# Neural Disambiguation of Causal Lexical Markers Based on Context

**Eugenio Martínez-Cámara**[†], **Vered Shwartz**[‡], **Iryna Gurevych**[†], **Ido Dagan**[‡]
[†]Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
[‡]Bar-Ilan University, Ramat-Gan, Israel
{camara, gurevych}@ukp.tu-darmstadt.de, vered1986@gmail.com,
dagan@cs.biu.ac.il

**Abstract**

Causation is a psychological tool of humans to understand the world and it is projected in natural language. Causation relates two events, so in order to understand the causal relation of those events and the causal reasoning of humans, the study of causality classification is required. We claim that the use of linguistic features may restrict the representation of causality, and dense vector spaces can provide a better encoding of the causal meaning of an utterance. Herein, we propose a neural network architecture only fed with word embeddings for the task of causality classification. Our results show that our claim holds, and we outperform the state-of-the-art on the AltLex corpus. The source code of our experiments is publicly available.[1]

## 1 Introduction

Causation is a psychological tool of humans to understand the world independently of language, and it is one of the principles involved in the construction of the human mental model of reality (Neeleman and van de Koot, 2012). Following the words of Reinhart (2002), causal relations are imposed by humans on input from the world, and the (computational) linguist's task is to understand what it is about language that enables speakers to use it to describe their causal perceptions.

Due to the importance of modelling causality, in this paper we present a study of the classification of the causal meaning of an utterance. The computational treatment of causality requires a computational definition, which should be grounded in a philosophical theory. There are two broad categories of theories modelling causality: dependency and production theories. The dependency theories define causality as a relation of dependence between two eventualities, and the counterfactual theory of Lewis (1973) is their main representative theory. Differently, production theories define causation as the transmission of forces with the sense that the transmission of the force of the causing event enables the provoked effect. In this case, the force dynamic theory of Talmy (1988) stands out.

The computational treatment of causality can be addressed by the study of the causal relation between a predicate and its arguments (lexical causality) or towards the analysis of the relation between propositions (propositional causality). Copley and Wolff (2014) argue that production theories may fit with lexical causality, because the transmission of a force may be better projected by the relation between a predicate and its arguments. Otherwise, propositional causality is explicitly projected by some lexical markers, such as *because*, that represent the dependency relation between propositions, so dependency theories may suit propositional causality.

We argue that the dependency theories restrict the phenomenon of causality, because they require the satisfaction of some conditions to establish the causal relation, hindering the representation of the human causal reasoning, which is sometimes based on assumptions instead of empirical evidences. In contrast,

[1] https://github.com/UKPLab/iwcs2017_disambiguation_causality_lexical_markers

production theories encompass those causal relations that are not founded on facts, can incorporate the general knowledge of humans into the model and may represent abstract causality. We additionally argue that the transmission of forces can be produced not only between the arguments of a predicate, but between the eventualities expressed in different propositions too. Thus, we follow the production theories, and we extend their definition in order to model propositional causality.

The computational treatment of causality could benefit from the existence of specific linguistic constructions of causality, and the matching of those linguistic constructions with any philosophical theory of causality. However, Neeleman and van de Koot (2012) assert that causation lacks of unarguable syntactic constructions, and Copley and Wolff (2014) conclude that there is not an agreement between the link of linguistic and philosophic theories of causation. Nonetheless, there are some lexical units that project the meaning of causality of an utterance. Those lexical units can be verbs (*cause*), prepositions (*because*), adverbs (*consequently*) or expressions like *as a result of*. However, those lexical markers are ambiguous, which means that they can also be used without a causal meaning. Moreover, the set of lexical markers with a causal meaning is not limited, which increases the need of their disambiguation.

Due to the lack of unambiguous linguistic construction of causality, we claim that the use of linguistic features may restrict the representation of causality meaning. Also, the use of dense vector spaces, roughly speaking word embeddings, can provide a better representation of causality, and can improve the disambiguation of the causal meaning of lexical markers. Hence, we propose a neural network architecture with two inputs as sequences of word embeddings, encoding the left and the right context of the lexical marker. We evaluate our proposal on the AltLex corpus (Hidey and McKeown, 2016), which is a corpus of causal relations signalled by lexical markers with a wider coverage of those kind of expressions than the Penn Discourse TreeBank Corpus (PDTB) (Prasad et al., 2008). Empirical results on the AltLex corpus show that our claim indeed holds and our system outperforms the state-of-the-art on that corpus.

## 2 Related Work

Previous works in the task of causal language classification are mainly focused on lexical or propositional causality, explicit causality and some of them were restricted to a narrow kind of syntactic constructions.

The works of Girju (2003); Riaz and Girju (2013, 2014) were focused on the classification of lexical causality conveyed between verbs and nouns. Recently, Kruengkrai et al. (2017) proposed a system for the classification of propositional causality in Japanese. The system incorporates background knowledge for enhancing the learning process through the use of multi-column convolutional neural networks.

Regarding the classification of explicit causality, Khoo et al. (1998) proposed a rule-based system grounded in regular expressions for the classification of explicit causal relations, whereas Mirza and Tonelli (2016) presented a supervised system based on the use of lexical, syntactic and semantic features from WordNet. The proposal of Bethard and Martin (2008) is similar to that of Mirza and Tonelli (2016), but it was focused only on conjunction constructions, namely conjoined events. The three approaches suffer from the ambiguity of the lexical markers, the limited coverage of the linguistic resources and the constraint to a specific syntactic construction.

In contrast, our proposal tries to cover lexical and propositional causality independently of whether it is explicit or implicit, and we do not restrict the study to a specific syntactic construction.

## 3 Causality classification

The next sections present the task definition (§ 3.1), the corpus (§ 3.2) and the proposed system (§ 3.3).

### 3.1 Definition

The cause dynamics model of Wolff (2007) is the main basis for our definition of causality, because we can adapt it to the causal language that we find on real data. The dynamics model states that causation is

| Sentence | Type |
|---|---|
| (Cathay Pacific delayed both legs of its quadruple daily Hong Kong to London route)$_{e_1}$ [due to]$_l$ (this disruption in air traffic services.)$_{e_2}$ | Explicit |
| The factory was not well equipped to handle (the gas)$_{e_1}$ (created by the sudden addition of water to the MIC tank.)$_{e_2}$ | Implicit |

Table 1: Explicit and implicit causal relations.

| Sentence | Meaning |
|---|---|
| An undercroft is traditionally a cellar or storage room, often brick-lined and vaulted, and used for storage in buildings *since* medieval times. | Temporal |
| Additionally if one is to use a large scan range then sensitivity of the instrument is decreased due to performing fewer scans per second *since* each scan will have to detect a wide range of mass fragments. | Causal |
| In stark contrast to his predecessor, five days *after* his election he spoke of his determination to do what he could to bring peace. | Temporal |
| Bischoff in a round table discussion claimed he fired Austin *after* he refused to do a taping in Atlanta. | Causal |

Table 2: Different meanings of the prepositions *since* and *after* taken from the AltLex corpus.

an interaction between two entities, namely *affector* and *patient*. This definition can be extended in order to fit our definition. The affector and the patient will be the causing ($e_1$) and the caused ($e_2$) events, and the interaction between them will correspond to the lexical marker ($l$) in case of explicit causal relations, and the context in implicit scenarios. Table 1 shows the difference between explicit and implicit causal relations. So, following the cause dynamics model of Wolff (2007), we define causality as $e_1 \xrightarrow[l]{CAUSE} e_2$. The next sentence from the test set used in our experiments is an example of our definition of causality:

*A government affidavit in 2006 stated that (the leak)$_{e_1}$ ([caused]$_l$ 558,125 injuries, including 38,478 temporary partial injuries and approximately 3,900 severely and permanently disabling injuries.)$_{e_2}$*

The presence of a lexical marker does not ensure that the meaning of an utterance is causal, because they are usually ambiguous. An example is the adverb *since*, which can have a temporal or a causal meaning, as Table 2 shows.

We define the task of causality classification as a task composed of two subtasks: *causal meaning classification* and *causal arguments identification*. Given two events, the task of causal meaning classification is to disambiguate the causal meaning (`Causal` or `Non Causal`) of the relation of those two events. The task of causal argument identification focuses on the identification of the causing ($e_1$) and caused ($e_2$) events. We contribute to the first subtask.

## 3.2 Data

According to the definition of causal meaning classification, we need a corpus in which the events are annotated, as well as the lexical markers that can trigger the causal meaning in case of explicit relations, for the classification of the causal meaning of an utterance. Thus, our method requires that the input utterance is composed of the two events of the relation and optionally the lexical marker ($e_1$, $l$, $e_2$).

The AltLex corpus (Hidey and McKeown, 2016) meets the requirements of our task. The corpus was built on the idea that causality can be expressed by different types of linguistic constructions. This is validated by the fact that in PDTB there are explicit causal lexical markers, and other kinds of expressions that have a discourse meaning, which are called AltLex (Alternative Lexicalization). The relations signalled by an AltLex expression are a kind of implicit relation, in which the causal meaning is projected by an expression that is not part of common discourse connectives. The relations with an AltLex expression are those ones in which the annotators did not find an appropriate lexical marker to insert between

| Corpus | Version | Causal | Non-Causal | Total |
|---|---|---|---|---|
| Training | non-bootstrapped | 7,606 | 79,290 | 86,896 |
| | bootstrapped | 12,534 | 88,210 | 100,744 |
| Dev. | | 181 | 307 | 488 |
| Test | | 315 | 296 | 611 |

Table 3: Number of instances in the AltLex corpus.

| | Non-Boots. | Boots. |
|---|---|---|
| Unambiguous in `Non Causal` class | 7171 | 7673 |
| Unambiguous in `Causal` class | 922 | 1034 |
| Ambiguous | 121 | 147 |
| Total | 8214 | 8854 |
| `Causal` in train, `Non Causal` in test | 0 | 27 |
| `Non Causal` in train, `Causal` in test | 0 | 8 |

Table 4: Distribution of the lexical markers.

the events, because the meaning of the causal relation is entailed by an AltLex expression. From the existence of AltLex expressions in PDTB one can deduce that there are more expressions that entail the causal meaning of an utterance. Hence, the authors of the AltLex corpus developed a method to identify a larger amount of AltLex expressions that can trigger the causal meaning of an utterance. The corpus construction leveraged Simple Wikipedia by aligning sentences from Wikipedia that consist of unknown lexical causal markers with sentences from Simple Wikipedia that contain corresponding known lexical causal markers. The result was a set of sentences with expressions that trigger their causal meaning. Once a first set of causal and non-causal sentences were identified, a *bootstrapping* method was applied to enhance the corpus. We call the first version of the corpus "non-bootstrapped" and the second one "bootstrapped". The corpus statistics are in Table 3. More details in Hidey and McKeown (2016).

Since one feature of the corpus is the annotation of the lexical markers that may express causation, we studied their class distribution. Table 4 shows the class distribution of the lexical markers, as well as the number of unarguable ones and the number of lexical markers with mostly a different meaning in the training and the test set. According to Table 4, there are few ambiguous lexical markers: 121 in the "non-bootstrapped" corpus and 147 in the "bootstrapped" version. However, there is an important difference between the two versions of the corpus: the class distribution of the lexical markers in the training and test set is the same in the "non-bootstrapped" version, whereas in the "bootstrapped" version is not. This fact means that the instances of the "bootstrapped" version of the corpus present a higher difficulty for the classifier, because there are some lexical markers with a dissimilar class distribution in the training and test set. We show that our system works on those instances in § 4.3.2.

### 3.3 Disambiguation of the Causal Meaning

According to our definition of causality as the relation of two events ($e_1 \xrightarrow[l]{CAUSE} e_2$), we propose a neural network architecture with two inputs (see Figure 1). The first input matches the first event ($e_1$), and the second one corresponds to the lexical marker ($l$) and the second event ($e_2$), which are separated by a special character. In case there is no lexical marker (implicit relation), the second input is composed of a special character and the second event.

The classification starts with the tokenization of the two inputs. The lengths ($n$, $m$) of the instances of each input are not necessarily the same, so in order to make their lengths equal, three zero-padding strategies were assessed, namely the maximum, the mean and the mode of the lengths ($t$) of the components of the inputs (see Equation 1).

For each word, its corresponding word vector of 300 components ($d$) was looked up in the 840b cased Glove embeddings (Pennington et al., 2014). Subsequently, the concatenated word embeddings get passed through an encoding Long Short-Term

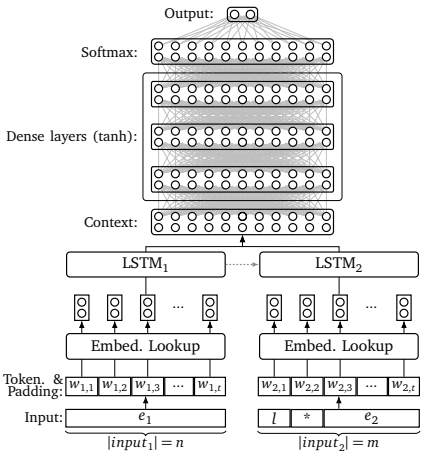


Figure 1: Neural model, where $e_1$ is the first event, $l$ is the lexical marker and $e_2$ is the second event.

Memory (LSTM) recurrent neural network (RNN) (Hochreiter and Schmidhuber, 1997) layer. We decided to use LSTM because of its ability to encode sequential and contextual information (Melamud et al., 2016). We assume some sort of relation exists between $e_1$ and $e_2$, so we first evaluated the performance of the connection of the two LSTM layers through the initialization of the second LSTM with the end state of the first one (dashed arrow in Figure 1). We call this model "Stated_Pair_LSTM". We assessed the same model but without the connection of the two LSTMs for evaluating our assumption. We call it "Pair_LSTM" (no dashed arrow in Figure 1).

The two outputs of the encoding layer are transformed to a vector of length 100 by a dense layer with a *tanh* activation function. The context of the causal relation is represented by the concatenation of the two vectors (see Equation 2). The output of Equation 2 is processed by three dense layers activated by a *tanh* function. The last layer is composed of the *softmax* operation.

$$\begin{aligned} &\forall e_1 \in \mathbb{R}^{n \times d}, \forall e_2 \in \mathbb{R}^{m \times d} \text{ and } r \in \{n, m\} \\ &pad(e) : \mathbb{R}^{r \times d} \rightarrow \mathbb{R}^{t \times d} \\ &t \in \{max(e), mean(e), mode(e)\} \end{aligned} \quad (1)$$

$$\begin{aligned} &\forall W \in \mathbb{R}^{100 \times nd} \text{ and } \forall b \in \mathbb{R}^{100} \\ &vec(e) : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}^{nd} \\ &tanh(W \cdot e + b) \\ &context : (vec(e_1), vec(e_2)) \end{aligned} \quad (2)$$

The performance of two learning optimizers with their default learning rates was evaluated, specifically Adadelta (Zeiler, 2012) and Adam (Kingma and Ba, 2015). Different values for dropout ($[0.5, 0.75]$) and $L^2$ regularization ($\left[8 \cdot 10^{-3}, 8 \cdot 10^{-6}\right]$) were evaluated to avoid overfitting.

# 4 Experiments and Results

As far as we know, the AltLex corpus has been only used in Hidey and McKeown (2016), so we consider their classification method as the state-of-the-art on that corpus. We have used the two versions of the corpus ("non bootstrapped" and "bootstrapped") in our experiments.

## 4.1 Baselines

We compare our proposal with two baselines. The first one (B1) assigns the most common class of each lexical marker in the training data, and it is similar to the baseline used in Hidey and McKeown (2016). The second baseline (B2) is the system of Hidey and McKeown (2016), which is based on SVM with a large set of features generated from the original parallel corpus and some lexical resources (WordNet, VerbNet and PropBank). Since relying on lexical resources restricts the recall of the system to their linguistic coverage, we propose a neural model fed only by a set of word embedding vectors.

## 4.2 Results

Table 5 displays the performance of the different configurations of our system and the baselines.[2] The precision, recall and F1 values were used to measure the performance of the system in the `Causal` class (C), and the accuracy to measure the overall performance of the system.

The performance of B1 defines a hard baseline for the two versions of the corpus, which might indicate that the training corpus is composed of few ambiguous causal connectives, which is expected given the statistics of the corpus in Table 4. However, the proposed systems outperform B1, which means that the systems learn beyond the class distribution of the lexical markers in the training data.

Those configurations that use Adadelta as optimizer outperform the system B2 in the "non bootstrapped" version of the corpus. The low value of precision of B2 means that it returns a large number of false positives. Rather, the high precision of our proposed approach indicates a good classification of sentences with unambiguous lexical markers, as it is expected since there are very few ambiguous lexical markers (see Table 4). The best configuration ("Pair_LSTM_Max_Adadelta") uses the `max` operation for

[2]For the sake of brevity, those systems that performed worse than the three baselines are not listed in Table 5.

| Training corpus | Method | Precision C. | Recall C. | F1 C. | Accuracy |
|---|---|---|---|---|---|
| Non-bootstrapped | B1 | 68.92% | 54.92% | 61.13% | 63.99% |
| | B2 | 70.28% | **77.60%** | 73.76% | 71.86% |
| | Stated_Pair_LSTM_Mean_Adadelta | **90.04%** | 60.31% | 72.24% | 76.10% |
| | Stated_Pair_LSTM_Mode_Adadelta | 89.23% | 63.17% | 73.97% | 77.08% |
| | **Pair_LSTM_Max_Adadelta** | 88.46% | 65.71% | **75.40%** | **77.90%** |
| | Pair_LSTM_Mean_Adadelta | 89.33% | 63.80% | 74.44% | 77.41% |
| Bootstrapped | B1 | 74.38% | 86.66% | 80.05% | 77.74% |
| | B2 | 77.29% | 84.85% | 80.90% | 79.58% |
| | Stated_Pair_LSTM_Max_Adadelta | 78.69% | 84.44% | 81.47% | 80.19% |
| | Stated_Pair_LSTM_Mean_Adam | 80.00% | 82.53% | 81.25% | 80.36% |
| | Stated_Pair_LSTM_Mode_Adam | 80.24% | 82.53% | 81.37% | 80.52% |
| | Pair_LSTM_Max_Adadelta | 78.07% | 84.76% | 81.12% | 79.86% |
| | Pair_LSTM_Mean_Adadelta | 78.48% | 85.71% | 81.94% | 80.52% |
| | **Pair_LSTM_Mean_Adam** | **80.30%** | 82.85% | 81.56% | **80.68%** |
| | Pair_LSTM_Mode_Adam | 77.24% | **87.30%** | **81.96%** | 80.19% |

Table 5: Results of the baselines and the different configurations of our neural model.

| Trainig corpus | Method | Precision C. | Recall C. | F1 C. | Accuracy |
|---|---|---|---|---|---|
| Non-bootstrapped | B1 | 68.92% | 54.92% | 61.13% | 63.99% |
| | B2 | 70.28% | 77.60% | 73.76% | 71.86% |
| | Pair_LSTM_Max_Adadelta | 88.46% | 65.71% | 75.40% | 77.90% |
| | Pair_LSTM_0dense_Max_Adadelta | **88.84%** | 65.71% | 75.54% | 78.06% |
| | **Pair_LSTM_4dense _Mode_Adadelta** | 88.52% | **68.57%** | **77.28%** | **79.21%** |
| Bootstrapped | B1 | 74.38% | **86.66%** | 80.05% | 77.74% |
| | B2 | 77.29% | 84.85% | 80.90% | 79.58% |
| | Pair_LSTM_Mean_Adam | 80.30% | 82.85% | 81.56% | 80.68% |
| | Pair_LSTM_2dense_Mode_Adam | 79.57% | 84.12% | 81.79% | 80.68% |
| | Pair_LSTM_0dense_Mean_Adam | 80.74% | 82.53% | 81.63% | 80.85% |
| | Pair_LSTM_1dense_Mean_Adam | 80.18% | **86.49%** | 81.80% | 80.85% |
| | Pair_LSTM_4dense_Mode_Adam | 80.06% | 84.12% | **82.04%** | 81.01% |
| | **Pair_LSTM_0dense_Mode_Adadelta** | **81.44%** | 82.22% | 81.83% | **81.17%** |

Table 6: Results of the evaluation of the influence of the number of dense layers.

the zero-padding strategy and Adadelta as learning optimizer. Our proposal improves B2 by 2.13% and 7.72% according to F1 and accuracy respectively.

Our assumption about the relation of the meaning of the two inputs does not hold, due to the better performance of the architecture "Pair_LSTM" with both versions of the corpus. Accordingly, it is better to independently encode each argument and then to measure the relation between the arguments by using dense layers. When the "bootstrapped" version of the corpus is used as training data, the optimizer Adam returns more homogeneous results between precision and recall, which indicates a better disambiguation of different expressions of causality. Although B2 outperforms our method in terms of recall by 2.41%, overall our system performs better in terms of F1 score, as B2 tends to classify many instances as false positives. To conclude, the best configuration ("Pair_LSTM_Mean_Adam") uses the `mean` strategy for zero-padding and the optimizer Adam, and yields an improvement of 3.89% in precision over B2.

According to Conneau et al. (2017), the number of dense layers influences the performance of neural models in text classification tasks, so we also made that analysis in the task of causality classification. We evaluated the performance of the best neural model, "Pair_LSTM", with a different number of dense layers, specifically from 0 to 4. The results in Table 6 show that 1) the combination of four dense layers and `mode` as padding strategy substantially increases the results compared to "Pair_LSTM_Max_Adadelta" when the "non bootstrapped" corpus is used as training set, and 2) the efficiency of the method is also improved because the `mode` operation reduces the length of the input. When the "bootstrapped" version of the corpus is used as training data, the number of layers also influences the performance. In this case,

the best performance is reached when no dense layers are used, and the `mode` operation is the non-zero padding strategy. This configuration is more efficient than "Pair_LSTM_Mean_Adam" because the length of the input vector of the model is shorter and less dense layers are used. Therefore, we conclude that the number of dense layers influences the performance of causality classification.

## 4.3 Analysis

In this section we present analyses of the proposed model from several points of views: an evaluation of the proposal in a balanced version of the datasets (§ 4.3.1), and a qualitative analysis of the classification of three different groups of lexical markers (§ 4.3.2).

### 4.3.1 Balanced Training Set

The `Causal` class is only the 8.7% and the 12.4% of all the instances in both versions of the corpus respectively (see Table 3). This big difference between the two classes may affect the performance of the classification, because it may separate the two classes and hence may ease the classification. So, we reduced the number of instances of the `Non Causal` by a factor of ten with the aim of evaluating our system with a more balanced dataset (see Table 8). The results of the evaluation of B1 and our best configurations are in Table 7. The results show that B1 follows a similar trend in the two versions of the corpus, which is a big difference between the precision and the recall. In contrast, our proposal not only outperforms the baselines, but it also yields a better balance between precision and recall. Our proposal significantly improves B1 according to McNemar's test ($p < 0.001$ and $p < 0.05$ respectively).

### 4.3.2 Distribution of Lexical Markers

Three groups of lexical markers were identified in the test set: 1) ambiguous lexical markers, those ones that are not in the training set or the difference of the probability to belong to each class is less than 10% according to their distribution in the training set (`Ambiguous`); 2) opposite meaning lexical markers, those ones whose probability distribution in the training set is opposite to their probability distribution in the test set (`Opposite`); and 3) the rest of lexical markers. We compare the performance of B2 and our best system in those clusters in order to know the strengths and weaknesses of our proposal.

Table 9 shows that B2 reaches a better performance with `Ambiguous` lexical markers, which means that we have to continue working on improving the representation of the context for the classification of unseen lexical markers. On the other hand, our neural model performs better with those lexical markers of the `Opposite` cluster, which means that our proposal rightly leverages the context of each lexical marker, and results in a higher capacity of generalization than B2. Our proposal also tends to classify better those instances with lexical markers that are not `Ambiguous` or `Opposite`.

Table 10 shows some instances from the test set of AltLex corpus whose lexical markers belong to the cluster `Opposite`, so they mostly have a different class in the training and the test set. Those examples are correctly classified by our best configuration ("Pair_LSTM_0dense_Mode_Adadelta") using the "bootstrapped" corpus as training data, and they are misclassified by B2. Table 10 shows the class of the instances in the training set (Training column) and in the test set (Test column), as well as the output of B2 and our proposal. The case of the verb *break* is noteworthy since *break* is a causative verb, and it mostly has a causative interpretation in the training data. However, there are other uses of *break* without a causal meaning, as the example shows in Table 10. *Make* is another example of a causative verb, and our proposal also correctly disambiguates it, while B2 does not. These positive results with causative verbs encourage us to research the subtleties of these kind of verbs. Due to the better performance of our proposal on the examples showed in Table 10 and the comparison of Table 9, we can conclude that our neural model learns beyond the class distribution of the training instances, so it has the ability of generalizing the causal meaning. The last conclusion allow us to confirm our claim, i.e. the use of linguistic features restricts the representation of causality, and a neural model only fed with word embeddings performs better in the task of causality classification.

| Trainig corpus | Method | Precision C. | Recall C. | F1 C. | Accuracy |
|---|---|---|---|---|---|
| Non-bootstrapped | B1 | 63.70% | 84.12% | 72.50% | 67.10% |
| | **Pair_LSTM_4dense_Mode_Adadelta** | **73.96%** | 79.36% | **76.56%** | **74.95%** |
| Bootstrapped | B1 | 67.34% | 94.28% | 78.57% | 73.48% |
| | **Pair_LSTM_0dense_Mode_Adadelta** | **72.27%** | 88.57% | **79.60%** | **76.56%** |

Table 7: Results of B1 and our best configurations with the downsampled version of the corpus.

| Corpus | Causal | Non Causal | Total |
|---|---|---|---|
| Non-bootstrapped | 7,606 | 7,929 | 15,534 |
| Bootstrapped | 12,534 | 8,821 | 21,354 |

Table 8: Size of the reduced version of the AltLex corpus.

| Lexical markers | Total | B2 | Our proposal |
|---|---|---|---|
| Ambiguous | 344 | 303 | 284 |
| Opposite | 92 | 48 | 64 |
| Rest | 181 | 127 | 150 |

Table 9: Lexical markers correctly classified by B2 and our proposal using the bootstrapped corpus for training.

| Sentence from the test set | Training | Test | B2 | Our proposal |
|---|---|---|---|---|
| The United States decided to *break* off economic relations with Cuba (which means that they would stop buying things from them). | Causal | Non Causal | Causal | Non Causal |
| Although Roosevelt had promised *to keep* the United States out of the war, he nevertheless took concrete steps to prepare for war. | Causal | Non Causal | Causal | Non Causal |
| Mary spent the next 18 years in confinement, but proved too dangerous *to keep* alive, as the Catholic powers in Europe considered her, not Elizabeth, the legitimate ruler of England. | Causal | Non Causal | Causal | Non Causal |
| Greatly alarmed and with Hitler *making* further demands on the Free City of Danzig, Britain and France guaranteed their support for Polish independence; when Italy conquered Albania in April 1939, the same guarantee was extended to Romania and Greece. | Causal | Non Causal | Causal | Non Causal |
| They are purely written languages and are often *difficult* to read aloud. | Causal | Non Causal | Causal | Non Causal |

Table 10: Some correctly classified examples by our best configuration that were misclassified by B2.

# 5 Conclusions and Future Work

We divided the task of causation classification into two subtasks: causal meaning classification and causal argument classification. The paper focused on the task of causal meaning classification, and we claim that the encoding of the two events of the relation is required for a suitable disambiguation of causality. We proposed an encoding system based on a neural network with two inputs, one for the first event and the other for the lexical marker and the second event. Our proposed system outperforms the state-of-the-art on the AltLex corpus. We also showed the success of the system in some non-causative sentences but with commonly causative verbs (see Table 10).

The task of causality classification lacks corpora not restricted to specific syntactic constructions (see § 2) and balanced corpora with a good coverage of causal instances (see § 4.3.1). Therefore, for future work, we plan the creation of a new corpus for the two subtasks of causality classification, namely causality disambiguation and causality argument classification.


# Acknowledgements

This work was supported by the German Research Foundation through the German-Israeli Project Cooperation (DIP, grant DA 1600/1-1 and grant GU 798/17-1). Calculations for this research were conducted on the Lichtenberg high performance computer of the TU Darmstadt.

## References


Bethard, S. and J. H. Martin (2008). Learning semantic links from a corpus of parallel temporal and causal relations. In *Proceedings of the 46th Annual Meeting of the ACL on Human Language Technologies: Short Papers*, HLT-Short '08, Stroudsburg, PA, USA, pp. 177–180. Association for Computational Linguistics.

Conneau, A., H. Schwenk, L. Barrault, and Y. Lecun (2017, April). Very deep convolutional networks for text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Valencia, Spain, pp. 1107–1116. Association for Computational Linguistics.

Copley, B. and P. Wolff (2014). *Causation in Grammatical Structures*, Chapter Theories of causation should inform linguistic theory and vice versa, pp. 11–57. Oxford Scholarship Online.

Girju, R. (2003). Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering - Volume 12*, MultiSumQA '03, Stroudsburg, PA, USA, pp. 76–83. Association for Computational Linguistics.

Hidey, C. and K. McKeown (2016, August). Identifying causal relations using parallel wikipedia articles. In *Proceedings of the 54th Annual Meeting of the ACL (Volume 1: Long Papers)*, Berlin, Germany, pp. 1424–1433. Association for Computational Linguistics.

Hochreiter, S. and J. Schmidhuber (1997, November). Long short-term memory. *Neural Comput. 9*(8), 1735–1780.

Khoo, C. S. G., J. Kornfilt, R. N. Oddy, and S. H. Myaeng (1998). Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. *Literary and Linguistic Computing 13*(4), 177–186.

Kingma, D. P. and J. Ba (2015). Adam: A method for stochastic optimization. In *3rd International Conference for Learning Representations, San Diego, 2015*.

Kruengkrai, C., K. Torisawa, C. Hashimoto, J. Kloetzer, J.-H. Oh, and M. Tanaka (2017, February). Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI-17)*, San Francisco, California, USA, pp. to appear.

Lewis, D. (1973). Causation. *Journal of Philosophy 70*(17), 556–567.

Melamud, O., J. Goldberger, and I. Dagan (2016). context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pp. 51–61.

Mirza, P. and S. Tonelli (2016, December). Catena: Causal and temporal relation extraction from natural language texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan, pp. 64–75. The COLING 2016 Organizing Committee.

Neeleman, A. and H. van de Koot (2012, May). The linguistic expression of causation. In M. Everaert, M. Marelj, and T. Siloni (Eds.), *The Theta System: Argument Structure at the Interface*, pp. 20–51. Oxford University Press.

Pennington, J., R. Socher, and C. D. Manning (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

Prasad, R., N. Dinesh, A. Leeand, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber (2008, may). The penn discourse treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Reinhart, T. (2002). *The Theta System: Syntactic Realization of Verbal Concepts*. Cambridge, Mass: The MIT Press.

Riaz, M. and R. Girju (2013, August). Toward a better understanding of causality between verbal events: Extraction and analysis of the causal power of verb-verb associations. In *Proceedings of the SIGDIAL 2013 Conference*, Metz, France, pp. 21–30. Association for Computational Linguistics.

Riaz, M. and R. Girju (2014, June). In-depth exploitation of noun and verb semantics to identify causation in verb-noun pairs. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, Philadelphia, PA, U.S.A., pp. 161–170. Association for Computational Linguistics.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science 12*(1), 49–100.

Wolff, P. (2007). Representing causation. *Journal of experimental psychology: General 136*(1), 82.

Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. *CoRR abs/1212.5701*.