the training set of Re-DocRED, we finetune a model that we probe for factual biases. To mitigate random errors, we train with five random seeds and report the median over each metric.

To assess a model's factual bias, we need to observe how its behavior changes when presented with counterfactual data. Following Paranjape et al. (2022), we use *pairwise consistency* as our measure. Pairwise consistency is the accuracy of the model on those counterfactuals whose factual counterparts (the original facts) were predicted correctly.

3.1 Results

Table 1 shows performance of the trained models on Re-DocRED test. As expected, the model that is only trained on factual data performs well on similar factual data. However, it only shows 68.6% consistency (for Re-DocRED test and RE-DOCRED-CF test) – more than 30% of the correctly predicted output is based on entity and factual biases. Figure 1 shows the original text (top) and the text after replacement of entities (bottom). We see that COVERED, our replacement algorithm, did a good job here: the original entities were replaced with similar entities (which occur in similar contexts and with similar relations), but of course the new triples are nonfactual. The model correctly predicted the triple (Cleanin' Out My Closet, part of, The Eminem Show) from the original document. However, for the document with replaced entities – even though the relation ("part of") is still the same, only the entities have changed – the model fails to extract the correct triple, which would be: (The Ultimate Collection, part of, London Calling). See Appendix B for more examples. This result corroborates other similar analysis that was done on the sentence level (Wang et al., 2022, 2023).

To evaluate the effectiveness of COVERED in generating plausible examples, we conducted a human evaluation on a sample subset of the data. We randomly selected 50 triplets from the test set and found that 45 of them were deemed plausible. This indicates that 90 percent of the counterfactual triplets accurately reflect relationships that are evident from the counterfactual version of the document.[4]

Our hypothesis is that we can increase robustness against entity and factual biases by training on counterfactual data. We finetune a separate model with its own separate random seed for each of the five parts of RE-DOCRED-CF train. Table 1 shows that consistency increases with a >20% gap compared to only using factual data (89.5% vs 68.6%). This shows that counterfactuals improve the model's robustness against entity and factual biases. However, training on counterfactuals only also deteriorates performance on factual test data (5.6 drop on F1, 72.4 vs 78.0). Since the real-world use case of DocRE models is factual data, we need to devise a solution that is both performant and consistent.

Therefore, we conduct a third experiment in which we mix each of the five subsets of RE-DOCRED-CF train with Re-DocRED train. To keep the number of training steps equal, we halve the number of epochs of training that we used in the other two experiments ($30 \rightarrow 15$). As shown in Table 1, the resulting model shows both a high performance with minimum drop in F1 (only -1.7, 76.3 vs 78.0) while also being consistent (88.3% vs 68.6% for the "factual-training-only" model). This means that the counterfactual RE-DOCRED-CF dataset helps the model to learn the task based on the context and mitigates bias issues while having the factual dataset alongside keeps the model performant on factual data.

4 Conclusion

In this work, we present a method for generating counterfactual examples for document-level relation extraction. Our approach searches for suitable entity replacements over a document and applies them to a point where most of the relations are affected by these replacements. By generating a counterfactual test set, we demonstrate the high

[4]For this evaluation, we excluded counterfactual examples where the original counterparts were already mislabelled (e.g., due to entity linking errors or misannotation of non-evident relations). This ensures that our analysis focuses on evaluating the plausibility of our pipeline on correctly labeled examples.

level of inconsistency DocRE model have when trained only with factual data. Adding counterfactuals to the training sets improves consistency by a large margin while keeping performance high. We make our pipeline COVERED and dataset RE-DOCRED-CF publicly available We hope our findings and resources will raise awareness and support future efforts in addressing entity and factual biases in document relation extraction.

Limitations

The main limitation of this work is its requirement of a seed DocRE dataset. This means to extend this approach to either other domains or languages we need a document RE dataset provided. Here, we measured consistency and performance levels on KD-DocRE, one of the recent and high-performing methods. However, other solutions might yield different performance results. Our aim in this work is to provide a document-level RE dataset for consistency. Also, the improvements in robustness against factual bias were gained in an equal setup.

Acknowledgements

This work was funded by Deutsche Forschungsgemeinschaft (Project SCHU 2246/14-1).

References

Haotian Chen, Bingsheng Chen, and Xiangdong Zhou. 2023. Did the models understand documents? benchmarking models for language understanding in document-level relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6418–6435, Toronto, Canada. Association for Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

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. *arXiv preprint arXiv:1907.11692*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Bhargavi Paranjape, Matthew Lamm, and Ian Tenney. 2022. Retrieval-guided counterfactual generation for QA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1670–1686, Dublin, Ireland. Association for Computational Linguistics.

Kun Qian, Ahmad Beirami, Zhouhan Lin, Ankita De, Alborz Geramifard, Zhou Yu, and Chinnadhurai Sankar. 2021. Annotation inconsistency and entity bias in MultiWOZ. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 326–337, Singapore and Online. Association for Computational Linguistics.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022a. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1672–1681, Dublin, Ireland. Association for Computational Linguistics.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022b. Revisiting DocRED - addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8472–8487, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hengzhu Tang, Yanan Cao, Zhenyu Zhang, Jiangxia Cao, Fang Fang, Shi Wang, and Pengfei Yin. 2020. Hin: Hierarchical inference network for document-level relation extraction. In *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part I 24*, pages 197–209. Springer.

Yiwei Wang, Muhao Chen, Wenxuan Zhou, Yujun Cai, Yuxuan Liang, Dayiheng Liu, Baosong Yang, Juncheng Liu, and Bryan Hooi. 2022. Should we rely on entity mentions for relation extraction? debiasing relation extraction with counterfactual analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3071–3081, Seattle, United States. Association for Computational Linguistics.

Yiwei Wang, Bryan Hooi, Fei Wang, Yujun Cai, Yuxuan Liang, Wenxuan Zhou, Jing Tang, Manjuan Duan, and Muhao Chen. 2023. How fragile is relation extraction under entity replacements? In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 414–423, Singapore. Association for Computational Linguistics.

Zhao Xiaoyan, Deng Yang, Yang Min, Wang Lingzhi, Zhang Rui, Cheng Hong, Lam Wai, Shen Ying, and Xu Ruifeng. 2023. A comprehensive survey on deep learning for relation extraction: Recent advances and new frontiers. *arXiv preprint arXiv:2306.02051*.

Nan Xu, Fei Wang, Bangzheng Li, Mingtao Dong, and Muhao Chen. 2022. Does your model classify entities reasonably? diagnosing and mitigating spurious correlations in entity typing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8642–8658, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14612–14620.

A Alternative Entities Search Algorithm

Algorithm 2 is the detailed pseudocode of our approach in finding suitable alternatives for an entity node.

B Extra Counterfactual Examples

In Figure 2, we demonstrated three other examples of a factual bias failure of a DocRE model. Some examples also include relations that are spanning across multiple sentences which a DocRE model should be capable to extract. However, after entity replacement the model (which is trained only on factual data) only manages to predict on the original set.

Original Document:
Tippy Walker (born February 19, 1947) is a former **American** child actress, best known for her role in the film The World of Henry Orient (1964). Born **Elizabeth Tipton Walker** in **New York City**, her father… After appearing in several television shows, such as Doctor Kildare and Peyton Place, and the female lead role in the film **Jennifer on My Mind**…

RE Model Prediction (TP): (Doctor Kildare, cast member, Tippy Walker)

Counterfactual Document w. Entity Replacement:
Elizabeth Sterling Haynes (born February 19, 1947) is a former **Canadian** child actress, best known for her role in the film The World of Henry Orient (1964). Born **Elizabeth** in **Amsterdam**, her father… After appearing in several television shows, such as Doctor Kildare and Peyton Place, and the female lead role in the film **Think Like a Man Too**…

Missed Prediction (FN): (Doctor Kildare, cast member, Elizabeth Sterling Haynes)

Original Document:
Watts Station is a train station built in 1904 in **Watts**, Los Angeles, **California**. It was one… and Long Beach. It was the only structure that remained intact when stores along 103rd Street in **Watts** were burned in the 1965 **Watts Riots**…

RE Model Prediction (TP): (103rd Street, located in the administrative territorial entity, California)

Counterfactual Document w. Entity Replacement:
Oslo Airport Station is a train station built in 1904 in **Alabama**, Los Angeles, **Costa Mesa**. It was one… and Long Beach. It was the only structure that remained intact when stores along 103rd Street in **Alabama** were burned in the 1965 **French Revolution**…

Missed Prediction (FN): (103rd Street, located in the administrative territorial entity, Costa Mesa)

Original Document:
James Whitman "Jim" McLamore (May 30, 1926 – August 9, 1996) was with David Edgerton responsible for the expansion of the **Burger King** fast food franchise… The pair sold the business to **Pillsbury** in 1967 and McLamore served as **Burger King**'s president until 1970 , and was…

RE Model Prediction (TP): (Burger King, owned by, Pillsbury)

Counterfactual Document w. Entity Replacement:
Ellison (May 30, 1926 – August 9, 1996) was with David Edgerton responsible for the expansion of the **Cheeseburger in Paradise** fast food franchise… The pair sold the business to **Orange S.A.** in 1967 and **Ellison** served as **Cheeseburger in Paradise**'s president until 1970, and was…

Missed Prediction (FN): (Cheeseburger in Paradise, owned by, Orange S.A.)

Figure 2: Three other examples of original documents and their counterfactual counterparts. In all three we observe a failure in predicting the counterfactual, while all information required for the relation to be extracted are present (Underlined).

Algorithm 2 Find Suitable Alternatives for a given Entity Node e_i

Input:

e_i: Input entity node

$\mathbb{E}$: Entity candidates pool

$\tau_{\text{e[MAX]}}$, $\tau_{\text{e[MIN]}}$: Maximum and minimum entity mention similarity threshold

τ_{c}: Context similarity threshold

Output:

$\mathcal{E}$: List of sets of alternative entity mentions for the given entity node

function GETALTS(e_i,$\mathbb{E}$,$\tau_{\text{e[MAX]}}$,$\tau_{\text{e[MIN]}}$,τ_{c})
Initialize: $\mathcal{E} \leftarrow [\,]$, $\mathcal{E}' \leftarrow [\,]$
for $\tilde{e}$ in $\mathbb{E}$ **do**
$r_{\text{sim}} = |\tilde{e}[\text{rel\_maps}] \cap e_i[\text{rel\_maps}]|$
if $r_{\text{sim}} = 0$ **then**
continue
else if $\tilde{e}[\text{doc\_title}] = e_i[\text{doc\_title}]$ **then**
continue
else if $\tilde{e}[\text{type}] \cap e_i[\text{type}] = \varnothing$ **then**
continue
end if
Set: $m_{\text{sim}} \leftarrow 0$
for m_i in e_i[mentions], $\tilde{m}$ in $\tilde{e}$[mentions] **do**
$\text{sim} = \cos(\tilde{m}[\text{emb}], m_i[\text{emb}])$
$m_{\text{sim}} \leftarrow \max(m_{\text{sim}}, \text{sim})$
end for
Set: $c_{\text{sim}} \leftarrow 0$
for c_i in e_i[contexts], $\tilde{c}$ in $\tilde{e}$[contexts] **do**
$\text{sim} = \cos(\tilde{c}[\text{emb}], c_i[\text{emb}])$
$c_{\text{sim}} \leftarrow \max(c_{\text{sim}}, \text{sim})$
end for
if $\tau_{\text{e[MIN]}} < m_{\text{sim}} < \tau_{\text{e[MAX]}}$ **and** $\tau_{\text{c}} < c_{\text{sim}}$ **then**
Add $(\tilde{e}, r_{\text{sim}}, m_{\text{sim}}, c_{\text{sim}})$ to $\mathcal{E}'$
end if
end for
Sort $\mathcal{E}'$ by $r_{\text{sim}}, m_{\text{sim}}, c_{\text{sim}}$
for $\mathbf{e}$ in $\mathcal{E}'$ **do**
Add $\mathbf{e}[0]$[mentions] to $\mathcal{E}$
end for
Drop any set in $\mathcal{E}$ that is a subset of another set in $\mathcal{E}$
return $\mathcal{E}$
end function