Proceedings of the First Workshop on Natural Language Processing for Political Sciences (PoliticalNLP), pages 46–55
Marseille, 24 June 2022

Cause and Effect in Governmental Reports: Two Data Sets for Causality Detection in Swedish

Luise Dürlich[1,2], Sebastian Reimann[1,3], Gustav Finnveden[1], Joakim Nivre[1,2], and Sara Stymne[1]
[1]Department of Linguistics and Philology, Uppsala University, Sweden
[2]RISE Research Institutes of Sweden, Kista, Sweden
[3]Department for German Language and Literature, Ruhr-Universität Bochum, Germany
luise.durlich@ri.se, sebastian.reimann@ruhr-uni-bochum.de
gustav.finnveden@gmail.com, {joakim.nivre,sara.stymne}@lingfil.uu.se

Abstract

Causality detection is the task of extracting information about causal relations from text. It is an important task for different types of document analysis, including political impact assessment. We present two new data sets for causality detection in Swedish. The first data set is annotated with binary relevance judgments, indicating whether a sentence contains causality information or not. In the second data set, sentence pairs are ranked for relevance with respect to a causality query, containing a specific hypothesized cause and/or effect. Both data sets are carefully curated and mainly intended for use as test data. We describe the data sets and their annotation, including detailed annotation guidelines. In addition, we present pilot experiments on cross-lingual zero-shot and few-shot causality detection, using training data from English and German.

Keywords: test analysis, causality, causality detection, annotation, cross-lingual transfer

1. Introduction

The analysis of large volumes of text is an important task for political scientists and governmental agencies. In our project the end goal is to enable impact assessment of governmental reports, where the identification of causal relations is a key element. One scenario in this area is that a user wants to investigate potential causes and/or effects related to a specific concept, such as unemployment or pollution. In such a scenario we need a system that can rank matches mentioning a causal relationship with respect to a given concept. A more basic task is binary relevance classification of sentences with respect to causality, which can feed into a more advanced system. In this paper we focus on creating data sets for causality detection, enabling the development of methods for causality detection and ranking, which in turn can feed into more ambitious projects on impact assessment.

Our focus is on Swedish governmental reports. While these reports are publicly available, they are not available in a format directly suitable for text processing, since the focus is on page layout rather than document structure. We release a processed version of this corpus, with extracted texts. One additional obstacle faced in this project was the lack of annotated data for causality detection, since there were no previously available data sets for Swedish. We have addressed this lack of data by annotating two small data sets for Swedish causality detection, which we present in this paper. The data sets are carefully curated, with the main purpose to serve as test data. We focus on two different subtasks. The first is binary identification of sentences as causal or non-causal. The second is a ranking task with respect to a query sentence containing a given cause and/or effect, such as *traffic causes pollution* or *X causes cancer*, where the task is to decide which of a pair of extracted sentences is more relevant to the query. We focus on the sentence level, using sentences as the unit for identification and ranking. All data sets are based on sentences from the processed corpus of Swedish governmental reports and are publicly available under the CC BY 4.0 license.[1]

There are a few data sets available for other languages, like English (Mariko et al., 2020) and German (Rehbein and Ruppenhofer, 2020). However, these data sets were created for different purposes, with different label sets, granularity, and guidelines. Despite this, they are ideal to use for experiments on cross-lingual causality detection. We report results from pilot experiments on binary causality detection with zero-shot transfer into Swedish, showing how we can handle variations of the annotation schemes of these resources. In addition we investigate a few-shot scenario where we add a limited amount of Swedish training data. We leave experiments on ranking causal sentences to future work.

2. Related Work

As noted by Dunietz et al. (2015), causality is a complex topic, which has been discussed in many fields, including psychology and philosophy. In this work, we follow the approach of Dunietz et al. (2015) to focus only on causality which is explicitly expressed linguistically, by the use of some causal connective. A causal connective is any type of linguistic expression that is used to express a causal relation, for instance, verbs

[1]https://github.com/UppsalaNLP/Swedish-Causality-Datasets

like *cause*, conjunctions like *because*, nouns like *effect*, and different types of multi-word expressions like *be a result of*. This can be contrasted to some other annotation schemes, such as Girju (2003a), who rely on a common sense intuition of real-world causality.

There are some data sets annotated for causality available for other languages than Swedish. SemEval-2010 Task 8 (Hendrickx et al., 2010) focuses on classifying semantic relations between pairs of entities. It has nine different classes, of which cause-effect is one. Examples were collected using a pattern-based web search, with a high number of patterns per class. Each example was annotated by two annotators, followed by a consolidation phase. The data sets for the FinCausal 2020 shared task (Mariko et al., 2020) on the other hand concentrate exclusively on causal relations. They provide data sets for two subtasks: binary labelling of examples as causal or not, and extraction of causes and effects. The examples for both subtasks were taken from financial news. The annotation scheme only considers examples where the effect is a quantitative fact, which is a stricter definition than in other data sets. Examples were first annotated by a single annotator, then revised and discussed by two additional annotators until agreement.

Rehbein and Ruppenhofer (2020) provide a data set for causality in German. Their annotation scheme is an extension of Dunietz et al. (2015). They focus on causal language, only considering relations that are signaled by some causal connective. The annotations are on the token level according to the participant roles in a causal relation (cause, effect, actor, affected) and the types of causation (consequence, motivation, purpose). Each example was annotated by at least two annotators, and in a final phase all disagreements were resolved by two expert annotators. There are also other annotation efforts targeting causal relations among other types of relations. Mirza et al. (2014) annotate both temporal and causal relations between events in the TempEval-3 corpus. The Penn Discourse treebank includes annotations of causal discourse relations (Prasad et al., 2008). Mihăilă et al. (2016) describe an annotation effort for causal relations in biomedical texts.

While we are not aware of any work focusing on cross-lingual causality detection, there is some work on identifying discourse connectives based on parallel corpora and word alignments (Rehbein and Ruppenhofer, 2017; Versley, 2010). However, work on monolingual causality detection is more abundant, much of it focusing on English. Early work used rule-based methods (Garcia, 1997), decision trees (Girju, 2003b), and SVMs (Hendrickx et al., 2010). As for many other tasks, neural networks have recently become dominant. For the recent FinCausal shared task, the most common approach was based on pretrained language models. While the best models used ensembling architectures (Gordeev et al., 2020), also the simpler baseline model based on only an English BERT-based model had a strong performance (Mariko et al., 2020). While cross-lingual learning has not been used for causality detection, there has been much work on other tasks. A viable approach to many tasks is to fine-tune a pretrained language model on task data from some transfer language, which can then be applied in a zero-shot setting to some other language (Wu and Dredze, 2019; Conneau et al., 2020). Adding even a little bit of target language data, in a few-shot setting, can often improve the results considerably (Lauscher et al., 2020).

3. Data Processing

In this section we describe the creation of a corpus of Swedish governmental reports, which was the data source for the causality data sets. We describe the preprocessing and sentence segmentation of this corpus. We also describe the definition of a set of Swedish causality keywords, which were evaluated based on an initial annotation effort.

3.1. Governmental Report Corpus

The source of all our data sets is a subset of 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. These are typically produced by either a committee or a single investigator appointed by the Swedish government. At the time of extraction, only a subset of the reports were available digitally in PDF and HTML format, covering mostly reports from 1994 to the present (fall 2020).[2]

We extracted the HTML versions with the intent of exploiting the structure of the markup. However, the HTML markup turned out to encode style elements focused on describing the page layout rather than document structure. Elements like titles, subtitles, headers, footers and larger structures such as tables of contents or lists were not identifiable as such through the HTML markup, although their font type and size were encoded through style attributes. This also meant that paragraphs of text were often split in half by headers and footers at page boundaries. Another issue with this representation was the formatting of running text in parallel columns in certain sections, where the text sections mainly appeared line by line from left to right rather than as blocks of text representing a column at a time. This type of formatting largely appeared in sections concerning legislative proposals and was used to present a revised wording of the law in one column, with the previous version for reference in the other column, and the two columns thus often contained two very similar pieces of text on the surface. Since this was challenging to process in a way that produced cohesive text and was also likely to introduce near duplicates in the data if it had worked as intended, such text was omitted from the final documents.

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

To deal with this form of markup varying in layout and style between documents, we conducted a rule-based extraction to distinguish between running text and structural elements. In this process, only text and corresponding titles were kept and saved as HTML. Most documents[3] are preceded by a summary of their content, which we chose to split from the main document and save as a standalone file. These summaries could be written in English, simplified Swedish, or regular Swedish. Concerning the actual reports, some included sections written in other languages. The extraction script included a language detection part using langdetect[4], to verify that a given section was in Swedish. All text classified as non-Swedish was omitted. In some cases, the extracted text contained additional white space, which made it difficult to disambiguate hyphenation from cases of word-wrapping. The resulting corpus of 3,558 reports and 3,434 summaries is publicly available.[5] We refer to this corpus as the SOU corpus.

3.2. Sentence Extraction

From the SOU corpus, we sample individual sentences or sentence pairs to create the two data sets. To extract text samples for annotation, we split the text paragraphs from our cleaned HTML corpus into sentences. We segmented the text into sentences using a combination of SpaCy pipelines (Honnibal et al., 2019) for Swedish[6] and some rules to correct for frequent errors such as unrecognised sentence boundaries for abbreviations at the end of the sentence and issues with possessive or plural marking for acronyms, which are typically preceded by a colon (e.g. *SOU:er* 'SOUs') that were generally treated as a sentence boundary by the pipeline.

3.3. Causality Keywords

A first step was to define a set of causality keywords, to be used in the remainder of the project. Causality keywords correspond to causal connectives. We proposed a set of 21 causality keywords including single words and multi-word expressions that typically convey causal relations, shown in Table 1. To evaluate which of these expressions typically express causality, we performed a small annotation to investigate how often sentences containing these expressions were considered causal. This was a quick annotation effort by three annotators, without specific guidelines. This data set, which we call the binary trial data set, could then also be used as additional Swedish training data in a cross-lingual setting.

For each of the 21 keywords, we randomly extracted 10 sentences from the SOU corpus. Inflections of the terms were generated using the inflector provided by the Granska tool for Swedish grammar checking (Domeij et al., 2000). Multi-word terms were matched with at most two words in between each individual word.[7] Three experts annotated the sentences as causal, non-causal, or uncertain, without the use of any specific guidelines. The resulting data set contains 210 examples with annotations from three annotators.

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.

The main purpose of this annotation was to identify a set of keywords that reliably expresses causal relations. We thus excluded keywords that either tended to be ambiguous or to refer to causality in a more abstract or hypothetical manner, for example, without really relating to any specific cause or effect. The final set of 13 keywords, the top 13 terms in Table 1, very frequently expressed causality. The remaining 8 keywords had a lower proportion of causal sentences. Note that all nouns are in this group. The selected 13 causality keywords are verbs (e.g. *orsaka* 'cause'), phrasal verbs (e.g. *leda till* 'lead to'), multi-word prepositions (e.g. *till följd av* 'due to'), and one verbal multiword expression (*vara ett resultat av* 'be a result of').

4. Causality Data Sets

In this section we describe the two curated data sets created in the project, which are briefly summarized in Table 2. For both data sets we watned to include some additional context to the annotators, and thus included

[3]Some SOUs are divided into multiple parts and span multiple documents.

[4]https://pypi.org/project/langdetect/

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

[6]https://github.com/Kungbib/swedish-spacy

[7]We found that longer distances between the different parts of a term often did not match the correct structure but rather unrelated cases, where *till* and *på* acted as prepositions rather than verb particles.

Data set	Extraction	Annotators/ex	Size	%causal
Ranking	Causality keywords	2–3	800	–
Binary	Cause/effect pairs	2–3	330	48.5

Table 2: Overview of the causality data sets

Cause		Effect	
avskogning	deforestation	växthuseffekt	greenhouse effect
klimatanpassning	climate change adaptation	investeringsbehov	investment needs
klimatförändring	climate change	*investering*	investment
befolkningstillväxt	population growth	bostadsbrist	housing shortage
befolkningsmängd	population size	konsumtion	consumption
biltrafik	car traffic	luftförorening	air pollution
åskväder	thunderstorm	villabrand	house fire
regnväder	rainy weather		
reporäntan	bank rate	bolånekostnad	mortgage cost
arbetslöshet	unemployment	brottslighet	crime
utbildningsnivå	level of education	inkomst	income
rökning	smoking	blodtryck	blood pressure
droger	drugs	missbruk	abuse
radon	radon	cancer	cancer
luftföroreningar	air pollution	sjukdomar	diseases

Table 3: Cause and effect pairs. Terms marked with italics are alternatives to the original term above it, and 'rainy weather' was used only as a cause, not paired with an effect.

four context sentences, two before the target and two after. The annotators focused on the target sentences, but could use the context sentences for disambiguation when needed. The final data sets include the context sentences. The annotators are the authors of this paper, who are either native speakers of Swedish or native speakers of German with a good command of Swedish. For each data set, a subset of three annotators worked on it, always including two native Swedish speakers.

4.1. Binary Data Set

The binary data set is designed with the task of binary causality detection in mind. Specifically, the task is to decide on the sentence level, whether a given sentence contains a cause and effect related by some causal keyword.

Sentences were extracted from the SOU corpus based on a set of cause and effect terms, suggested by two political scientists, shown in Table 3. We extracted sentences containing both terms of a potential cause-effect pair, such as *cancer* and *radon*. The matching was done with stemmed versions of the terms. The motivation for this extraction method was that we wanted to allow other means of expressing causality than the limited set of causality keywords in Table 1. In the final annotation we did not require the sentences to express a causal relation with respect to the term pair used for extraction (which was not shown to annotators). A causal relation between any concepts was allowed.

The annotation was performed in three phases. In a first round, three annotators performed an annotation of 30 sentences without any guidelines. Based on this experience, initial guidelines were drawn up, which were used in a second phase. The guidelines were largely based on those for German by Dunietz et al. (2015), with the exception that we did not divide causality into different subtypes. This procedure increased the inter-annotator agreement from a Fleiss' kappa of 0.38 to 0.56. After this phase the guidelines were modified into the final version in Figure 1. In the final annotation phase, there were two annotators per sample, and a kappa score of 0.5. After the annotation, all examples from phase two and all disagreements from the final phase were consolidated by at least two annotators, to increase agreement. While unsure annotations were allowed, there were very few such annotations used, and they were all resolved to either positive or negative labels in the consolidation phase. In phase two, we used 10 sentences from 3 term pairs (the three bottom term pairs in Table 3), and in the final phase, we sampled 300 sentences equally from the remaining term pairs, filtering out duplicate and near duplicate sentences. The final data set contains 330 sentences, of which 48.5% are causal.

4.2. Ranking Data Set

We define the second task as ranking two sentences by their relevance to a causal query, where a query consists of either a single term specifying a cause or an effect, or a cause–effect term pair. Figure 2 gives an example of a ranking pair extracted for the prompt '[MASK] causes greenhouse effect'. In this example both of the sentences are relevant, but the second sentence is considered more relevant since it explicitly mentions *greenhouse effect* from the prompt. The motivation behind ranking pairs of sentences rather than ranking a longer list was that it is easier to define and create general guidelines for such a task.

A sentence S is said to contain a causal relation CR, if and only if:

- S contains a unit at word level, or above, a connective, which explicitly states a CR.
- This connective does not have any meaning other than causality (in S).
- S contains references to at least two entities for which the stated CR holds; a cause and an effect.
- Causes and effects are normally events or states of affairs, even though also an actor of a certain action can be metonymically considered to be a cause as well.

In addition:

- Modal causal sentences should be annotated (e.g. "X maybe causes Y").
- Negative causal statements should be annotated (e.g. "X does not cause Y").
- Causes and effects do not have to be explicit in S, they could instead be explicit in the context sentences (e.g. referred to by a pronoun).
- We require explicit causal connectives in the text; lexical causality like "kill" meaning "cause to die" should not be annotated.
- While the sentences are sampled based on "cause-effect word pairs", the annotation is not limited to causality with respect to this word pair, but a CR with respect to any two entities should be annotated as positive.

Annotation scheme

- **y**: yes, the sentence contains a CR
- **n**: no, the sentence does not contain a CR
- **?**: unsure/borderline case (avoid overusing)

Figure 1: Guidelines for the curated binary annotation.

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 2: Example of a ranking sentence pair for the prompt *[MASK] medför växthuseffekt* '[MASK] causes greenhouse effect'.

We extracted the ranking data set using the set of cause and effect pairs listed in Table 3. To find relevant text passages, we applied a semantic textual similarity model for Swedish, contrastive tension (Carlsson et al., 2021), based on the KB-BERT model for Swedish (Malmsten et al., 2020). The model was used to embed the subset of all sentences in the SOU corpus matching at least one of the 13 causal keywords in order to avoid matching too many sentences related in theme, but without explicit causal statements. For each query we also embedded a constructed prompt of the cause and/or effect using one of the causality keywords[8]. This prompt consisted of the causality keyword and either both cause and effect at the respective position around the keyword, or each of the two terms individually, with the other replaced by a MASK token. Among the embedded SOU sentences, we then selected the 500 nearest neighbours to the prompt embedding. To obtain pairs of sentences we randomly sampled the neighbours with replacement.

We conducted a first exploratory pilot annotation round on 9 prompts with 10 ranking examples each. Two to three annotators were tasked with determining out of a pair of two sentences the sentence that is the most relevant to a query consisting of a cause, an effect or both. In the course of this annotation round we observed almost no overlap in sentences between the ranking pairs, essentially providing us with mostly unconnected relevance judgments for each prompt. In order to increase the chances of observing the same sentence in multiple pairs, we randomly sampled a subset of 200 sentences out of the ranked list of neighbours that we then sampled pairs of sentences from. The goal of having the same sentence occur in multiple ranking pairs was to obtain a more connected ranking list in the end. Such a list allows us to verify that sentence annotations are consistent for connected sentences. Based on this pilot annotation, we created a set of guidelines.

Following another small pilot annotation round on a selection of the same 9 prompts with 10 new examples each, we observed that we were losing information in treating pairs where both sentences are relevant, but one more so than the other, the same as cases where only one sentence is relevant. To address this, we derived the final guidelines in Figure 3. According to the guidelines, the example in Figure 2 would be labeled as 5, i.e. both sentences are relevant, but the second sentence is more relevant to the query, since it uses a more specific term.

By following the guidelines on another annotation round with 30 more prompts inter-annotator, agreement improved from a Fleiss' kappa of 0.50 on the pilot data to 0.55 on the new examples. For the

[8]We tried applying each of the 13 keywords and ranking by relative frequency and rank of the match, but found that this did not really produce a better semantic ranking than just combining the term or terms with a single keyword. We picked *medföra* ('entail')

A sentence S is relevant in relation to a query Q with cause term C and/or effect term E if and only if the following two conditions hold:

1. At least one query term T in Q (T = C or T = E) is matched in S by a phrase M(T) that is either synonymous with or has a close semantic relation (hyponymy, hypernymy, meronymy) to T.
2. S can be understood as referring to (but not necessarily asserting) a causal relationship where M(C) is a cause or M(E) is an effect (or both).

When determining whether M(T) matches T (condition 1), the following heuristics may be applied:

1. More specific terms (hyponyms) always match more general terms. For example, "tea" and "herbal tea" both match "beverage". Added specificity may result from lexical hyponymy ("tea" – "beverage"), compounding ("herbal tea" – "tea") or modification ("tea with milk" – "tea").
2. More general terms (hypernyms) match more specific terms only if they are close in a semantic hierarchy. For example, "tea" and "beverage" match "herbal tea", but "liquid" does not. Added generality may result from lexical hypernymy ("beverage" – "tea"), decompounding ("tea" – "herbal tea") or dropped modification ("tea" – "tea with milk").
3. The interpretation of terms should be made in context, which means that contextual information may be used to, for example, resolve anaphoric reference, lexical ambiguity, or implicit modification. For example, a pronoun like "it" matches "beverage" if its antecedent matches "beverage", and "tea" matches "herbal tea" if the contextual information supports an inclusive interpretation but not if it makes clear that only "black tea" is relevant.

When ranking two relevant sentences in relation to a query Q with cause term C and/or effect term E, apply the following rules in order of decreasing priority:

1. Prefer sentences with a greater number of matching terms in the correct causal roles.
2. Prefer sentences with semantically closer matches of the query term(s). Specifically: exact match > synonym > hyponym > hypernym > meronym.
3. Prefer more specific and informative sentences. Specifically:
   (a) Prefer explicit statements of causality over implicit statements.
   (b) Prefer factual statements over modal statements.
   (c) Prefer positive statements over negative statements.
   (d) Prefer clausal statements over nominalizations.

Annotation Scheme:

0. both irrelevant
1. first sentence relevant, second sentence irrelevant
2. second sentence relevant, first sentence irrelevant
3. both sentences equally relevant
4. first sentence most relevant, second sentence also relevant
5. second sentence most relevant, first sentence also relevant

Figure 3: Guidelines for the ranking annotation.

30 prompts we chose to sample 20 sentence pairs per prompt. We found that some of the prompts — such as 'climate change adaptation entails investment needs', 'deforestation causes MASK" and 'thunderstorms cause MASK' — were overly specific and generated very few relevant matches with our extraction method. To account for this, we chose to re-rank them with respect to more thematically fitting prompts to the retrieved sentence pairs and added the climate change/investment and rainy weather terms. We also opted for adding the 90 pilot annotation examples (with the pairs: drugs/abuse, radon/cancer, air pollution/diseases). As these had been annotated with 4 instead of 6 labels, they were relabelled to fit the final annotation scheme. Each annotation with a disagreement was then consolidated by at least two annotators and checked for inconsistencies between overlapping pairs. The result is a set of 800 sentence pairs and their ranked relevance with respect to a specific causal prompt.

4.3. Comparison of Extraction Methods

In order to explore the connection between the two sentence extraction methods, where the ranking set was filtered based on causality keywords and the binary set was extracted based on term pairs, we investigated which causality keywords were used in the binary data set. To that end we automatically matched the causality keywords from Table 1, separating them into two groups, the selected group (top), and the filtered group (bottom). Half of the sentences, 80 sentences, had at least one such match, and in a few cases matched more than one keyword. We went through the remaining 80 sentences manually, marking the causal connective. Table 4 shows an overview of all connectives occurring at least four times. Both the selected and filtered causality keywords had a subset that occurred multiple times. For the remaining connectives, most of them were rare, with 49 connectives only occurring once. They are a mix of verbs, nouns, and different types of multi-word expressions. The most frequent keyword not on our keyword list is the verb *innebära* ('mean/imply'), which we could consider including in our set of causality keywords in future work. When we match our causality keywords towards the negative binary sentences, none of the selected 13 keywords match, which is a further validation that they can ex-

Type	Keyword	Translation	Frequency
Selected	påverka	to affect	18
	orsaka	to cause	9
	till följd av	due to	9
	leda till	to lead to	6
	bidra till	to contribute to	5
	på grund av	because of	5
	medföra	to entail	5
	bero på	to depend on	4
	Total		**64**
Filtered	effekt	effect	15
	följd	consequence	9
	därför	therefore	7
	eftersom	because	6
	orsak	cause	6
	Total		**49**
Other	innebära	to mean/imply	9
	om	if	4
	Total		**76**

Table 4: Causality keyword frequency in the curated binary data set, occurring at least 4 times. Totals also include less frequent keywords. Type refers to whether the causality keyword occurs in the list of causality keywords in Table 1.

tract causality with a high precision. Of the 8 filtered keywords, five of them occur in negative sentences, a total of eight times.

5. Pilot Experiments

In this section we report results on a pilot experiment on binary causality detection. Since we have no high-quality Swedish training data, we apply cross-lingual learning, using data in English and German. For testing we use the binary Swedish test set. We also apply few-shot learning, by adding the Swedish binary trial data set to the training set, showing that we need to address the imbalance of the data set in order for that approach to be useful.[9]

We base the cross-lingual experiments on the transformer-based, multilingual model XLM-Roberta (Conneau et al., 2020, XLM-R), using the architecture for sequence classification from the Transformer library (Wolf et al., 2020), with dropout and a linear layer for classification on top of it. We run our system for two epochs, with a learning rate of 2e-5, batch size 32, and maximum sequence length 256. The hyper-parameters were tuned by training on English data and testing on German development data (375 sentences provided by Rehbein and Ruppenhofer (2020)), which we believe is preferable to monolingual tuning.

We used both English and German source language training data. The English data is from the FinCausal data set by Mariko et al. (2020) and the SemEval-2010 data by Hendrickx et al. (2010), where the original multi-way annotations were transformed into binary annotations by considering all cause-effect relations to be positive examples and all other relations to be negative examples. For the German data, the annotations of Rehbein and Ruppenhofer (2020) were turned into binary annotations by considering all instances with both a cause and an effect to be causal. We noted in our experiments that the stricter guidelines of the FinCausal data, requiring quantifiable facts as effects, were problematic to us. Thus, we also opted to only use the positive examples from FinCausal, which we call FinCausal+. The size of these data sets are summarized in Table 5.

Data set	Train	%causal
SemEval	7,200	12.1
FinCausal	13,478	7.5
FinCausal+	1,010	100.0
German	3,104	50.5

Table 5: Size and proportion of causal examples for the English and German training data sets.

Data set	F1-macro	P	R
FinCausal	35.91	60.91	2.12
SemEval	63.11	69.01	50.38
SemEval +FinCausal	48.92	87.82	16.25
SemEval +FinCausal+	62.03	67.17	60.25
German	76.93	75,78	79.37
German +SemEval +FinCausal+	71.56	70.53	75,13

Table 6: F1-macro, and precision and recall for the causal class with different source language training data sets.

5.1. Zero-Shot Experiments

Table 6 presents the zero-shot results.[10] The results are averages over five runs with different random seeds. We show the F1-macro score, as well as precision and recall for the causal class. Here, the performance across different training data choices varies substantially. When looking at the F1-macro and the recall for the causal class of the models where a concatenation of the English source data or only the FinCausal data was used, we can see that the models failed to recognize many examples that actually expressed causality. This may be due to the strict annotation for the FinCausal data, since in the two experiments without the negative FinCausal examples the recall for the causal class and

[9] A subset of these experiments, and additional experiments with other embeddings, results for German, and additional analysis, are presented in Reimann (2021) and Reimann and Stymne (2022)

[10] As a further point of comparison, the monolingual performance of these systems when tested on a matching test set varies between 82.3 for German and 95.7 for FinCausal

		F1-macro	P	R
Consolidation by numerical scores	EN (SE + FC-c)	36.76	49.22	98.75
	DE	67.09	61.63	94.37
	EN + DE	53.90	54.70	98.12
Consolidation by majority vote	EN (SE + FC-c)	60.20	57.36	92.50
	DE	71.23	64.91	72.50
	EN + DE	71.86	66.83	84.83
Balanced Training Set	EN (SE + FC-c)	68.46	65.56	73.75
	DE	78.24	82.96	70.00
	EN + DE	77.46	79.45	72.50

Table 7: F1-macro and precision and recall for the causal class for the few-shot experiments.

the F1-macro were much higher.

Table 6 also demonstrates that finetuning on the German data clearly led to better results than doing so on the English data, even though the German training data contains substantially fewer examples. Combining German and English led to slightly worse results. A possible hypothesis for the superior performance with German may be that the underlying annotation guidelines for both the German data and the Swedish test data were relatively similar, which resulted in a notably better performance. Also, the German data, like the Swedish test data is balanced between positive and negative examples, unlike the English data. The findings of Turc et al. (2021), however, also hint that German in many cases may be generally more beneficial than English as a source language for cross-lingual NLP tasks. Both German and English are relatively closely related to Swedish, which might be a factor contributing to the reasonably good results, but further experiments would be required to investigate the effect of language relatedness.

5.2. Few-Shot Experiments

The evaluation of the causality keywords, described in section 3.3 led to the creation of the Swedish binary trial data set. We wanted to see if using this small and quickly annotated data set could improve results for Swedish. Since the trial data set contains separate annotations by three annotators, we needed to define ways of consolidating the three annotations. We applied three variants of consolidation. In the numerical scoring scheme, a positive annotation received a score of 0.2, an unclear annotation a score of 0.1 and a non-causal annotation a score of zero. We considered the causal label for examples that reach a score of 0.3 or more. The motivation for this scheme was that such sentences have at least some causal signal. This scheme led to 81% causal examples. A stricter alternative is a simple majority vote, where all sentences are considered causal if two of the three annotators agreed on that. However, even for consolidation through majority vote the distribution still is skewed towards the causal class, with 68% causal examples. Thus, we created a third, balanced, variation including all the negative examples as defined by majority vote plus a sample of positive examples from the training data, which has the same size. While this balanced the data set, it reduced the number of examples from 210 to 134.

Table 7 shows the results. For the target language data set, where annotations were calculated through the numerical scheme, the performance was surprisingly low. A clear overuse of the causal class can be observed. Interestingly, this problem seems to become less obvious when using the Swedish training data where the annotations were consolidated by majority vote, with fewer instances of the causal class. Note that neither of these two schemes led to any improvements over zero-shot learning. When we balance the Swedish training data, precision improved further, at some cost to recall, and we see the overall best scores. In all cases, the F1-macro scores are better than the corresponding zero-shot experiments. Again, there are clear differences between the transfer language choice, with German giving the best results in this setting as well, but with the gap to English somewhat reduced.

6. Conclusion

In this paper we present two curated data sets for Swedish causality detection. One data set is focused on binary identification of sentences containing causal expressions, whereas the second data set is focused on ranking of causal sentences with respect to a target cause and/or effect. These resources are mainly considered as test sets for Swedish causality detection. As such they enable the exploration and evaluation of causality detection and causality–theme ranking in Swedish. In addition we release a quickly annotated binary trial data set.

In a set of pilot experiments we explore cross-lingual causality detection, using training data from German and English and one of our new data sets for evaluation. We show that performance varies between three different training data sets in English and German. While we can get some improvements by adding our Swedish trial data set to the training data, this requires balancing the data in the trial set.

This work is a first step towards enabling impact assessment of Swedish governmental reports. The presented data sets will enable further work on both binary causality classification and ranking of causal sentences with respect to a theme, which could then feed into more advanced systems for impact assessment.

Acknowledgements

This work was funded by Vinnova in the project 2019-02252: Datalab for results in the public sector. We would like to thank Sven-Olof Junker and Martin Sparr at The Swedish National Financial Management Authority for valuable discussions and for providing the list of cause and effect pairs in Table 3. The computations were enabled by resources in project UPPMAX 2020/2-2 at the Uppsala Multidisciplinary Center for Advanced Computational Science.

7. Bibliographical References

Carlsson, F., Gyllensten, A. C., Gogoulou, E., Hellqvist, E. Y., and Sahlgren, M. (2021). Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Domeij, R., Knutsson, O., Carlberger, J., and Kann, V. (2000). Granska–an efficient hybrid system for Swedish grammar checking. In *Proceedings of the 12th Nordic Conference of Computational Linguistics (NODALIDA 1999)*, pages 49–56, Trondheim, Norway, December. Department of Linguistics, Norwegian University of Science and Technology, Norway.

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

Garcia, D. (1997). Coatis, an nlp system to locate expressions of actions connected by causality links. In Enric Plaza et al., editors, *Knowledge Acquisition, Modeling and Management*, pages 347–352, Berlin, Heidelberg. Springer Berlin Heidelberg.

Girju, R. (2003a). Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*, pages 76–83, Sapporo, Japan, July. Association for Computational Linguistics.

Girju, R. (2003b). Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*, pages 76–83, Sapporo, Japan, July. Association for Computational Linguistics.

Gordeev, D., Davletov, A., Rey, A., and Arefiev, N. (2020). LIORI at the FinCausal 2020 shared task. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 45–49, Barcelona, Spain (Online), December. COLING.

Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., and Szpakowicz, S. (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, July. Association for Computational Linguistics.

Honnibal, M., Montani, I., Honnibal, M., Peters, H., Landeghem, S. V., Samsonov, M., Geovedi, J., Regan, J., Orosz, G., Kristiansen, S. L., McCann, P. O., Altinok, D., Roman, Howard, G., Bozek, S., Bot, E., Amery, M., Phatthiyaphaibun, W., Vogelsang, L. U., Böing, B., Tippa, P. K., jeannefukumaru, GregDubbin, Mazaev, V., Balakrishnan, R., Møllerhøj, J. D., wbwseeker, Burton, M., thomasO, and Patel, A. (2019). explosion/spaCy: v2.1.7: Improved evaluation, better language factories and bug fixes, August.

Lauscher, A., Ravishankar, V., Vulić, I., and Glavaš, G. (2020). From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online, November. Association for Computational Linguistics.

Malmsten, M., Börjeson, L., and Haffenden, C. (2020). Playing with words at the national library of sweden - making a swedish BERT. *CoRR*, abs/2007.01658.

Mariko, D., Abi-Akl, H., Labidurie, E., Durfort, S., De Mazancourt, H., and El-Haj, M. (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), December. COLING.

Mihăilă, C., Ohta, T., Pyysalo, S., and Ananiadou, S. (2016). BioCause: Annotating and analysing causality in the biomedical domain. *BMC Bioinformatics*, 14(2).

Mirza, P., Sprugnoli, R., Tonelli, S., and Speranza, M. (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, April. Association for Computational Linguistics.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Rehbein, I. and Ruppenhofer, J. (2017). Catching the common cause: Extraction and annotation of

causal relations and their participants. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 105–114, Valencia, Spain, April. Association for Computational Linguistics.

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

Reimann, S. and Stymne, S. (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.

Reimann, S. M. (2021). Multilingual zero-shot and few-shot causality detection. Master's thesis, Uppsala University, Sweden.

Turc, I., Lee, K., Eisenstein, J., Chang, M.-W., and Toutanova, K. (2021). Revisiting the primacy of english in zero-shot cross-lingual transfer.

Versley, Y. (2010). Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In *Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora AEPC 2010.*, pages 83–92, Tartu, Estonia.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Wu, S. and Dredze, M. (2019). Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China, November. Association for Computational Linguistics.