Exploring Cross-Lingual Transfer to Counteract Data Scarcity for Causality Detection

Sebastian Reimann
Uppsala University
Uppsala, Sweden
Ruhr-Universität Bochum
Bochum, Germany
Sebastian.Reimann@ruhr-uni-bochum.de

Sara Stymne
Uppsala University
Uppsala, Sweden
Sara.Stymne@lingfil.uu.se

ABSTRACT

Finding causal relations in text is an important task for many types of textual analysis. It is a challenging task, especially for the many languages with no or only little annotated training data available. To overcome this issue, we explore cross-lingual methods. Our main focus is on Swedish, for which we have a limited amount of data, and where we explore transfer from English and German. We also present additional results for German with English as a source language. We explore both a zero-shot setting without any target training data, and a few-shot setting with a small amount of target data. An additional challenge is the fact that the annotation schemes for the different data sets differ, and we discuss how we can address this issue. Moreover, we explore the impact of different types of sentence representations. We find that we have the best results for Swedish with German as a source language, for which we have a rather small but compatible data set. We are able to take advantage of a limited amount of noisy Swedish training data, but only if we balance its classes. In addition we find that the newer transformer-based representations can make better use of target language data, but that a representation based on recurrent neural networks is surprisingly competitive in the zero-shot setting.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**; *Information extraction*; *Neural networks.*

KEYWORDS

Causal Relations, Causality Detection, Cross-Lingual Transfer

ACM Reference Format:
Sebastian Reimann and Sara Stymne. 2022. Exploring Cross-Lingual Transfer to Counteract Data Scarcity for Causality Detection. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3487553.3517136

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9130-6/22/04.
https://doi.org/10.1145/3487553.3517136

1 INTRODUCTION

Event detection is an important step in analyzing large bodies of text. One important relation type is causal relations, expressions of how a cause can lead to an effect. The Swedish example 1 illustrates this as it contains a cause *åsknedslag* ('thunderstorms'), its consequence *bränder* ('fires') and this relation is made explicit through the use of a causal connective *till följd av* ('as a consequence of').

(1) *Efter bränder till följd av åsknedslag och*
After fires as consequence of thunder and
övriga kända brandorsaker fanns soteld
other known fire-causes existed chimney-fire
och övriga eldstadsrelaterade bränder som nästa
and other fireplace-related fires as next
kategori.
category.
'After fires as a consequence of thunderstorms and other known causes of fire, there were chimney fires and other fires related to fireplaces as a next category.'

Identifying such sentences can, for instance, be an important step in impact assessment of governmental reports, which is our end target. Identifying causal relations in a large, ever-growing body of textual data might allow governmental agencies to track down, analyze, and predict developments within the public sector and, consequently, within society. However, properly annotated data for training supervised neural models on this task still represents a challenge as data is often scarce or not available at all.

This is for example the case for Swedish, which theoretically may be considered a medium-resourced language, with a large Wikipedia giving a fairly good representation in resources like mBERT [5]. Unfortunately, it lacks any annotated data for causality detection. However, there are available data sets, albeit not perfectly matching our target, for the related languages English and German, allowing us to explore cross-lingual learning.

Transfer learning according to [23] aims at improving the performance on a task in a target domain by making use of knowledge obtained from a source domain. Cross-lingual transfer represents a case of such transfer learning where the source domain is represented through a source language (sometimes also called a transfer language) and the target domain through a specific target language and which may help here to address the lack of training data. In this paper we explore two settings, zero-shot learning where we only use training data from the source language and few-shot learning, where we also add a limited amount of training data from the target language. Multilingual representations such as mBERT [5] have

been shown to yield good results across a range of tasks [24, 32] in zero-shot settings where no training data is available for the target language and task. However using few-shot learning, even with only a small amount of in-language annotated data for the target task may lead to competitive results [21].

In this paper our main focus is on sentence-level causality detection for Swedish. We investigate zero-shot transfer with training and testing on different languages and investigate potential gains from adding a small amount of target language data. Previous work [11, 21] has suggested that just a few hours of annotation work may greatly improve results. Thus, we also investigate how a quickly annotated small Swedish data set can be used for training. In order to evaluate our methods, we annotate a test set for Swedish sentence-level causality detection, where the task is to decide whether a sentence contains a causal relation or not.

The disparity of sources presents an additional challenge. There are considerable differences between both the two English data sets, and the German data set we use for training our cross-lingual models, as well as differences between our Swedish data sets. In previous work on cross-lingual argument mining for example [9], domain mismatches between data sets from different sources led to problems. In the data sets used in our study, the guidelines for when to consider a sentence causal differ, as well as the level of annotation, and the balance between causal and non-causal examples. In all cases it was possible to convert the data sets to sentence-level causality detection, so that we could use them for our goals. To shed further light on cross-lingual causality detection, we also perform exploratory experiments with German as the target language.

Our main goals are to:

- Investigate which type of multilingual pretrained embedding may be the most beneficial, comparing the transformer-based mBERT [5] and XLM-RoBERTa [3] with LASER [1] embeddings based on recurrent neural networks
- Explore the impact of annotation schemes from different annotation projects, since data is still relatively scarce for the present task
- Investigate to what extent cross-lingual transfer for causality detection may benefit from additional target language data in training

In summary, we find that LASER for zero-shot transfer and causality detection is often able to at least compete with the newer Transformer-based models. It, however, is not able to make much use of additional training data, which on the other hand has a great impact on the performance of mBERT and XLM-R, although not always in a positive way. Our results also further stress the importance of consistent annotation guidelines.

2 PREVIOUS WORK

2.1 Causality Detection

Early attempts to automatically detect causal relations were rule-based and focused on lexical patterns as well as morphological and morphosyntactic clues [10, 17]. Later work started using machine learning, such as decision tree classifiers taking different lexical and syntactical features as input to learn rules by which causality could be detected [12]. Besides decision trees, also support vector machines (SVM) have been proven useful. Most of the contributions to the SemEval-2010 Task 8 on the classification of semantic relations, including causal or cause-effect relations, made use of SVMs [15], including the best performing contribution [27].

More recently, neural networks have been used to detect causal relations, for instance [4], a BiLSTM that takes GloVe word vectors plus linguistic features based on POS-tags, dependency relations and WordNet noun hierarchies as input for finding causal sentences and extracting the respective cause, effect and causal connective. In the first subtask of the FinCausal 2020 shared task [22], focusing on binary classification of causal relations, pretrained language models based on the Transformer architecture [30], and BERT [5] in particular, were dominating, with ensemble architectures [13, 28] being among the best performing systems. However, also the provided baseline [22] using only the English BERT-based model with a dropout layer and a linear regression layer on top of it resulted in a strong performance.

2.2 Multilingual Pretrained Embeddings and Cross-Lingual Transfer

LASER (Language Agnostic Sentence Representations) [1] represents one attempt to provide sentence representations that can be used in NLP tasks for a wide range of languages. These sentence representations are based on a recurrent neural network, in this case a BiLSTM [14, 16]. Sentence representation are obtained through a BiLSTM-encoder, pretrained on a machine translation task by translating sentences from 93 different languages into English and Spanish. The evaluation on multilingual natural language inference (XLNI) and document classification with English as a source language in all experiments resulted in accuracies between 62% and 72% for the former and 60% and 85% for the latter.

Devlin et al. [5] provided a multilingual version of the transformer-based BERT (mBERT) besides the monolingual version that was for example also successful for causality detection [22]. It was pretrained on a masked language modeling and a next sentence prediction task involving the entire Wikipedia dumps of 104 languages. In zero-shot experiments on NER and the languages English, German, Dutch and Spanish, mBERT achieved accuracies between 65% and 75%. For zero-shot POS-tagging with mBERT, the accuracies were even ranging from 80% to 90%. These experiments were later extended, covering 38 languages and zero-shot document classification and XNLI [32]. It was also found that the layers of mBERT were able to recognize the respective language with an accuracy of 96% on average when tested on sentences of 99 different languages, which suggests that each layer at least to some extent contains language-specific information [32].

Another pretrained transformer-based model is XLM (X-Lingual Model), pretrained on an additional machine translation task besides masked language modelling [19]. XLM eventually outperformed mBERT on zero-shot XNLI. Conneau et al. introduced a further improved version of XLM, XLM-RoBERTa (XLM-R), which was pretrained using only the masked language modelling task, on a larger corpus, a larger vocabulary and a higher number of hidden states [3]. It outperformed both mBERT and XLM on zero-shot NER, XNLI and question answering.

It has further been shown that finetuning mBERT and XLM-R on various sizes of target language data, ranging from 10 to 1000 additional sentences can lead to improvements [21]. For most tasks, improvements were seen even with as little as ten sentences. This improvement was more pronounced for languages that were not related to English. Most work on zero-shot transfer has used English as a source language. However, it has been shown that it is often beneficial to use other languages, even when the other language is machine translated from English [29]. Here, in particular, German and Russian often worked well as source languages across tasks and target languages.

3 DATA

In this section we describe the newly annotated Swedish data, as well as the existing data sets for English and German, including the modifications we had to make to them to fit the Swedish target. Tables 1 and 2 show an overview of the data.

3.1 English

The English data is composed of the data sets of two previous shared tasks: SemEval-2010 Task 8 [15] and the FinCausal 2020 shared task [22]. The latter already had a binary labelling according to whether a sentence expresses causality or not. The original annotations of the SemEval-2010 data set contain nine different semantic relations, which were replaced with a binary labelling, where all sentences expressing a cause-effect relation were considered as causal and all other sentences received the label non-causal.

The two English data sets differ with respect to their annotation schemes. For FinCausal [22] a modification to the effect role is proposed so that only a quantified fact, a fact that is directly connected to a measure and that expresses a number or quantity, can be an effect. Neither SemEval, nor the German and Swedish data sets include this constraint in their annotation guidelines. To see the effects of this difference, we used only the FinCausal or SemEval data respectively, a concatenation of both (SE+FC in Table 1), and, since the FinCausal annotations are stricter and it thus can be assumed that a causal example according to the FinCausal annotation would also be causal according to the SemEval annotation scheme, a combination of the FinCausal positive data and all SemEval data (SE+FC-c in Table 1). Since no development set was provided for SemEval, we used 10% of the provided SemEval training data as a development set in experiments where only the SemEval data was involved. In the setting with the SemEval data plus the positive FinCausal examples, we simply add the positive examples from the FinCausal development set to these 800 sentences.

3.2 German

All German data used in the present study is taken from a data set of causal language [25], who base their annotation scheme on [8]. It contains sentences from the TIGER corpus [6] of German newspaper texts and the Europarl corpus [18] that consists of proceedings from the European parliament. Contrary to the English data sets, they annotated the data on the token level with two different types of tags, indicating the role of a participant in a causal relation (cause, effect, affected and actor) and the specific type of causation (consequence, motivation, purpose). We transformed their annotations to binary sentence labels in order to bring them into accordance with the other data sets by considering all examples that contain both a cause and an effect as causal relations and consequently, sentences in which one of these tags or both were missing, were considered to be non-causal. Additionally, we considered all the different types of causation here as a single causal class.

3.3 Swedish

All Swedish data was extracted from a corpus of governmental reports, written between 1994–2020.[1] This data was extracted from HTML-format, and pre-processed to retrieve headers and running text, removing non-text elements such as tables, headers, and footers, and merging sentences and words across line and page breaks.[2]

From the Swedish corpus, we sample sentences by searching for term pairs that potentially express a causal relation, such as *radon → cancer* and *car traffic → pollution*. The sentences were annotated by three annotators with a background in computational linguistics, two native speakers and one native German speaker with a high level of Swedish, according to whether they express causality. In an initial pilot annotation round, the inter-annotator agreement was relatively low, with a Fleiss' Kappa κ of 0.38, which expresses only fair agreement [20]. Thus, specific guidelines inspired by Dunietz et al. [8], similar to the guidelines of the German data presented in Section 3.2, were drawn. These guidelines focused on explicit causality, which requires, besides a cause and an effect, the presence of a causal connective that exclusively expresses causality. Thus, temporal relations that implicitly express causality are for example ruled out, but modal and negative causality was annotated. Unlike [7] we did not consider different types of causality, but grouped all their types into a single causal class. The guidelines led to an improved inter-annotator agreement of 0.58 in a second pilot round on 30 sentences. The final annotation round covered 300 sentences, which were combined with the 30 sentences from the final pilot round, giving 330 sentences in the final data set. Each sentence was annotated by at least two annotators. The inter-annotator-agreement after the annotation phase, was at 0.5, thus similar to the second pilot. To further improve annotation quality, all disagreements were consolidated by discussion among the annotators.

The small Swedish training set was collected by a quick annotation phase with three annotators. Ten sentences were sampled from the governmental report corpus for each of 21 terms which could potentially express causality such as *orsaka* ('cause'), *på grund av* ('because of') and *resultat* ('result').[3] This extraction procedure led to a large proportion of positive examples. Each annotator decided for each sentence if it expresses causality, if it is an unclear case or if it does not express causality. Similarly to the first test set pilot without guidelines, only fair inter-annotator-agreement [20] of 0.45 was reached. Additionally, no consolidation was performed on this data.

We attempted several methods to transform the annotation of the three annotators into one single annotation for each sentence. Here, it can be argued that if for example one annotator decides on

[1] Statens offentliga utredningar, available from http://data.riksdagen.se/data/dokument/.
[2] Available from https://github.com/UppsalaNLP/SOU-corpus
[3] The full set of causality terms and causal term pairs can be found in [26].

Table 1: Distribution of the English training and development data including the percentage of causal examples

	SE+FC	%	FinCausal	%	SemEval	%	SE+FC-c	%
Train	21,478	9.37	13,478	7.49	7,200	12.06	8,210	22.87
Dev	8,629	6.62	8,629	6.62	800	16.88	1,369	51.42

Table 2: Distribution of the German and Swedish data, including the percentage of causal examples

	German	%	**Swedish**	%
Train	3,104	50.48	210	see Tab. 3
Dev	375	51.73	–	–
Test	905	50.60	330	48.49

Table 3: Label distribution of the three different variants of the Swedish training data

	Pos.	Neg.	% Pos.
Numerical Scores	170	40	80.95
Majority Vote	143	67	68.10
Balancing	67	67	50.00

a sentence to be causal and another one decides that this sentence represents an unclear case, then the sentence conveys at least some notion of causality that would make a negative annotation too strict. A method that takes this into account would be the assignment of numerical scores where 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. This presents a relatively low threshold since only one positive and one unclear annotation would be needed for the causal label. A stricter alternative is a simple majority vote, where all sentences are considered causal if two of the three annotators agreed on that. For both the numerical weighting and the majority vote, the distribution is strongly skewed towards the causal class, as Table 3 shows. Thus, we create a third variation including the negative examples from the Swedish training data plus a sample of positive examples from the training data, which has the same size.

Both Swedish causality data sets are publicly available under the CC-BY license.[4]

4 EXPERIMENTAL SETUP

4.1 Model Architectures and Embeddings

One goal of this paper is to compare different pretrained multilingual embeddings on their performance on causality detection, especially by comparing the BiLSTM-based LASER with the transformer-based mBERT and XLM-R. The embedding types and the classifier architectures that make use of them will be introduced in the following. Tuning hyperparameters for all models, if not stated otherwise, was conducted through training or finetuning the respective models on English and testing them on the German development set since it seems to be intuitive to choose hyperparameters by evaluating them on a setting where some cross-lingual transfer is involved, rather than in a monolingual setting.

LASER. The LASER sentence embeddings [1] are used in order to include a non-finetuning based approach. As recommended in [1], we first tokenize our sentences, and then transform them to BPE subword units. We finally feed them to the encoder to obtain the respective embeddings. For classification, we use the provided multi-layer perceptron (MLP) classifier, with two hidden layers. We train the MLP classifier for 100 epochs with a learning rate of 0.001, 10 nodes in the first and 8 nodes in the second hidden layer, a batch size of 12 and dropout of 0.1.

Multilingual BERT. In our experiments involving mBERT, we use the BertForSequenceClassification architecture provided by HuggingFace [31], in which dropout and an additional linear layer for classification are added on top of the regular BERT architecture. No preprocessing is carried out before encoding the text through the BERT tokenizer. The mBERT-based classifier is finetuned for 3 epochs with a learning rate of $2e-5$ and batch size of 32. The maximum length of the inputs is 256.

XLM-R. Here again, the implementation from the HuggingFace Transformer library for sequence classification with an addtional linear layer was used. We finetune XLM-R for 2 epochs with a learning rate of $2e-5$, a batch size of 32 and a maximum length of 256.

4.2 Experiments

Our goal is it to compare different embedding types as well as to measure the impact of additional target language data.[5] Consequently, we perform both zero-shot experiments and few-shot experiments, with all three multilingual representations involved. The source language for German as a target language will be English in all experiments. In experiments with Swedish as a target language, we additionally use German as a source language, also in combination with English. We also investigate the influence of the different annotation schemes, especially with respect to the stricter annotation for the FinCausal data. Thus, for experiments with English as a source language, we try out all the different combinations of the English data sets.

For measuring the effect of target language data in training, we follow an approach similar to [21] by using varying amounts of training examples. The Swedish data is scarce, which limits our possibilities to experiment with different sizes. To measure the

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

[5] Additional experiments as well as a more in-depth analysis can be found in [26]

effect of larger portions of target language data, we thus use the German data set, which is notably larger in size compared to the Swedish data, but still considerably smaller than all English data sets and use 210 sentences (6.8%,; the size of the Swedish training data), as well as 12.5%, 25%, 50%, 75% of it, as well as the full German training data. For Swedish we also experiment with different ways of combining the training data from the three annotators.

We consider the variation when running systems, and repeat all zero-shot experiments five times. For evaluation we present the F1-macro score, and in addition the precision and recall for the causal class.

5 RESULTS AND DISCUSSION

5.1 Zero-Shot Experiments

Table 4 presents the results of the zero-shot experiments with English as a source language. For all combinations of target and source languages, the differences across embedding types were relatively small with LASER often matching or even outperforming mBERT and XLM-R. However, differences can be observed with respect to the training data choices. When looking at the F1-macro and the recall for the causal class of the models where the joint English data set or only the FinCausal data were used, we can see that the models in the end failed to recognize a wide range of examples in the test data that actually expressed causality. Here, it indeed seems that the aforementioned problem of a stricter annotation for the FinCausal data, compared to the other data sets, caused problems, especially when considering that in the two experiments without the negative FinCausal examples the recall for the causal class and the F1-macro were much higher.

In both the English–German and the English–Swedish zero-shot scenario, adding the positive FinCausal examples to the SemEval data improved recall substantially as well as the performance overall. It, however, did so at the price of precision. The models then were exposed to a broad range of additional positive examples with vocabulary from the financial domain and expressions of quantity. This consequently led to difficulties in generalization, which is expressed through a decrease in precision when comparing the results here to the results when only the SemEval data is involved.

Additionally, there were several issues concerning the experiments where we used English as a source language and German as a target language which can be linked to differences between the two languages. German differs with respect to adpositions since, besides prepositions, it also offers a range of causal postpositions that follow the verb, instead of preceding it, such as *halber* ("for the sake of"). Moreover, several German prepositions that express causality like *mangels* ("out of a lack of") require more complex constructions when being translated to English. Problems with these two phenomena occurred with all different English data sets in the zero-shot scenario when being tested on German.

Table 5 presents results for Swedish, where we use German as a source language, either on its own, or in combination with English. When comparing the results of Table 4 and Table 5 it becomes visible that for mBERT and XLM-R, 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. Possible hypotheses for this 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, and that the German training data, unlike the English, is balanced. The findings of [29], however, also hint that German in many cases may be generally more beneficial than English as a source language for cross-lingual NLP tasks, which may be the case here as well. With LASER, though, English performs slightly better than German as source language. Combining German and English led to slightly worse results than for only German with mBERT and XLM-R, and to higher recall but lower precision compared to a single language for LASER. However, no clear patterns of problems with linguistic phenomena were observed when applying zero-shot cross-lingual transfer from English to Swedish.

5.2 Few-Shot Experiments

Table 6 gives an overview of the results in the few-shot experiments involving additional Swedish target language data. For the Swedish data set where the final annotations were calculated through the numerical scheme, the performance was surprisingly low, especially for the two transformer-based models. For both mBERT and XLM-R and for all source languages, 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. Here only mBERT in combination with the English data set was performing poor.

When we balance the Swedish training data in the previously described manner, however, we were even able to see improvements over the average of the zero-shot experiments, similar to previous findings [21, 33]. This is interesting since in our case, we use only a little bit more than half of the available Swedish training data. Previous work, e.g. [2] has found that mBERT and XLM-R capture more language-specific information compared to models pretrained on a machine translation task such as the BiLSTM encoder for LASER. A possible hypothesis may thus be that, since a vast majority of the Swedish examples is causal, the models appear to wrongly consider some general aspects of the Swedish language as decisive for the decision on causality and consequently overgeneralize and classify a vast amount of Swedish examples as causal. This is also supported by the fact that the problem is notably more pronounced for English as a source language, where, on the other hand, positive causal examples in the training data are scarce and that this issue of overgeneralization does not really affect the LASER embeddings. The Swedish few-shot results have demonstrated that an improvement similar to the findings in [21] can be achieved for multilingual causality detection, at least when efforts are taken to counteract data imbalance.

Figure 1 shows results for few-shot learning with German as a target, and a varying amount of German training data. For mBERT as well as XLM-R, few-shot transfer led to a similar improvement over the zero-shot setting. Moreover, we can see that larger additional target language data sets may even lead to a greater improvement for both transformer-based models, even though, when using more than half of the data set, the differences become smaller again. For LASER on the other hand more target-language data only leads to minor improvements and using the full data set leads

Table 4: F1-macro scores and precision / recall values for the causal class on average in zero-shot experiments when the respective model is trained or finetuned on English data

	Model	German			Swedish		
		F1	P	R	F1	P	R
SemEval +FinCausal	LASER	41.85	75.03	9.05	52.21	96.35	19.38
	mBERT	42.89	74.92	10.55	46.07	78.32	13.50
	XLM-R	42.63	75.33	10.24	48.92	87.82	16.25
FinCausal	LASER	34.58	49.37	1.37	34.96	58.33	1.00
	mBERT	35.97	41.24	2.71	36.26	61.09	2.50
	XLM-R	35.06	25.10	2.08	35.91	60.91	2.12
SemEval	LASER	53.38	79.70	24.13	65.16	74.55	47.50
	mBERT	49.50	76.51	19.20	62.19	72.23	51.13
	XLM-R	52.64	79.81	23.28	63.11	69.01	50.38
SemEval +FinCausal (causal)	LASER	63.79	65.90	57.25	67.59	63.04	85.62
	mBERT	57.29	62.43	42.84	57.17	69.54	44.38
	XLM-R	56.76	64.33	41.73	62.03	67.17	60.25

Table 5: F1-macro scores and precision / recall values for the causal class on average in zero-shot-experiments involving finetuning on German and testing on Swedish

Data	Model	F1	P	R
German	LASER	66.65	62.40	82.90
	mBERT	72.51	73.09	72.75
	XLM-R	76.93	75,78	79.37
German + SemEval + FinCausal (causal)	LASER	66.59	61.78	87.12
	mBERT	71.44	73.59	65.50
	XLM-R	71.56	70.53	75.13

Table 6: F1-macro scores for few-shot experiments involving Swedish as a target language. The English data is SemEval+FinCausal (causal).

Method	Source	LASER			mBERT			XLM-R		
		F1	P	R	F1	P	R	F1	P	R
None (Zero-Shot)	EN	67.59	63.04	85.62	57.17	69.54	44.38	62.03	67.17	60.25
	DE	66.65	62.40	82.90	72.51	73.09	72.75	76.93	75.78	79.37
	EN+DE	66.59	61.78	87.12	71.44	73.59	65.50	71.56	70.53	75.13
Numerical	EN	67.06	59.41	88.75	32.65	48.48	100	36.76	49.22	98.75
	DE	60.69	57.72	56.76	56.76	56.32	97.50	67.09	61.63	94.37
	EN+DE	65.46	60.67	90.62	55.24	55.16	96,88	53.90	54.70	98.12
Majority	EN	68.25	63.38	84.38	35.22	49.08	100	60.20	57.36	92.50
	DE	64.78	60.44	85.00	67.53	62.13	91.25	71.23	64.91	72.50
	EN+DE	66.65	61.78	86.88	72.43	75.74	64.38	71.86	66.83	84.83
Balanced	EN	70.57	67.40	76.25	59.30	56.68	87.50	68.46	65.56	73.75
	DE	68.58	64.47	79.37	71.34	66.84	81.87	78.24	82.96	70.00
	EN+DE	67.52	62.67	85.00	75.07	80.62	65.00	77.46	79.45	72.50

even to a worse performance compared to when using 75% or more of the available German training data. Given findings of [2], that the LASER embeddings contain less language-specific information compared to the embeddings of mBERT and XLM-R, it could thus

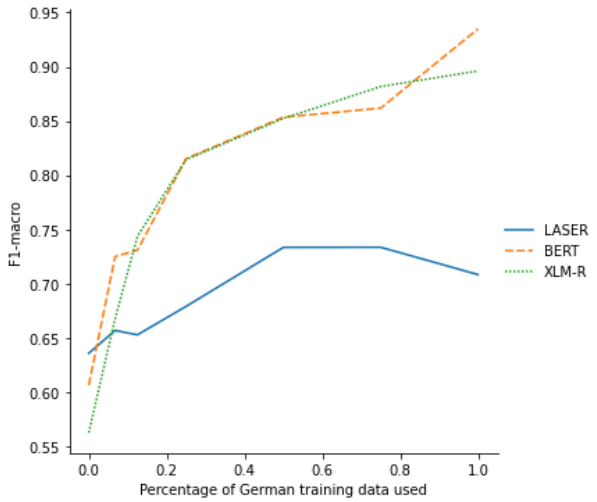

Figure 1: F1-macro scores when including target language data of different sizes through concatenating target and source language training data

be that LASER may be worse at taking advantage from additional German and Swedish data.

6 CONCLUSION

We explored cross-lingual causality detection for German with English as a source language as well as for Swedish with English and German as source languages. A challenge is that there are several differences between the existing data sets in English and German, as well as our new Swedish data sets. In particular, the strict annotation scheme of the FinCausal data [22] led to shortcomings in the zero-shot scenario, which, however, could be smoothed over by removing the negative examples, hinting at a possible confusion related to the different annotation guidelines. We also needed to balance our small Swedish training set, in order to be able to take advantage of it.

We also compared the BiLSTM-based LASER to the transformer-based mBERT and XLM-R. We showed that LASER was surprisingly competitive, especially in the cross-lingual zero-shot setting from English, where it had the best performance. For Swedish, it was preferable to transfer from German with XLM-R and mBERT, which gave the overall best results, whereas the difference was small for LASER. This preference for German could be traced back to, on the one hand, similar annotation guidelines for the respective data sets and a more balanced data set, but potentially also to German being a better source language for cross-lingual transfer [29].

The findings of [21] regarding few-shot transfer also apply here since additional target language data led to notable improvements for mBERT and XLM-R. However, the findings of the Swedish few-shot experiments demonstrate that caution is advised with regards to class imbalance. Even small samples of target language data that are skewed towards one class may lead to overgeneralization, possibly because the model interprets some actually language specific characteristics as relevant for the decision on causality.

So far, we only attempted cross-lingual transfer for causality detection between languages of the same family. A possible line for future research may thus be to explore how well the multilingual representations used in the present study are able to transfer between more distantly related languages. One more possible line of research, not only related to causality detection but generally to multilingual NLP would be to further explore the observed phenomenon of overgeneralization, especially its links to the information learned by mBERT and XLM-R.

ACKNOWLEDGMENTS

We would like to thank Luise Dürlich, Gustav Finnveden, and Joakim Nivre for annotation work and for insightful discussions, and 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. The computations were enabled by resources in project UPPMAX 2020/2-2 at the Uppsala Multidisciplinary Center for Advanced Computational Science. Sara Stymne was funded by Vinnova in the project 2019-02252: Datalab for results in the public sector.

REFERENCES

[1] Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics* 7 (March 2019), 597–610. https://doi.org/10.1162/tacl_a_00288

[2] Rochelle Choenni and Ekaterina Shutova. 2020. What does it mean to be language-agnostic? Probing multilingual sentence encoders for typological properties. arXiv:2009.12862 [cs.CL]

[3] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[4] Tirthankar Dasgupta, Rupsa Saha, Lipika Dey, and Abir Naskar. 2018. Automatic Extraction of Causal Relations from Text using Linguistically Informed Deep Neural Networks. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, Melbourne, Australia, 306–316. https://doi.org/10.18653/v1/W18-5035

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[6] Stefanie Dipper, Thorsten Brants, Wolfgang Lezius, Oliver Plaehn, and George Smith. 2001. The TIGER Treebank. In *The Workshop on Treebanks and Linguistic Theories, TLT'01*. 24–41.

[7] Jesse Dunietz, Greg Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and Dave Ferrucci. 2020. To Test Machine Comprehension, Start by Defining Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7839–7859. https://doi.org/10.18653/v1/2020.acl-main.701

[8] Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2015. Annotating Causal Language Using Corpus Lexicography of Constructions. In *Proceedings of The 9th Linguistic Annotation Workshop*. Association for Computational Linguistics, Denver, Colorado, USA, 188–196. https://doi.org/10.3115/v1/W15-1622

[9] Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2018. Cross-lingual Argumentation Mining: Machine Translation (and a bit of Projection) is All You Need!. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 831–844. https://www.aclweb.org/anthology/C18-1071

[10] Daniela Garcia. 1997. COATIS, an NLP system to locate expressions of actions connected by causality links. In *Knowledge Acquisition, Modeling and Management*, Enric Plaza and Richard Benjamins (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 347–352.

[11] Dan Garrette and Jason Baldridge. 2013. Learning a Part-of-Speech Tagger from Two Hours of Annotation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Atlanta, Georgia, 138–147. https://www.aclweb.org/anthology/N13-1014

[12] Roxana Girju. 2003. Automatic Detection of Causal Relations for Question Answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*. Association for Computational Linguistics, Sapporo, Japan, 76–83. https://doi.org/10.3115/1119312.1119322

[13] Denis Gordeev, Adis Davletov, Alexey Rey, and Nikolay Arefiev. 2020. LIORI at the FinCausal 2020 Shared task. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*. COLING, Barcelona, Spain (Online), 45–49. https://www.aclweb.org/anthology/2020.fnp-1.6

[14] Alex Graves, Abdel-rahman. Mohamed, and Geoffrey Hinton. 2013. Speech Recognition with Deep Recurrent Neural Networks. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. Vancouver, Canada, 6645–6649. https://doi.org/10.1109/ICASSP.2013.6638947

[15] 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*. Association for Computational Linguistics, Uppsala, Sweden, 33–38. https://www.aclweb.org/anthology/S10-1006

[16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (nov 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[17] Christopher S. G. Khoo, Jaklin Kornfilt, Robert N. Oddy, and Sung H. Myaeng. 1998. Automatic Extraction of Cause-Effect Information from Newspaper Text without Knowledge-Based Inferencing. *Literary and linguistic computing* 13, 4 (1998), 177–186.

[18] Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X: Papers*. Phuket, Thailand, 79–86.

[19] Guillaume Lample and Alexis Conneau. 2019. Cross-lingual Language Model Pretraining. arXiv:1901.07291 [cs.CL]

[20] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174. http://www.jstor.org/stable/2529310

[21] Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 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)*. Association for Computational Linguistics, Online, 4483–4499. https://doi.org/10.18653/v1/2020.emnlp-main.363

[22] 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*. COLING, Barcelona, Spain (Online), 23–32. https://www.aclweb.org/anthology/2020.fnp-1.3

[23] Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 10 (2010), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

[24] Matúš Pikuliak, Marián Šimko, and Mária Bieliková. 2021. Cross-lingual learning for text processing: A survey. *Expert Systems with Applications* 165 (2021), 113765. https://doi.org/10.1016/j.eswa.2020.113765

[25] Ines Rehbein and Josef Ruppenhofer. 2020. A New Resource for German Causal Language. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 5968–5977. https://www.aclweb.org/anthology/2020.lrec-1.731

[26] Sebastian Michael Reimann. 2021. *Multilingual Zero-Shot and Few-Shot Causality Detection*. Master's thesis. Uppsala University, Department of Linguistics and Philology.

[27] Bryan Rink and Sanda Harabagiu. 2010. UTD: Classifying Semantic Relations by Combining Lexical and Semantic Resources. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Uppsala, Sweden, 256–259. https://www.aclweb.org/anthology/S10-1057

[28] Zsolt Szántó and Gábor Berend. 2020. ProsperAMnet at FinCausal 2020, Task 1 & 2: Modeling causality in financial texts using multi-headed transformers. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*. COLING, Barcelona, Spain (Online), 80–84. https://www.aclweb.org/anthology/2020.fnp-1.13

[29] Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the Primacy of English in Zero-shot Cross-lingual Transfer. arXiv:2106.16171 [cs.CL]

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. arXiv:1706.03762 [cs.CL]

[31] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

[32] Shijie Wu and Mark Dredze. 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)*. Association for Computational Linguistics, Hong Kong, China, 833–844. https://doi.org/10.18653/v1/D19-1077

[33] Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2020. A Closer Look at Few-Shot Crosslingual Transfer: Variance, Benchmarks and Baselines. arXiv:2012.15682 [cs.CL]