# What Causes Unemployment? Unsupervised Causality Mining from Swedish Governmental Reports

**Luise Dürlich**[1] **Joakim Nivre**[1,2] **Sara Stymne**[2]
[1]RISE Research Institutes of Sweden, Kista, Sweden
[2]Department of Linguistics and Philology, Uppsala University, Sweden
{luise.durlich,joakim.nivre}@ri.se, sara.stymne@lingfil.uu.se

## Abstract

Extracting statements about causality from text documents is a challenging task in the absence of annotated training data. We combine pattern matching of causal connectives with semantic similarity ranking, using a language model fine-tuned for semantic textual similarity, to create a search system for causal statements about user-specified concepts. Preliminary experiments on Swedish governmental reports, using a small manually annotated test set, show promising results in comparison to two simple baselines.

## 1 Introduction

Extracting causal relations from natural language text is a popular task that has been tackled using a variety of techniques (Ali et al., 2021; Yang et al., 2022). Most approaches are based on supervised machine learning and presuppose annotated training data, which is lacking for many languages and domains.

In a recent project dealing with the analysis of Swedish governmental reports, we were faced with the task of building a system that would allow analysts to search for statements about potential causes and/or effects related to specific concepts, such as *pollution* or *unemployment*. In such a scenario we need a system that can retrieve sentences that make causal claims involving the specific concepts and rank them by relevance to the original query.

Since there were no pre-existing data sets with causality annotation available for Swedish, we could not make use of supervised learning but instead had to explore a combination of techniques involving keyword matching and pre-trained language models fine-tuned for semantic textual similarity (STS). The idea is to use keywords corresponding to causal connectives – such as the verb *cause* or the prepositional expression *because of* – to construct templatic sentences with masked tokens corresponding to the sought cause or effect, for example: "**MASK** causes pollution", "unemployment because of **MASK**". The textual semantic similarity model can then be used to find the sentences in a document collection that are semantically similar to the template sentences and are likely to include instantiations of the sought cause or effect.

For evaluation, we created a small test dataset for ranking causal sentences, previously presented in Dürlich et al. (2022). Preliminary experiments show that the unsupervised method combining keyword matching and semantic similarity search improves over two simple baselines.

## 2 Task and Approach

Causality mining refers to a broad class of tasks that involve extracting information about causality from natural language text. The specific task addressed in our project can be defined as follows: Given a document collection and an input query specifying a potential CAUSE, a potential EFFECT, or both, return a list of sentences describing causal relations matching the query, ranked in order of decreasing relevance. For example, if the input query is [CAUSE: pollution], the system should return sentences describing causal effects of pollution; if the query is [EFFECT: unemployment], the system should return sentences talking about the causes of unemployment; and if the query is [CAUSE: recession, EFFECT: unemployment], the system should return sentences discussing whether recession causes unemployment.

Since we do not have access to annotated training data for this task, we instead leverage a pretrained masked language model tuned for STS. The main idea is to first convert the query to one or more *query prompts*, that is, templatic sentences with masked tokens corresponding to empty slots, and then search for semantically similar sentences

| Causality keywords | English translations |
|---|---|
| bero på | depend on / be due to |
| bidra till | contribute to |
| leda till | lead to |
| på grund av | because of / due to |
| till följd av | due to / as a consequence of |
| vara ett resultat av | be a result of |
| framkalla | induce / evoke |
| förorsaka | cause |
| medföra | entail / involve |
| orsaka | cause |
| påverka | affect / influence |
| resultera | result |
| vålla | cause / inflict |
| därför | therefore / consequently |
| eftersom | because |
| effekt | effect |
| följd | consequence |
| orsak | cause |
| resultat | result |
| förklara | explain |
| rendera | render |

Table 1: Causality keywords. The top 13 keywords were selected to be used in our main data sets.

in the document collection. For example, the query [CAUSE: pollution] could be converted to a query prompt such as "**MASK** causes pollution" or "pollution is the result of **MASK**" with the hope that semantically similar sentences make claims about specific phenomena that cause pollution. In the following, we describe the creation of query prompts and the semantic search procedure in more detail. A key component in both is a set of *causality keywords*, which are used both to create template sentences and to filter sentences in the search procedure. We therefore begin by describing the selection of causality keywords.

### 2.1 Causality Keywords

We initially proposed a set of 21 causality keywords including both single words and multi-word expressions that typically convey causal relations. To evaluate these keywords, we performed a small annotation study to investigate how often sentences containing these expressions were considered causal. For each of the 21 keywords, we randomly extracted 10 sentences from the SOU-corpus, described in Section 3.1, allowing inflectional variants. The three authors annotated these sentences as causal, non-causal, or uncertain, without the use of any specific guidelines.

Based on the results, we excluded keywords that either tended to be ambiguous or to refer to causality in a more abstract or hypothetical manner, for example, without relating to any specific cause or effect. The final set of 13 keywords, in the upper section of Table 1, very frequently expressed causality, while the remaining 8 keywords, which include all nouns, had a lower proportion of causal sentences.

### 2.2 Query Prompt Generation

Based on the 13 keywords, we define a set of 15 query prompt templates, in which the position of cause and effect are made explicit by defining two distinct slots around the keywords – the two multi-word prepositions *på grund av* and *till följd av* each map to two very similar versions of this, one just adding the slots directly around the keyword ("CAUSE because of EFFECT") and one adding in the verb *händer* ("CAUSE happens because of EFFECT"), whereas all other keywords only produce a single template. For each query we generate 15 prompts by filling in one or both of the slots in the prompt template. If only one of cause and effect is defined, we replace the missing slot with the **MASK** token.

### 2.3 Semantic Similarity Search System

A first step in preparing the search is applying the causality keywords to filter the text collection of interest. The filtered text collection, which we assume now only contains sentences mentioning causality, is embedded sentence by sentence using the Swedish STS model trained using the contrastive tension (CT) technique by Carlsson et al. (2021), which had given state-of-the-art performance for English STS at the time when the experiments were run.

CT evades the issue of limited training data for STS tasks by focusing on the contrast between completely identical and randomly matched sentences, which allows for the automatic creation of large training data sets. Two instances of the same pretrained language model – KB-BERT (Malmsten et al., 2020) in our case – are trained jointly to each embed a sentence in the pair and to maximise the dot product between the sentence representations for identical sentences and minimize it for the random pairs. The CT model used here is the one performing better during evaluation on SentEval (Conneau and Kiela, 2018) machine-translated to Swedish.

We store the sentence embeddings generated by the CT model along with document and section IDs for each sentence and fit a nearest neighbour model to the embeddings. Once a user specifies a search

| | |
|---|---|
| Sentence 1 | *Flera av teknikerna bedöms resultera i långsiktig inbindning av koldioxid.*<br>'Several of the techniques are considered to result in long-term sequestration of carbon dioxide.' |
| Sentence 2 | *Exempelvis ger koldioxidutsläpp inga lokala skador, utan bidrar till växthuseffekten.*<br>'For example, carbon dioxide emissions do not cause local damage, but contribute to the greenhouse effect.' |

Figure 1: Example of a sentence pair to be ranked for the query [CAUSE: greenhouse effect].

query, it is converted into a query prompt and embedded by the CT model. The nearest neighbour model provides us with 300 candidates per prompt in terms of cosine distance. To get a combined ranking for all 15 prompts we sum the individual cosine distances of each neighbour over all prompts – the underlying assumption being that a relevant sentence should rank highly for all prompts – and rank the resulting list by ascending distance.

Note that the CT model itself is not fine-tuned for causality, which is why we restrict the nearest neighbour model to only consider sentences containing one of the previously established causal keywords. Without this restriction, the broad notion of semantic similarity captured by the model would include many non-causal statements that share some (other) aspect of meaning with the query prompts.

## 3 Data Sets

In this section we briefly describe the corpus of Swedish governmental reports that we used in our experiments and the test set for causality ranking, which is based on the same corpus. For more details on the corpus and data set creation we refer to Dürlich et al. (2022).

### 3.1 Governmental Report Corpus

The texts used in this work are the Swedish Government Official Reports, *Statens offentliga utredningar* (SOU) in Swedish, a series of reports with the goal of introducing legislative proposals and investigating complicated matters in the legislative process. We used reports from 1994–2020, available from Riksdagen.[1] Unfortunately the available documents focused on page layout and not on document structure, and the reports did not use a standardized format. We thus pre-processed the reports to extract titles and paragraphs. The resulting SOU-corpus of 3,558 reports and 3,434 summaries is publicly available.[2]

### 3.2 Causality Ranking Test Set

To evaluate the system we should ideally have a list of (manually) ranked sentences for each test query, but to create such lists would be a very difficult annotation task. To make the annotation easier and more reliable, we therefore let annotators classify pairs of sentences for relevance and rank the two sentences internally. Our annotation scheme encompasses six categories covering the case of none of the two sentences being relevant, both being equally relevant, only the first/second being relevant, and the first/second being more relevant than the other.

Figure 1 gives an example of a ranking pair extracted for the query [CAUSE: greenhouse effect]. In this example both of the sentences are considered relevant, but the second sentence is considered more relevant since it explicitly mentions *greenhouse effect*. We extracted the ranking data set using a set of 15 cause and effect pairs like *radon/cancer* and *unemployment/crime*, which were used to construct queries involving only the cause, only the effect, or both.

To find relevant text passages, we used a simplified version of the system described in Section 2.3 and selected the 500 nearest neighbours to the corresponding query prompts. Instead of combining 15 separate prompts, as in the full system, we only sampled from the neighbors to a single prompt based on the template "CAUSE medför (entails) EFFECT". We obtained pairs of sentences by randomly sampling the neighbours with replacement. In total, the test set consists of 800 sentence pairs and their ranked relevance with respect to 43 unique causal prompts.

## 4 Experiments

In this section we compare our system, where we rank causal sentences with CT models, to two baselines: **Random**, which just randomly shuffles the sentences we consider for ranking, and **KB-BERT**, where sentence embeddings are obtained by mean-pooling the hidden states of KB-BERT. For the **Random** baseline we take the average of 10 different random seeds. Besides the CT model trained

[1] https://riksdagen.se/sv/dokument-lagar/?doktyp=sou

[2] https://github.com/UppsalaNLP/SOU-corpus

| Model | p@5 | p@10 | MAP | ACC |
|---|---|---|---|---|
| Baseline (Random) | 0.40 | 0.51 | 0.41 | 0.51 |
| Baseline (KB-BERT) | 0.49 | 0.49 | 0.43 | 0.51 |
| RISE CT | 0.57 | 0.60 | 0.55 | 0.62 |
| CT SOU only (1) | 0.55 | 0.63 | 0.53 | 0.66 |
| CT SOU only (2) | 0.59 | 0.61 | 0.56 | 0.65 |
| RISE CT + SOU (1) | **0.60** | **0.64** | **0.57** | **0.70** |
| RISE CT + SOU (2) | 0.59 | 0.62 | 0.55 | 0.66 |

Table 2: Ranking results using different kinds of semantic representations. The best result for each metric is marked in bold.

by RISE, we also train two additional in-domain CT models, which we describe next.

### 4.1 Domain-specific CT training

In addition to the pre-trained CT model, we investigated the effect of fitting the model on in-domain data from the SOU-corpus. We consider two approaches. The first one (CT SOU only) is to train KB-BERT with the CT objective on SOUs only. The second (RISE CT + SOU) is to run another CT training round with the Swedish model by Carlsson et al. (2021) (RISE CT) on the SOUs. Since we only had one of the two models available, we initialised both models as RISE CT. The data for both variants is sampled from the filtered sentences in the SOU-corpus. At each epoch during training we validate both models on SentEval and stop training as soon as the validation performance drops. We report ranking results for both models trained in a single CT training session.

### 4.2 Evaluation

During evaluation, we do not fit a full nearest neighbour model, but simply take the cosine distances between the set of annotated sentences per query in the test set and the respective query. We evaluate the ranking using the following evaluation metrics:

**Precision at *k* (P@k):** The number of relevant sentences among the top *k* nearest neighbours. Here we exclude queries with less than *k* relevant sentences.

**Mean average precision (MAP):** The mean of the average precision over all 43 queries in the test set.

**Accuracy (ACC):** the percentage of sentence pairs where the model ranks the pair consistently with the human ranking – not including pairs where the sentences were considered equally (ir)relevant. For P@k and MAP, we converted the pair-wise human relevance judgements in the test set into binary scores over the set of matched sentences per query. That is, we considered all sentences that had been judged as relevant, even when they were considered less relevant than another sentence, as relevant, and all other sentences as irrelevant.

### 4.3 Results

Table 2 shows the results of the evaluation. It can clearly be seen that all CT-models perform better than both baselines. The RISE CT + SOU (1) achieves the best results in all five metrics, followed closely by both its partner model and CT SOU only (2). While the domain-specific training seems to have helped somewhat, the difference to the RISE CT-model is quite small. We find it interesting that training CT only on the small in-domain SOU-corpus (CT SOU only), is at least as good as the original CT-model trained on a much larger out-of-domain Wikipedia corpus. KB-BERT performs either slightly worse than the random baseline or only marginally better, clearly not being a good fit for this task.

Our results indicate that around six out of ten matches in a ranked list would be relevant. We think this can be useful in our target scenario with a human in the loop, but there is still room for improvement. For instance, when testing the search system and during annotation, we noticed that the system often confused the roles of causes and effects, an issue that can be addressed in future work.

## 5 Conclusion

In this paper we describe an initial exploration on causality mining with respect to specific concepts, such as *pollution* or *unemployment*, in Swedish governmental reports. We present the task in detail, and note that there is no available training data. We thus design a search system based on the combination of keyword matching and semantic similarity ranking, which can give reasonable results for a human-in-the-loop scenario.

Although the preliminary results look promising, further evaluation on a larger test set as well as on other document collections will be needed to assess the viability of the approach. It would also be interesting to explore whether syntactic or semantic parsing could be used to improve the model's capacity to distinguish the direction of causality and prevent the confusion of causes and effects.

## Acknowledgements

This work was funded by Vinnova in the project 2019-02252: Datalab for results in the public sector. We thank Sven-Olof Junker, Martin Sparr, Fredrik Carlsson, Sebastian Reimann and Gustav Finnveden for valuable discussions. The computations were enabled by resources in project UPPMAX 2020/2-2 at the Uppsala Multidisciplinary Center for Advanced Computational Science.

## References

Wajid Ali, Wanli Zuo, Rahman Ali, Xianglin Zuo, and Gohar Rahman. 2021. Causality mining in natural languages using machine and deep learning techniques: A survey. *Applied Sciences*, 11.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2021. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations*.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Luise Dürlich, Sebastian Reimann, Gustav Finnveden, Joakim Nivre, and Sara Stymne. 2022. Cause and effect in governmental reports: Two data sets for causality detection in Swedish. In *Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences*, pages 46–55, Marseille, France. European Language Resources Association.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the National Library of Sweden - making a Swedish BERT. *CoRR*, abs/2007.01658.

Jie Yang, Soyeon Caren Han, and Josiah Pong. 2022. A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems*, 64:1161–1186.