What Causes Unemployment? Unsupervised Causality Mining from Swedish Governmental Reports

Luise Dürlich[1,2] **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 create a search system for causal statements about user-specified concepts by combining pattern matching of causal connectives with semantic similarity ranking, using a language model fine-tuned for semantic textual similarity. Preliminary experiments on a small test set from Swedish governmental reports 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 to causality mining are based on supervised machine learning and presuppose annotated training data, which is lacking for many languages and domains. In this paper, we describe a system for exploring Swedish governmental reports, where users can search for statements about potential causes and/or effects related to specific concepts, such as *pollution* or *unemployment*. The system should then retrieve sentences that make causal claims involving the specific concepts and rank them by relevance to the original query.

In order to delimit the scope of causality mining, we follow the approach of Dunietz et al. (2015) and focus on causality that is explicitly expressed linguistically, by the use of some causal connective, rather than implicitly expressed causality. A causal connective is any type of linguistic expression that is used to express a causal relation, for instance, verbs like *cause*, conjunctions like *because*, nouns like *effect*, and different types of multi-word expressions like *be a result of*. While there is plenty of work on the identification of causal sentences or entities, and the extraction of causal relations from text (for an overview, see Yang et al., 2022), we are not aware of any work focusing on the ranking of extracted causal sentences. It could be viewed as an information retrieval task, for which work is abundant (see, e.g., Mitra and Craswell, 2018). There are also efforts for more specific structured queries, such as Taub Tabib et al. (2020).

Most approaches to causality mining are based on supervised machine learning and presuppose annotated training data, which is available for languages like English (Hendrickx et al., 2010; Mariko et al., 2020; Mirza et al., 2014; Prasad et al., 2008; Mihăilă et al., 2016) and German (Rehbein and Ruppenhofer, 2020). Since the single annotated data set available for Swedish (Dürlich et al., 2022) is much too small to use as training data, we instead 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 STS model, in our case contrastive tension (Carlson et al., 2021), 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. We believe that this system design can be useful for other tasks where a domain is under-resourced for the target language.

We evaluate our approach on a small test set for ranking causal sentences (Dürlich et al., 2022). Preliminary experiments show that the unsupervised method combining keyword matching and semantic similarity search improves over two simple baselines. As far as we are aware, these are the first published results for this task and data set.

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.

Facing the lack of annotated training data for this task, we instead leverage a pre-trained 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 in the document collection. For example, the query [CAUSE: pollution] could be converted to a query prompt such as "pollution causes **MASK**" or "**MASK** is the result of pollution" with the hope that semantically similar sentences make claims about specific phenomena caused by 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 set of causality keywords used.

2.1 Causality Keywords

For our approach, we need a set of causality keywords, or connectives. These are linguistic expressions of causality, which could include verbs, phrasal verbs, prepositions, and other types of expressions. We choose the set of causality keywords previously used for the creation of the causality test set (Dürlich et al., 2022). Those keywords were selected from an initial pool of 21 candidate expressions. For each of these expressions, a set of candidate sentences containing them were extracted, and annotated by three annotators, without any specific guidelines. Based on the annotation, a set of 13 keywords that consistently expressed causal relations were selected, shown in Table 1.

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

Table 1: Causality keywords (Dürlich et al., 2022)

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 multiword 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 our project was started.

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 pre-trained language model – KB-BERT

Sentence 1	*Flera av teknikerna bedöms resultera i långsiktig inbindning av koldioxid.* '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.* '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 [EFFECT: greenhouse effect] (Dürlich et al., 2022).

(Malmsten et al., 2020) in our case – are trained jointly to each embed a sentence in the pair and to maximize 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 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 other aspects of meaning with the query prompts.

3 Data Sets

We use a previously published test set (Dürlich et al., 2022), which includes sentences from the Swedish Government Official Reports, *Statens offentliga utredningar* (SOU) in Swedish, from 1994–2020. For more details on the corpus and data set creation, we refer to the original paper.

Ideally we would evaluate a full list of ranked sentences for each test query. Instead, the test set frames an easier annotation task: to classify pairs of sentences for relevance and rank the two sentences internally. For a sentence pair, like the one in Figure 1, the task was to assess their relevance to a given query. The annotation scheme has six categories covering the following cases for sentences 1 and 2:

1. 1 and 2 are equally relevant.
2. 1 and 2 are both relevant but 1 more so.
3. 1 and 2 are both relevant but 2 more so.
4. 1 is relevant but 2 is not.
5. 2 is relevant but 1 is not.
6. 1 and 2 are both irrelevant.

In the example in Figure 1, both sentences are considered relevant, but the second more so, since it explicitly mentioned the term *greenhouse effect* from the query; hence it is classified as case 3. The test set consists of 800 sentence pairs and their ranked relevance with respect to 43 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** (Malmsten et al., 2020), 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 original CT model, CT-Orig, released by Carlsson et al. (2021), initialized with KB-BERT, and then trained on Wikipedia data with the CT objective, we also train two additional in-domain CT models, which we describe next.

4.1 Domain-Specific CT Training

We investigate the effect of fitting the CT model on in-domain data from the SOU corpus, considering two approaches. For the first one (CT-SOU), we initialize it with KB-BERT and then only run the CT objective on a subset of the SOUs – only sentences containing causal connectives. For the second one (CT-Orig+SOU) we instead initialize with CT-Orig, and then run another round of CT training round on the SOUs subset. The subset of SOUs contained 490K sentences (14M tokens). Since we only had

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
CT-Orig	0.57	0.60	0.55	0.62
CT-SOU$_1$	0.55	0.63	0.53	0.66
CT-SOU$_2$	0.59	0.61	0.56	0.65
CT-Orig+SOU$_1$	**0.60**	**0.64**	**0.57**	**0.70**
CT-Orig+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.

one of the two original models available, we initialized both models as CT-Orig. 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 judgments 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. For the two domain-specific CT-models, both instantiations from CT training are shown (with subscripts). It can clearly be seen that all CT-models perform better than both baselines. The CT-Orig+SOU$_1$ achieves the best results in all five metrics, followed closely by both its partner model and CT-SOU$_2$. While the domain-specific training seems to have helped somewhat, the difference to the original CT model (CT-Orig) is quite small. We find it interesting that training CT only on the small in-domain SOU corpus (CT-SOU) 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 it leaves room for improvement. For instance, we noticed that the system often confused the roles of causes and effects, an issue that can be addressed in future work.

5 Conclusion

We describe an initial exploration of 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. This work can be viewed as a first step towards enabling impact assessment of Swedish governmental reports. Our system for ranking causal sentences with respect to a theme could potentially feed into more advanced systems for impact assessment, for instance with the goal of exploring trends across sources and over time.

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. Another direction would be to use less aggressive methods than causal keywords for filtering causal sentences. One possibility could be to utilize available data from other languages, to train a cross-lingual model for identifying Swedish causal sentences, as proposed in Reimann and Stymne (2022).

Acknowledgments

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(21).

Keith Carlson, Allen Riddell, and Daniel Rockmore. 2021. Unsupervised text style transfer with content embeddings. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 226–233, Held Online. INCOMA Ltd.

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 Language Resources Association (ELRA).

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2015. Annotating causal language using corpus lexicography of constructions. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 188–196, Denver, Colorado, USA. Association for Computational Linguistics.

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.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with Words at the National Library of Sweden – Making a Swedish BERT. arXiv:2007.01658v1 [cs.CL].

Dominique Mariko, Hanna Abi-Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. 2020. The financial document causality detection shared task (FinCausal 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 23–32, Barcelona, Spain (Online). COLING.

Claudiu Mihăilă, Tomoko Ohta, Sampo Pyysalo, and Sophia Ananiadou. 2016. BioCause: Annotating and analysing causality in the biomedical domain. *BMC Bioinformatics*, 14(2).

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the TempEval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19, Gothenburg, Sweden. Association for Computational Linguistics.

Bhaskar Mitra and Nick Craswell. 2018. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval*, 13(1):1–126.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Ines Rehbein and Josef Ruppenhofer. 2020. A new resource for German causal language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5968–5977, Marseille, France. European Language Resources Association.

Sebastian Reimann and Sara Stymne. 2022. Exploring cross-lingual transfer to counteract data scarcity for causality detection. In *Proceedings of the Web Conference 2022 (WWW '22 Companion); The 3rd International Workshop on Cross-lingual Event-centric Open Analytics (CLEOPATRA 2022)*, Virtual Event, Lyon, France.

Hillel Taub Tabib, Micah Shlain, Shoval Sadde, Dan Lahav, Matan Eyal, Yaara Cohen, and Yoav Goldberg. 2020. Interactive extractive search over biomedical corpora. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 28–37, Online. Association for Computational Linguistics.

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.