IDENTIFICATION OF CAUSAL DISCOURSE RELATIONS IN FRENCH TEXT USING MACHINE-TRANSLATED TRAINING RESOURCES

JEAN CONSTANTIN

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

8,117 words

STUDENT NUMBER

401349

COMMITTEE

Grzegorz Chrupała
Gonzalo Nápoles

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

July 15, 2022

IDENTIFICATION OF CAUSAL DISCOURSE RELATIONS IN FRENCH TEXT USING MACHINE-TRANSLATED TRAINING RESOURCES

JEAN CONSTANTIN

Abstract

Causal discourse relations are fundamental building blocks of a coherent speech. Identifying these relations automatically could help deepen text analysis, and improve natural language understanding systems. Despite progresses in English, little research has been dedicated to the subject in other languages due to a lack of annotated training data. This thesis evaluates a low-cost solution to this problem: machine-translation. A standard English dataset for discourse relation is automatically translated to French and used to train French causal relation classification models. These models obtained F1-scores well-above random baselines in identifying causal relations in original French text with a recurrent neural network-based model (RNN) scoring 0.314 and a BERT-based model scoring 0.599. Specifically, the RNN model was only able to identify explicit causal relations while the BERT model obtained high scores on both explicit and implicit relations. These results suggest that machine-translated data can indeed be used to train French deep-learning models for complex tasks like causal relations classification. However, larger and more advanced models like BERT appear to be better equipped to process implicit relations.

Data source and ethics statement

Work on this thesis did not involve collecting data from human participants or animals. The datasets and codes mentioned in this thesis were acquired from free and open-access sources, the author of this work does not have any legal claim on them. The code and resources used and developed in this thesis are publicly available at the following repository: https://github.com/JeanConstantin/causality-detection. The figures and graphs were all created by the author.

1 PROBLEM STATEMENT & RESEARCH GOALS

1.1 *Context*

Causality refers to the logical link existing between two connected arguments: a cause and its consequence. This logical concept is one of the foundations of scientific reasoning and argumentation. *"What led to this event? Why did this event happen? What were its consequences?"* are all examples of causal thinking which could apply to any field. Beyond the natural intuition behind causality, it is necessary to give a more precise definition. A common framework to understand causality is "counterfactual reasoning": two events are causally related if the second event could not have happened, had the first event not taken place. In texts, interlocutors or readers assess whether two arguments exhibit causal links by asking *"What if the first argument had not happened?"*, *"Can the second argument exist without the first one?"*. Causality is an essential element in the construction of meaningful argumentation and can be expressed in many ways. For instance, in scientific publications, causal mechanisms can be described using schemas and logical symbols. However, in most cases, causality is conveyed through argumentation and the use of specific grammatical structures.

The automatic identification of such arguments in texts falls under the more general task of discourse relation classification. Discourse relations are a set of possible logical relations characterizing the link between two meaningful chunks of texts or arguments. Discourse relation types may vary across models and languages, but some categories such as causality, contrast or concession seem to be shared by most research projects (Prasad et al., 2008). This exercise is recognized as particularly difficult, yet essential for natural language understanding (NLU) (Roze et al., 2019). Literature reviews like Yang et al. (2021) show that the causal discourse relation extraction task has been a research topic since the 1990's. Most of the recent research has relied on English annotated resources like the well-known Penn Discourse Tree Bank (PDTB) and has yet to be tested on other languages. Like many other complex natural language processing (NLP) tasks, these efforts have been impaired by a lack of quality annotated corpora in languages other than English. Annotated resources like discourse relations are very costly as they require linguistics expertise and long hours of manual work. Even though French is a widely spoken language, the only available resources annotated with discourse relations are too small to effectively train a deep learning model.

Despite these difficulties, causal relation extraction could open new research opportunities in many areas including social sciences and policy

research. Indeed, policy research heavily relies on causal reasoning to evaluate a policy's impacts, unforeseen consequences and mechanisms. These analyses traditionally rely on econometrics or sociological methods like interviews. Both show limits: while the first is only suited for tabular numeric data and often fails to capture more subtle information, the latter is labor intensive, small-scaled and inconsistent. To overcome these limits, governments have been gathering large amounts of text data through consultations and online platforms in which citizens share their opinions and policy propositions. This wealth of data contains very useful information for policymakers but is rarely exploited to its full extent due to its volume and highly unstructured nature. For example, several citizen consultations organized by the French Parliament yielded more than 100,000 free-form written answers with some participants developing long argumentations. However, these answers were not analyzed globally and only a few random excerpts were quoted in the final reports. Such limited analyses disincentivize participation and can raise doubts over the quality and representativity of the final conclusions. NLP systems could help governments better leverage text data and extract relevant information like causal arguments. However, such systems remain unavailable in "rarer languages" like French.

This thesis will explore NLP methods to automatically identify causal argument pairs in French texts. As mentioned above, the lack of French training data is a major obstacle to this task. To solve this issue, past researchers have relied on multilingual models (Kurfalı and Östling, 2021) as well as large English models and texts automatically translated to English from other languages (Isbister et al., 2021a; Vu and Moschitti, 2021). This thesis will seek to evaluate another potential solution: training a French causal relation classification model using a large annotated English corpus machine-translated to French. Continuous research on discourse relation classification has led to a rapidly growing number of approaches ranging from simple rule-based algorithms to state-of-the-art deep learning models. Like other NLP tasks, discourse relation classification has widely benefited from the development of transformer-based models like BERT. This work will assess the performances gains associated with BERT models compared to classical approaches like recurrent neural networks (RNN) on causal argument identification in French texts. While basic models have been demonstrated to work well on explicit discourse relations, mixed results have been observed for implicit relations. In many cases, the logical link between two arguments is explicited through connectives or conjunctions like "because", "hence", "as a result", etc. When such cues are absent, one may only rely on context to determine the nature of the relation between the two arguments. A model should be able to detect both

types of relations, hence this thesis will evaluate models' performances for both explicit and implicit relations.

1.2 *Research Questions*

RQ1. To what extent does a causal relation classifier trained with English data machine-translated into French transfer to original French text?

RQ2. How does such a model's performances vary when classifying explicit and implicit causal discourse relations in French?

RQ3. Do BERT models perform better than recurrent neural networks (RNN) models in identifying causality in French?

1.3 *Main findings*

Models trained with the machine-translated data and evaluated on original French texts all reported F1-scores above random baselines, showing that machine-translated data can indeed generalize to original French texts. The BERT models achieved the highest scores in all categories: reason, result, implicit and explicit relations. At the opposite, the RNN model was only able to capture the more obvious explicit relations. Overall, a significant drop in performances was observed for both models when transferring from machine-translated data to original French data. This suggests that machine-translated data transfers only partially to original French.

2 LITERATURE REVIEW

2.1 *Discourse relation classification: from rule-based algorithm to deep learning models*

Early research used explicit discourse connectives ("because", "since", etc.) to characterize the logical relations between two consecutive arguments. Pitler et al. (2008) noted that 46.75% of all causal discourse relations in the Penn Discourse Tree Bank (PDTB, Prasad et al. (2008)) are explicit. They also showed that most of the discourse connectives used in this type of relation are unambiguous, they are not used to express other relations. For instance, 100% of "because" instances indicate a causal relation, this proportion is 100%, 100%, 95.99% and 52.17% for "so", "thus", "if" and "since" respectively. In this way, by relying only on the connectives' distribution, they reached a F1-score of 0.89 in identifying explicit causal argument pairs.

However, rule-based classifications are unable to tackle cases where explicit connectives are absent or ambiguous. Not relying on explicit connectives requires classification models to understand two consecutive arguments' deeper meaning and interactions. Early attempts made use of statistical models like Naïve Bayes to find co-occurrences of semantic patterns in the training data. For example, the co-occurrence of 'disease' and 'death' is a frequent causal pattern which a model can identify. Most of the research has focused on improving arguments' representation to solve the sparsity problem. Pitler et al. (2009) extracted lexical features from arguments, while Biran and McKeown (2013) clustered words into categories. Models are then able to capture more generalizable patterns.

Zhang et al. (2015) introduced one of the first deep learning model aimed at characterizing implicit discourse relations between arguments and obtain an overall F1-score of 0.52 in identifying causal argument pairs. An important aspect of their strategy was the use of Word2Vec embeddings which were pre-trained on a large unlabeled corpus. Pre-trained word representations are particularly useful for implicit relations classification as they bring external knowledge to the model. Relations between words are encoded in their embeddings such that a model would be able to capture patterns more easily and with less data. Word embeddings are characterized with a structure that allows analogies operations like the famous "queen – woman + man = king". Similar analogies are also likely to characterize common relations of causality between words which a deep learning a model could capture. More sophisticated architectures were then developed. Cai and Zhao (2017), Bai and Zhao (2018) and Sun et al. (2019) all relied on a similar approach, where arguments are encoded using pre-trained word embeddings and fed to bi-RNN layers to contextualize the word representations. The two argument representations are concatenated and undergo convolutional layers to finally be sent to a classification layer. An important aspect of Cai and Zhao (2017) and Bai and Zhao (2018)'s approaches is the bi-argument attention module, which allows the two arguments to "look at each other" and highlight the parts interacting the most while reducing the noise.

Shi and Demberg (2019) were the first to apply a BERT model to the discourse relation classification task. They noted that BERT's next sentence prediction (NSP) training is particularly adapted to discourse relation classification as it forces the model to form discourse expectations, defined as "the typical causes, consequences, next events or contrasts following a given event described in the first argument" (p.2, Shi and Demberg (2019)). For example, a first argument such as: *"He missed his train"*, raises expectations for an explanation in the second argument: *"He woke up late"*, *"Road traffic was terrible"*, *etc.* Kishimoto et al. (2020) confirmed BERT's

efficiency in discourse relation classification and reported a F1-score of 0.72 in identifying causal relations. The current state-of-the-art performances were obtained by Ma et al. (2021)'s "graph attention network". Their architecture encodes arguments into BERT embeddings which are then sent to bi-argument attention layer. This layer differs from Bai and Zhao (2018) as it relies on a graph obtained from dependency parsing. The graph creates interactions between the arguments' elements which are grammatically and lexically connected.

Research on discourse relation classification in French has been comparatively rare with most attempts centered on explicit relations. Garcia (1997) developed an early tool to identify causal arguments in French based on a list of explicit verbs and cues. Following the same logic, Roze et al. (2012) released a list of 328 French connectives with their corresponding relations. 23% of connectives were found to be ambiguous, meaning that they were associated to several relation types.

2.2 *Approaches to the lack of training data*

The deep learning models cited above require vast amounts of training data. However, corpora annotated with discourse relations are still rare as they require expansive hand labelling. Most annotated resources are still small and in English, like the well-known Penn Discourse Tree Bank (PDTB). The lack of annotated data is a heavy constraint in other languages as illustrated by the almost inexistent research on French discourse relations. Many researchers have attempted to solve this issue by automatically building training datasets through distant supervision. Marcu and Echihabi (2001) were the first to introduce an unsupervised method to automatically mine training examples. They collected sentences with explicit connectives and masked them to create implicit training examples. They hypothesized that connectives are often redundant and can be removed without affecting the sentence meaning. Using a naïve bayes classifier, they obtained performances well above the baseline on the classification of argument pairs with masked connectives. Sporleder and Lascarides (2008) suggested, however, that this type of distant supervision fails to generalize to natural data due to structural differences between implicit and explicit relations which a simple masking of the connective fails to capture. Yet, Nie et al. (2019) showed that deep learning models like BERT are able to learn from explicit training examples and transfer to implicit relations. (Shi et al., 2019) took advantage of parallel corpora in multiple languages and automatically extracted natural implicit training examples by identifying pairs explicitly connected in one language but without such cues in the parallel language. Hence, distant supervision offers a particularly interesting solution to the

lack of training data in other languages like French. Braud (2016) mined French artificial training examples following Marcu and Echihabi (2001)'s protocol and observed that statistical models trained with this kind of data do not generalize well to natural implicit relations.

Another approach to train models for rarer languages has been multilingual models: language models trained on multiple languages at the same time. Lample and Conneau (2019) pre-trained XLM models with Wikipedia articles in 15 languages ranging from English to Bulgarian and Urdu. The first model was trained with a unsupervised corpus of texts in multiple languages, while the second model was trained using a corpus of texts with their parallel translations. The second dataset with parallel translations is thought to force the model to learn similar word representation across languages. Both models obtained state-of-the-art result in natural language inference tasks across all 15 languages. Kurfalı and Östling (2021) were the first evaluate the performances of multilingual models on discourse relation classification. They fine-tuned multilingual models for the discourse relation classification task using an English annotated corpus and then performed a "zero-shot" evaluation on other languages. "Zero shot" means that the model is able to perform discourse relation classification on languages it was not exposed to during training, thus transferring knowledge across languages. They showed that zero-shot classification obtains results above random baselines, but that the performance gap when transferring to new languages is high and significant. Surprisingly, they also noted a very weak correlation between the performance drop and the proximity between the training and the zero-shot evaluation languages. Isbister et al. (2021b) explored automatic translation to leverage the more powerful English models on other "low resources" languages. They translated Swedish, Danish, Norwegian and Finnish evaluation datasets into English and then performed a sentiment classification task using English models. They reported better performances than models trained in the original language for Swedish, Norwegian and Danish which are all close to English. In the case of Finnish, which belongs to a completely different linguistic family, the method did not improve performances compared to the native BERT model. Vu and Moschitti (2021) fine-tuned a transformer model to perform a question answering task using a machine-translated English dataset and reported an average 10% drop in accuracy in the target language compared to English.

The current state of the research in discourse relation classification shows that the task is still far from resolved, with causality as one of the most difficult relation to identify. The development of deep learning models, pre-trained language models, and transformer models like BERT have brought important breakthroughs. However, most of these progresses

have been concentrated on English due to a lack of annotated data in other languages like French. Solutions like distant supervision as well as multilingual models have been explored and have yielded mixed results so far. Machine-translated annotated data appears as a low-cost yet effective solution to train models in rarer languages. This approach has yet to be tested on a complex task such as causal discourse relation classification. This thesis will thus seek to provide a comprehensive assessment of the performances of French models trained with machine-translated data from English to French. It will also compare RNN based models to state-of-the-art BERT models. Finally, even though explicit discourse relation classification has been considered solved for many years, one may question the ability of machine-translated data to properly capture the grammatical characteristics of such relations in the target language. Thus, this thesis will also assess the performances of French models trained with machine-translated data on explicit and implicit causal relations distinctly.

3 METHODOLOGY AND EXPERIMENTAL SETUP

3.1 *Problem formulation*

The data is composed of N argument pairs with argument pair i described as:

$$Pair_i = \{A_{i,1}, A_{i,2}\}$$

More specifically, an argument A_i is a chunk of text of varying length which contains a free-standing idea, concept, description or action in the discourse. It is also known as an elementary discourse unit (EDU). In most cases, $A_{i,1}$ and $A_{i,2}$ are two consecutive arguments in the original raw text.

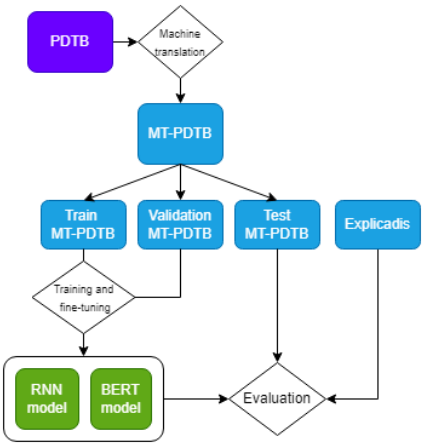
The causal discourse relation identification task is studied through two sub-tasks. In task 1, a model is trained to label an argument pair $Pair_i$ as either “causal” or “non-causal”. This binary classification problem is addressed by most research papers in discourse relation classification, in this way the models’ performances can be compared.

In task 2, models are trained to perform a more granular classification task. “Causal” arguments are split into two subclasses: “reason” and “result”. These two subclasses capture the direction of causality within the argument pair. The “result” class indicates that the first argument is the cause of the second argument, while “reason” points to the inverse relation. Task 2 is thus a three-way classification model, where a model labels an argument pair $Pair_i$ as “non-causal”, “reason” or “result”. Past research in discourse relation classification has focused on higher levels categories pooling "reason" and "result" types together like in task 1. However, a

model that does not make this distinction would have little practical use. Task 2 is thus motivated by the need for finer-grained categories.

For both tasks, models will be trained on an annotated English corpus which was machine-translated into French and will be evaluated on both translated data and original French data. Figure 1 schematically represents the complete methodology adopted in this thesis.

Figure 1: Methodology diagram



3.2 *Causal discourse relation classification models*

3.2.1 *RNN model*

The first model evaluated in this thesis (Fig. 1), is based on the architecture described by Cai and Zhao (2017), Bai and Zhao (2018) and Sun et al. (2019), which was the predominant approach before BERT models. The model is described in details below as it combines different elements of the mentioned approaches.

Arguments in argument pair $Pair_i = \{A_{i,1}, A_{i,2}\}$ are tokenized using SpaCy's transformer French model[1]. Their lengths are normalized to 50 tokens each through truncation and padding. Bai and Zhao (2018) transform words into continuous representations using a combination of pre-trained word and sub-word embeddings. Following their example, each token are converted into pre-trained FastText French embeddings with 300 dimensions. These French word embeddings[2] were pre-trained on internet text data from the Common Crawl dataset and on Wikipedia articles. FastText combines word and character n-gram representations,

[1] SpaCy model "fr_dep_news_trf" available at: https://spacy.io/models/fr
[2] Available at: https://fasttext.cc/docs/en/crawl-vectors.html

in this way, when an unknown word is encountered, an embedding can be obtained from the sub-word n-grams. This choice is motivated by the morphological complexity of the French language. Contrary to English, French has many conjugations which would require lemmatization to avoid unknown words. Lemmatization would, in turn, result in the loss of verb tenses which provide valuable information in chains of events like causal relations. The embedding layer output is:

$$Pair_i^{(emb)} = \{E_{i,1}, E_{i,2}\}$$

where $E_{i,1}, E_{i,2} \in \mathbb{R}^{50\times300}$ are the pre-trained continuous representations of $A_{i,1}$ and $A_{i,2}$. The two arguments representations are then sent to two argument-specific biGRU encoders to contextualize their representations. The biGRU encoder can be formulated as follows:

$$y_i = BiGRU(E_i) \in \mathbb{R}^{50\times(2\times300)}$$

The biGRU encoder outputs a forward and a backward sequence which are combined in a linear layer.

$$F_i = y_i \cdot W + b \in \mathbb{R}^{50\times300}$$

This is done for both arguments in parallel so that,

$$Pair_i^{(context)} = \{F_{i,1}, F_{i,2}\}$$

Similarly to Bai and Zhao (2018), bi-argument attention weights are computed. This module allows the two arguments to “look at each other" and highlights the parts which interact the most.

$$AttW_i = (W \cdot F_{i,1} + b) \cdot F_{i,2}^T \in \mathbb{R}^{50\times50}$$

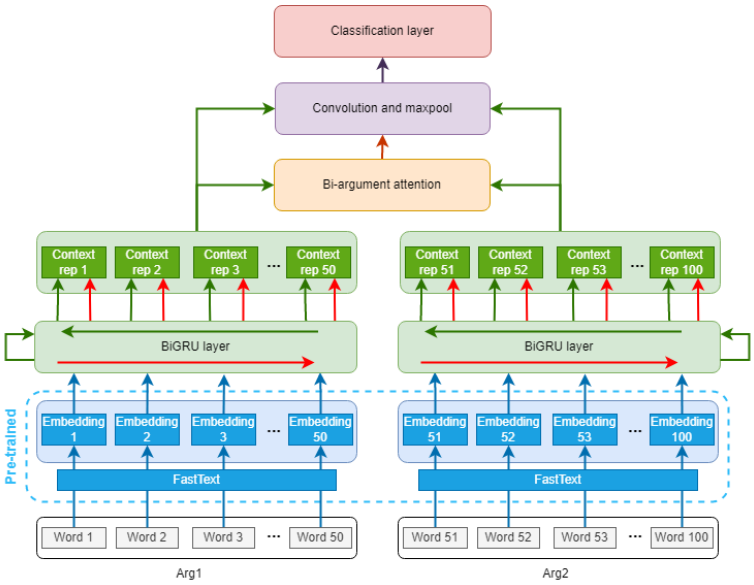
Attention weights are then applied to the arguments representations.

$$R_{i,1} = softmax(AttW_i^T) \cdot F_{i,1} \in \mathbb{R}^{50\times300}$$
$$R_{i,2} = softmax(AttW_i) \cdot F_{i,2} \in \mathbb{R}^{50\times300}$$
$$Pair_i^{(representation)} = \{R_{i,1}, R_{1,2}\}$$

The two arguments’ representations are then concatenated and fed to a series of convolutional layers and a final maxpooling layer. The maxpooling layer applies a top-K max-over-sequence operation where the K highest values in the sequence are selected for each dimension to obtain $O_i \in \mathbb{R}^{K\times300}$. The optimal K is determined by experience and is set to 10. O_i is flattened and sent to a final classification layer which outputs class

probabilities. The model is built and implemented using the Python library Pytorch, its main structure was adapted from Bai and Zhao (2018)'s freely available code. [3]

Figure 2: RNN model



3.2.2 *BERT model*

The second model evaluated in this thesis is the transformer-based model BERT following Kishimoto et al. (2020). BERT (Bidirectional Encoder Representations from Transformers) is a general model developed by Devlin et al. (2019). BERT relies on the transformer architecture introduced by Vaswani et al. (2017). A transformer is composed of an encoder and a decoder, both contain multiple stacked multi-attention heads modules. A simple classification layer is then added on top of the decoder in order to predict the next element of a decoded sequence. Contrary to previous NLP models which mostly relied on recurrence to analyze word sequences, transformers process an entire sequence at once. Thus, the position of each word in the sequence needs to be encoded with a "positional encoding" and fed to the model along with the text tokens.

BERT is a language model, it encodes word sequences into continuous representations. Hence, it is only composed of a transformer's encoder. BERT models are trained on two unsupervised tasks. In the masked

[3] Available under MIT Licence at: https://github.com/diccooo/Deep$_E$*nhanced*$_R$*epr*$_f$*or*$_I$*DRR*

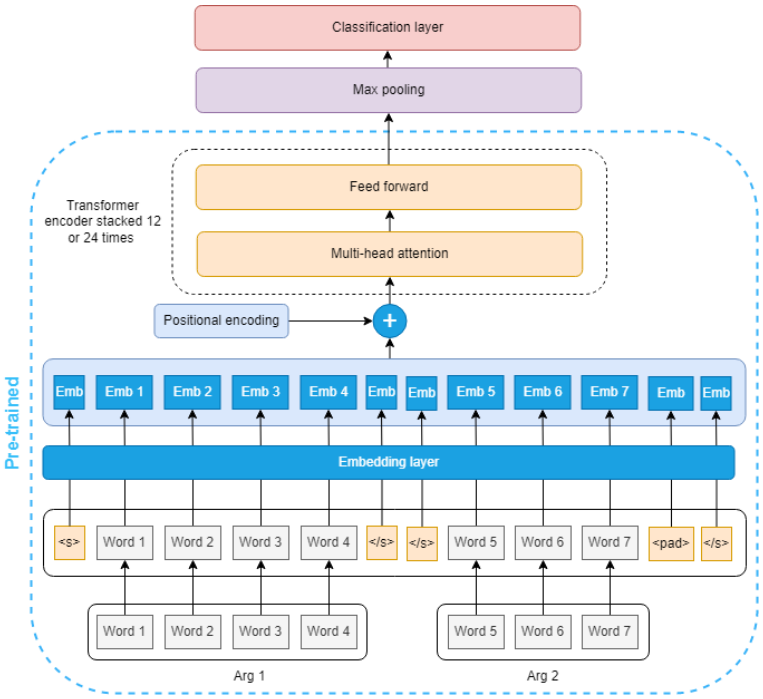
language model task (MLM), 15% of a sequence's words are randomly masked, the model then attempts to predict them. The second task, next sentence prediction (NSP), teaches the model to identify whether a second sentence originally followed the first or was taken from another part of the corpus. As argued by Shi and Demberg (2019), the NSP task is closely related to discourse relation classification. The original BERT models developed by Devlin et al. (2019) were pre-trained on more than 3 billions words and were found to be particularly powerful and versatile. BERT models have been fine-tuned to many NLP tasks and have delivered major breakthroughs outperforming the task-specific and often very complex previous models as the one described by Bai and Zhao (2018) for example.

In this thesis, a pre-trained French BERT model will be fine-tuned to detect causal discourse relations. Martin et al. (2020) released CamemBERT [4], a BERT model pre-trained on 32.7 billions tokens of raw French text obtained from internet sources. The main datasets used for the model's pre-training were the CommonCrawl and Oscar datasets which gather a wide range of topics and modes of expression. The model was trained using the masked language modelling (MLM) task. Two versions of CamemBERT were released by Martin et al. (2020) and are evaluated in this thesis: CamemBERT-base with 12 layers, 12 attention heads and 110 millions parameters, and CamemBERT-large with 24 layers, 16 attention heads and 335 millions parameters.

Argument pairs are pre-processed using a CamemBERT specific tokenizer. Words are separated into tokens, and unknown words are split into their sub-words parts. Once transformed into token sequences, the two arguments are concatenated and padded or truncated to 200 tokens in total. Special BERT tokens are added to the sequence to mark the arguments' edges: <s> marks the beginning of the sequence, while </s> marks the end of each argument. Tokens are then encoded into pre-trained embeddings and added to positional encodings. The sequence is then sent to a multi-head attention layer followed by a feed-forward layer which build contextualized word representations. This operation is repeated 12 or 24 times depending on the model used. The final output is maxpooled to reduce its size and fed to a classification layer which returns class probabilities. CamemBERT models are implemented and fine-tuned using the Python library Transformers.

[4] Available under MIT License at: https://camembert-model.fr/

Figure 3: CamemBERT model



3.3 *Datasets*

Training deep learning model for causal discourse relation classification requires large amounts of annotated data. However, discourse relation annotation is a long and complex task requiring linguistic expertise. Annotated corpora are thus rare and relatively small, particularly in other languages than English. Indeed, while the reference English corpus for discourse relations, the Penn Discourse Tree Bank (PDTB, Prasad et al. (2008)) contained more than a million words, the largest French discourse relation corpus, Annodis (Ho-Dac and Péry-Woodley, 2014), only contained 28,000 words. This thesis explores a low-cost and simple method to leverage English annotated data in French: machine-translation.

3.3.1 *Machine-translated Penn Discourse Tree Bank (MT-PDTB)*

Causal discourse relation classifiers will be trained using a version of the PDTB 2.0[5] translated automatically from English into French. The PDTB 2.0 was first published by Prasad et al. (2008) through the Linguistic Data Consortium. It is composed of 40 600 argument pairs drawn from Wall Street Journal articles and annotated with their discourse relations. Rela-

[5] Available under GPL-2.0 Licence at: https://github.com/cgpotts/pdtb2

tions are classified into 4 groups: temporal, contingency, comparison and expansion. The contingency group conveys the concept of causality and contains four subtypes: cause, pragmatic cause, condition and pragmatic condition. Condition relations express causality between hypothetical or unrealized events. *"If I get up early, I will not miss my train"* is a conditional statement where *missing the train* has not happened yet. This thesis focuses on realized events and will only consider the cause subtype which refers to non-conditional causal relations. The cause type is further split in two categories indicating the causal relation's direction: "result" indicates that the first argument is the cause and the second argument the consequence (Arg1 $\rightarrow$ Arg2), while "reason" indicates the inverse relation (Arg1 $\leftarrow$ Arg2). Finally, the PDTB also contains annotations on whether the relation was implicit or explicit as well as the connective used in explicit relations.

Explicit connectives had initially been removed for arguments pairs in the PTDB. The full arguments including these connectives were thus rebuilt from the PDTB's raw text. The PDTB's argument pairs were then automatically translated from English into French using LibreTranslate. LibreTranslate is a free and open-source machine-translation API [6]. It relies on OpenNMT, an attention-based seq-to-seq toolkit developed by Klein et al. (2020) and already used in several machine translation projects. Due to length limits set by LibreTranslate, 5,757 arguments pairs could not be processed leaving the machine-translated PDTB (MT-PDTB) with 34,843 argument pairs in French for training.

Table 1: Causal relations distribution in the MT-PDTB

	Reason	Result	Not causal	Total
Explicit	1,327	754	13,328	15,409
Implicit	2,143	1,504	15,787	19,434
Total	3,470	2,258	29,115	34,843

3.3.2 *Explicadis*

The causal discourse relation classifier is trained on the MT-PDTB and evaluated on the original French corpus Explicadis (Atallah, 2015). Explicadis is a corpus dedicated to causal discourse relations. It is based on Annodis, the first French corpus annotated with discourse relations (Ho-Dac and Péry-Woodley, 2014). Explicadis is composed of 110 texts including news articles, Wikipedia articles, scientific publications and geopolitical reports. This corpus' writing style is descriptive and formal which makes it compa-

[6] Available at: https://libretranslate.com/

rable to the PDTB. Explicadis contains 2,848 argument pairs annotated as either "result", "explanation" or "not causal". "Result" describes a situation in which the first argument caused the second argument (Arg1 $\rightarrow$ Arg2), while "explanation" corresponds to the PDTB's "reason" category and characterizes a relation where the second argument caused the first argument (Arg1 $\leftarrow$ Arg2). Explicadis also identifies explicit and implicit relations with their connectives. However, when the PDTB focused on a well-defined list of causal conjunctions, Explicadis' authors took a much broader approach and included other causal cues like expressions, verbs and verbal forms. In order to better align with the PDTB, implicit and explicit pairs were re-annotated for this thesis. The exhaustive list of French connectives[7] compiled by Roze et al. (2012) was used to flag potential explicit discourse relations, which were then reviewed by hand.

Table 2: Causal relations distribution in the Explicadis

	Reason	Result	Not causal	Total
Explicit	83	151	778	1,012
Implicit	100	42	1,694	1,836
Total	183	193	2,472	2,848

3.3.3 *Connectives in explicit relations*

The classification of explicit discourse relations in English is considered as a solved problem, as a result recent research has solely focused on implicit relations. However, natural text often contain both types of relations, and a causal relation classification model should be able to handle both cases. Moreover, Zufferey and Cartoni (2012) suggested that the expression of causality strongly differs between French and English in terms of connectives and grammatical structures. Connectives annotations in the PDTB and Explicadis offer an interesting assessment of the linguistics differences between French and English.

Figure 4 shows that in English, a large majority of explicit reason relations use the same connectives: "because" and "as". The result relation shows more diversity. In French (Fig. 5), the reason relation appears to rely on a more varied set of connectives than in English. This observation supports Zufferey and Cartoni (2012)'s conclusions that French tends to represent causality in more diverse ways. This quick analysis shows that the classification of explicit discourse relations in French may not be as trivial. It also raises doubts over the use of machine-translated data from

[7] Available under Creative Commons License at: http://www.linguist.univ-paris-diderot.fr/ croze/D/Lexconn.xml

Figure 4: Connective distribution in explicit causal relations (PDTB)

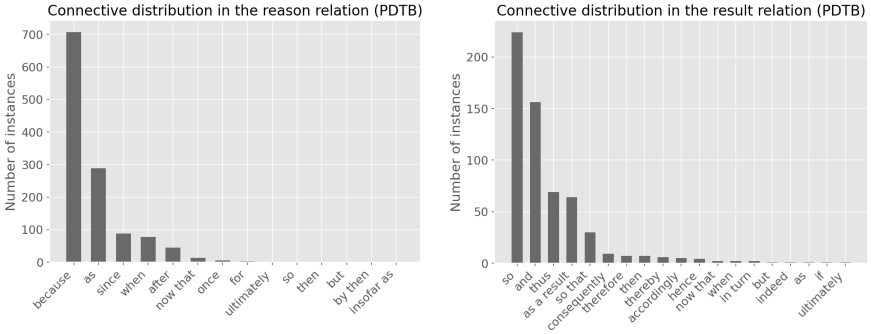
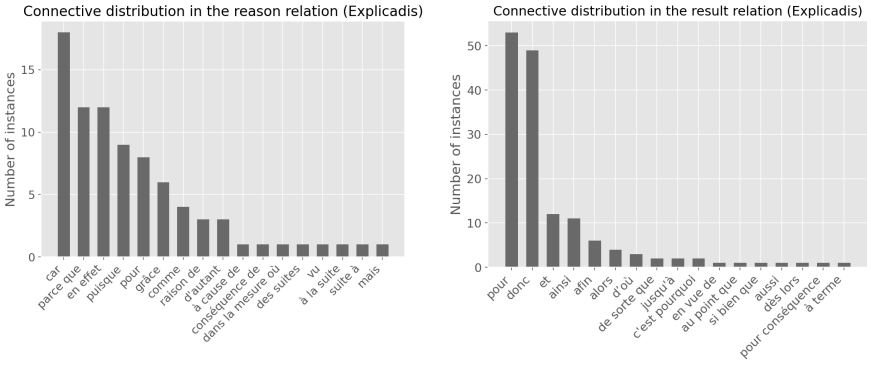


Figure 5: Connective distribution in explicit causal relations (Explicadis)



English to French for the model's training. Indeed, a simple translation may fail to properly reproduce the full range of causal expressions and grammatical structures in French.

3.4 *Evaluation methodology*

The two tasks studied in this thesis are classification problems which can be easily evaluated with standard indicators like accuracy. However, both the training and evaluation datasets are highly unbalanced with 83.6% of the MT-PDTB and 86.8% of the Explicadis dataset labelled as "not causal". Accuracy may not represent the models' performances appropriately. To counter this problem, the F1-score is used as the main performance indicator. The F1-score is computed by taking the harmonic mean of the precision and the recall scores.

$$F1score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$Precision = \frac{True\ positive}{True\ positive + False\ positive}$$

$$Recall = \frac{True\ positive}{True\ positive + False\ negative}$$

Precision is the proportion of real positive instances among instances predicted as positive. Recall, also known as sensitivity, captures the proportion of real positive values correctly predicted as such. Both indicators are combined with equal weights in the F1-score. For Task 1's binary classification problem, the F1-score is simply computed for the "causal" class. Task 2 is a three way classification, individual F1-scores are computed for the "reason" and "result" classes and are then averaged (macro average). Scores are compared to random baselines where classes are assigned randomly for all instances.

3.5 *Hyper-parameters fine-tuning*

Due to the size of the models and their necessary training times, the RNN and BERT models were fine-tuned using a simple holdout cross-validation approach. The MT-PDTB dataset was randomly shuffled and split into static training, validation and test datasets each accounting for 70%, 15% and 15% of the original dataset respectively. To deal with class imbalances, the splitting of the dataset is stratified so that class distribution is similar across the training, validation and test datasets. Models are trained using the training dataset and fine-tuned based on performances obtained on the validation dataset. To further address class imbalances, the training dataset was randomly resampled. For task 1, the "not causal" class was downsampled to 10,000 instances while the "causal" class was upsampled to 10,000 instances. Similarly for task 2, the "not causal" class was upsampled to 10,000 instances and the "reason" and "result" classes were upsampled to 5,000 instances each. The validation and test datasets were not resampled. Both models are trained using Google Colab's GPU and high-RAM environment. Grid-search outputs are available in the appendix (section A).

3.5.1 *Fine-tuning of the RNN model*

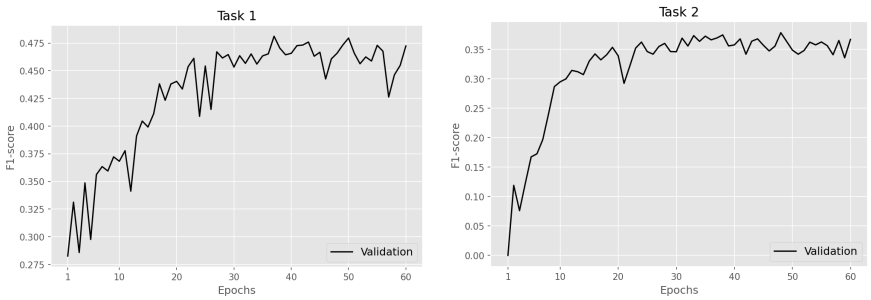
The first model is fitted on the resampled training dataset by optimizing a binary cross-entropy loss function for task 1 and a cross-entropy loss function for task 2. The selected optimizer is the widely used RMSProp (root mean square propagation). It is an adaptive learning algorithm which can speed up learning and reduce weight oscillations. Contrary to stochastic gradient descent, RMSProp keeps specific learning rates for each weight

and adapt their values to the gradient's movements. Weights are updated by mini-batches of a 100 instances for 20 epochs. The optimizer's learning rate and weight decay parameters were fine-tuned using a grid search approach. The learning rate determines the step size for the weights updates during the optimization. Weight decay, also known as L2 regularization, prevents overfitting by constraining the model's weights. Validation loss and F1-scores are reported in the appendix (tables 6 to 9) for all hyper-parameters combinations after 20 epochs. To reduce the impact of random variations, the displayed indicators are the average scores over the model's last three epochs.

For task 1's binary classification problem, validation loss is minimized and validation F1-score is maximized with a learning rate of 0.001 and a weight decay of 0.00005. The same hyper-parameters combinations are tested for task 2 via grid search in similar conditions. As expected, overall results on task 2's three way classification problem are lower than for task 1. The optimal set of hyper-parameters is also less obvious. A weight decay of 0.00005 and a learning rate of 0.001 minimize the loss, but lead to a poor F1-score. The selected combination maximizes the F1-score at 0.342, it corresponds to a learning rate of 0.0005 and a weight decay of 0.00005.

The number of training epochs is another important hyper-parameter to consider. A model trained for a too short amount of time would invariably underfit the data and score poorly on both the training and the validation data. At the opposite, a model trained for too long would overfit the data. As a result, the model would display high scores on the training dataset and low scores on the validation dataset. To estimate the optimal number of epochs for each task, models are trained on the resampled training dataset with the fine-tuned parameters obtained above over 60 epochs. Figure 6 shows the evolution of the validation F1-score at every epoch. For task 1, the validation F1-score reaches its maximum, around 0.45, after 30 epochs. For task 2, the validation F1-score stagnates around 0.35 after 35 epochs. Hence, the models' training are early-stopped at 30 and 35 epochs for task 1 and task 2 respectively.

Figure 6: F1-score evolution during RNN model's training



3.5.2 *Fine-tuning of the BERT model*

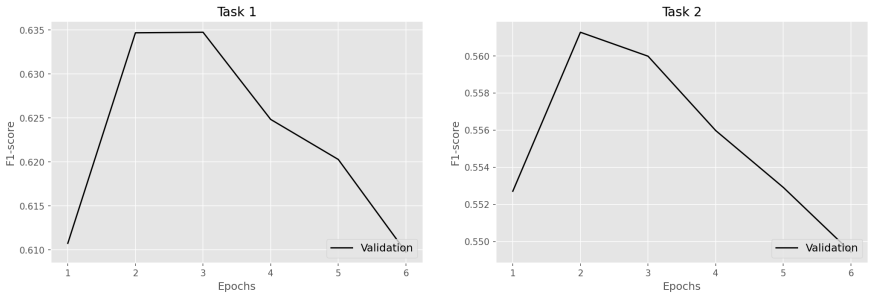
CamemBERT, a pre-trained French language model is imported and fine-tuned on the resampled training dataset to identify causal discourse relations. A cross-entropy loss function is optimized with the Adam algorithm over mini-batches of five instances. Adam maintains adaptive per-parameter learning rates. By combining aspects of both the momentum and the RMSProp approaches, gradient descent is accelerated toward the minimum while prevented from oscillating. Learning rate and weight decay are fine-tuned using a grid search approach. Results on the validation dataset are reported in the appendix (table 10 to 13) after two epochs using the smaller CamemBERT-base model.

In task 1, a combination of 0.00001 for the learning rate and of 0.01 for the weight decay appears as optimal as it maximizes the validation F1 score at 0.582 and yields a relatively low loss of 0.607. The CamemBERT model comes in two sizes, CamemBERT-base which was used in the fine tuning and contains 110 millions trainable parameters, and CamemBERT-large which contains 335 millions parameters. CamemBERT-large is trained using the optimal hyper-parameter combination obtained above and yields a F1-score of 0.640 for a loss of 0.691 on task 1. CamemBERT-large raises the F1-score by 10.0% compared to CamemBERT-base and will thus be used instead.

For task 2, a weight decay of 0.01 optimizes both the validation loss and the F1-score. A learning rate of 0.00001 maximizes the F1-score. Hence, the best hyper-parameter combination for this model is a learning rate of 0.00001 and a weight decay of 0.01 for task 2. CamemBERT-large is trained with the optimal set of hyper-parameters defined above for task 2. After 2 epochs, it obtains a loss of 0.850 and a validation F1-score of 0.561, which corresponds to a 11.1% performance increase relative to CamemBERT-base.

To determine the optimal training time, CamemBERT-large models are trained on the resampled training dataset with the fine-tuned hyper-parameters over six epochs (Fig. 7). For task 1, the validation F1-score reaches its maximal value, 0.635 on the third epoch. The model's performance follows a similar pattern for task 2 and is maximized on the second epoch at a value of 0.561. Thus, the models' training will be early-stopped after three and two epochs for task 1 and task 2 respectively.

Figure 7: F1-score evolution during BERT model's training



4 RESULTS AND DISCUSSION

4.1 *Results*

The RNN model and BERT model are evaluated on two annotated datasets. The first is a test subset originating from the MT-PDTB which the models have not been exposed to during either training or fine-tuning. The second dataset is the original French corpus Explicadis introduced previously.

4.1.1 *Results on task 1*

Task 1 is a binary classification problem where argument pairs are labelled as either "causal" or "non-causal". The RNN and BERT models are trained with the fine-tuned hyper-parameters defined in the previous section. Table 3 reports F1-scores obtained by both models on a test set from the MT-PDTB as well as the original French Explicadis corpus. The RNN model obtains a F1-score of 0.450 on the MT-PDTB test set and 0.314 on the Explicadis corpus. These performances are modest but are still well above random baselines of 0.253 and 0.207. Most of these results are driven by performances on explicit relations for which the RNN model scored 0.465 on the MT-PDTB and 0.502 on Explicadis. F1-scores obtained on implicit relations, albeit lower, were also above the random baseline for both datasets with 0.441 for the MT-PDTB and 0.177 for Explicadis. The RNN model appears to be more performant in identifying explicit causal relations. More precisely, the RNN model displays higher recall scores and lower precision[8] for both datasets. Hence, the model is able to identify most of the true causal relations but also seems to make many false positive mistakes.

The BERT model performed better than the RNN model on all scores. Specifically, it improved F1-scores by 38% on the MT-PDTB test set and

[8] Recall and precision scores can be found in the appendix

91% on Explicadis to reach 0.623 and 0.599 respectively. The BERT model also obtains F1-scores above the RNN model and the random baseline on explicit and implicit relations in both datasets. Similarly to the RNN model, its displays high performances on explicit relations with F1-scores of 0.763 on the MT-PDTB test set and 0.744 on Explicadis. Performances on implicit relations are lower but remain well above the baselines with F1-scores of 0.552 on the MT-PDTB and 0.419 on Explicadis. Contrary to the RNN model, the BERT model obtained higher precision scores than recall, which means that it makes relatively fewer mistakes in labelling relations as causal.

Overall, both models show performance drops when transferring from the machine-translated data to original French data. This trend appears to be particularly strong and significant for the RNN model. It is also interesting to note that the F1-score for explicit relation does not decrease between the MT-PDTB and Explicadis, showing that automatic translated data from English offers a correct representation of the expression of causality in French.

Table 3: Causal F1-scores (task 1)

	MT-PDTB			Explicadis		
	All	Explicit	Implicit	All	Explicit	Implicit
Random baseline	0.253	0.257	0.234	0.207	0.335	0.133
RNN model	0.450	0.465	0.441	0.314	0.502	0.177
BERT model	0.623	0.763	0.552	0.599	0.744	0.419

4.1.2 *Results on task 2*

In task 2, the models are trained with their respective fine-tuned hyperparameters to classify argument pairs as "reason", "result" or "not causal". Table 4 shows F1-scores for the "reason" class. The RNN model scored 0.461 on the MT-PDTB test set and 0.200 on the Explicadis dataset, both above their respective random baselines. Like in task 1, the "reason" class F1-scores are higher for explicit relations: 0.510 on the MT-PDTB and 0.387 on Explicadis. The RNN model scores above the random baseline for implicit reason relations in the MT-PDTB, but appears unable to detect implicit "reason" relations in Explicadis'original French text.

Once again, the BERT model outperforms the RNN model on all indicators, with F1-scores of 0.636 on the MT-PDTB test set and 0.598 on Explicadis. The BERT model displays high performances in detecting explicit "reason" relations with F1-scores of 0.756 on the MT-PDTB and 0.764

on Explicadis. As expected, lower performances, albeit widely above the random baseline, are observed for implicit "reason" relations with F1-scores of 0.563 and 0.463 on the MT-PDTB and Explicadis.

Table 4: Reason class F1-scores (task 2)

	MT-PDTB			Explicadis		
	All	Explicit	Implicit	All	Explicit	Implicit
Random baseline	0.158	0.176	0.155	0.106	0.103	0.107
RNN model	0.461	0.510	0.424	0.200	0.387	0.087
BERT model	0.636	0.756	0.563	0.598	0.764	0.463

Both models show lower performances in identifying "result" relations (table 5) compared to "reason" relations. The RNN model obtains F1-scores of 0.280 on the MT-PDTB test set and 0.279 on Explicadis. The model performs similarly on implicit and explicit relations in the MT-PDTB but display large a performance gap for Explicadis with F1-scores of 0.506 and 0.054 for explicit and implicit relations. Similarly to previous results, the model is unable to detect implicit "result" relations in Explicadis and scores under the random baseline. The BERT model achieves again the best results with an overall F1-score of 0.537 on the MT-PDTB and 0.496 on Explicadis. Performances on explicit and implicit relations are all well-above the random baseline, with the highest scores obtained on explicit relations: 0.693 in the MT-PDTB and 0.598 im Explicadis. The BERT model obtains relatively lower results on implicit relations.

Both models obtains higher precision than recall in identifying the "reason" class. Similarly to task 1, the RNN model scores high recall for the result class in both datasets, it obtains scores similar or higher than the BERT model. It is interesting to note that the RNN model seems to perform very differently for the reason and the result class. At the opposite, the BERT model appears to be more consistent with precision slightly above recall.

Table 5: Result class F1-scores (task 2)

	MT-PDTB			Explicadis		
	All	Explicit	Implicit	All	Explicit	Implicit
Random baseline	0.096	0.121	0.115	0.117	0.216	0.030
RNN model	0.280	0.288	0.281	0.279	0.506	0.054
BERT model	0.537	0.693	0.459	0.496	0.598	0.299

4.2 *Discussion and potential improvements*

4.2.1 *Research questions and results*

The RNN and the BERT models were trained on an English corpus which was machine-translated into French. Both models scored significantly above random baselines when identifying causal relationships in original French data. These results indicate that machine-translated texts from English to French can indeed be used to train a French causal discourse relation classifier (RQ.1). However, the significant performance drops observed between the MT-PDTB test set and Explicadis suggest that machine-translated training data does not fully transfer to original French text (-30.2% for the RNN model and -3.9% for the BERT model in the binary task). For the finer-grained "reason" and "result" categories, similar performance drops are observed. However, the relatively low performance drop observed for the BERT model also suggests that larger transformers architecture models are better equipped to capture subtle implicit causality cues.

Obvious grammatical connectives indicating the nature of discourse relations make explicit relations easy to identify. At the opposite, models are expected to yield lower results on implicit relations. As anticipated in the literature, average F1-scores on implicit relations drop significantly compared to explicit relations (RQ.2). The performance gap between explicit and implicit relations is particularly important for the RNN model (-5% in the MT-PDTB and -65% in Explicadis compared to -27% and -44% for the BERT model) and appears to be wider in the Explicadis dataset for both models. These performance gaps can also be found for the "reason" and the "result" classes. They appear as particularly marked for the "result" class with F1-score dropping by -89% for the RNN model and -50% for the BERT model on the explicadis dataset. Overall, the BERT model is the only model able to outperform the baseline in identifying implicit relations.

Causal discourse relations are identified using a set of often ambiguous cues and noisy signals. In line with conclusions from past research, this thesis confirms the wide advantage of BERT models in discourse relation classification. Larger and more sophisticated models like BERT appear as better equipped to capture causality's subtle patterns, particularly when relations are not marked by an explicit connective. On average, the BERT model outperformed the RNN model by 56% on the MT-PDTB test set, and by 123% on the Explicadis dataset (RQ.3).

4.2.2 *Error analysis*

Even though explicit relation classification has been considered a trivial task, both models displayed high rates of misclassifications for this cate-

gory. As argued by Zufferey and Cartoni (2012), the expression of causality in French varies significantly from English. French exhibits more diverse grammatical structures which makes the task relatively more complex. As expected, most of the misclassified explicit relations had ambiguous connectives used in other relation types like "et" (and) or "après" (after). Significant noise was also introduced in the training data during the PDTB's machine-translation from English to French. In many instances, explicit connectives were poorly translated or deleted altogether. Ambiguous English connectives like "since", "as", "after", etc. possess multiple potential translations in French depending on the context. For example, the English connective "since" can convey a meaning similar to "from" or to "because". However, its literal French translation, "depuis" can only be interpreted as "from" and never marks causality. Such poor translations have blurred the structure of argument pairs and can easily erase their causal meaning. In turn, these inaccurate training examples caused false positives in Explicadis as the models wrongly learned to associate "depuis" to causality. Finally, many false positive cases seem to have been caused by the presence of a causal connective in one of the argument, while this connective did not actually characterize the relation. An example can be found in the PDTB: "**Arg1**:[*analysts are skeptical of it because it's carrying a lot of debt*] **Arg2**:[*We've gotten our costs down and we're better positioned for any cyclical downturn than we've ever been*]". Here, the connective "because" indicates a relation contained within the first argument but has no relation to the second argument.

Models were in general able to identify implicit causal relations when the two arguments clearly mentioned the same entities or concepts. Many false negatives cases affected argument pairs that were not obviously connected. In some instances, the first argument mentioned an entity, and the second argument only referred to it using an undefined pronoun like in the following PDTB example: "**Arg1**:[*CNN is my wire service*] **Arg2**:[*they're on top of everything*]". In other cases, the two arguments discussed seemingly separated events which could only be understood as linked using the surrounding context. Such cases are particularly difficult and even a human annotator would fail to classify them without more information. For instance, the argument pair "**Arg1**:[*Mr. Miller vetoed that*] **Arg2**:[*Even I can't understand all the footnotes*] requires context about Mr. Miller to properly understand the relation linking the two arguments.

Finally, the argument pairs' annotated labels constituted another source of uncertainty. Discourse relation annotation is a complex task, particularly for implicit relations and fine-grained categories like "reason" and "result". Scholman et al. (2021) confirmed that the inter-annotator agreement for implicit relations is low and well-under the agreement for explicit relations.

Atallah (2015) noticed that causal relations in the French discourse corpus Annodis had low inter-annotator agreement, which motivated the creation of Explicadis, an annotated corpus dedicated to causal relations. Hence, it is not surprising that a significant share of false positive cases could be argued to be indeed causal. The following argument pair "**Arg1**:[*On the other hand, Valley National tumbled 24%*] **Arg2**:[*after reporting a sizable third-quarter loss*]" is annotated as "temporal.succession" in the PDTB and was misclassified as "reason" by the BERT model. These two categories have very close definitions and could overlap depending on the context. Such false positive cases were also observed in the original French corpus Explicadis where the argument pair "**Arg1**:[*Décès de Guy Hosneld*], **Arg2**:[*suite à une longue et douloureuse maladie*]" [9] was misclassified by the BERT model as "reason" while its Explicadis label is "elaboration".

4.2.3 *Potential improvements*

This thesis has shown that an English corpus machine-translated into French can indeed be used to train French models for complex tasks such as causal discourse relation classification. However, poor translations of connectives, expressions and grammatical structures have also introduced noise in the training data which has had a negative impact over performances. Translation quality thus appears as the key element. The same protocol should be reproduced using multiple machine-translation systems to determine the most appropriate. These considerations might be particularly important for very low resource languages for which machine-translation systems are not as developed as English-to-French.

Most of the research in discourse relation classification has separated implicit and explicit relations. This thesis took a different approach and attempted to process both implicit and explicit relations at the same time. Indeed, natural texts display both types of relations and a model should be versatile enough to process them equally. However, the presence of connectives in a large share of argument pairs could explain the models' high performances on explicit relations and low performances on implicit relations. These obvious cues could have led the models to become "lazy" and focus their attention on connectives while leaving out weaker signals. In this way, machine-translated data should be tested for the training of distinct implicit and explicit discourse relation classifiers.

Discourse relation annotations are a complicated task which can lack coherence. Depending on the context, causality can easily be confused with other closely related relation types. Causality can also assume different meanings and forms across fields and topics. Hence, domain specific anno-

[9] Translation: "Death of Guy Hosneld, following a long and painful disease"

tated corpora could help train and evaluate better models for particular tasks like government policy evaluation, etc. For example, several medical corpora were annotated with causes and effects related to diseases, and drugs' adverse effects (Karimi et al., 2015).

5 CONCLUSION

Causality is fundamental concept in sciences, which allows to build knowledge on the world, and on how systems evolves and interact. It is also omnipresent in natural language and is necessary for speakers to make sense of their environments, explain change and argue. Past research has explored causality as one of the many discourse relations which characterize the logical links between the different parts of a coherent text. Recent deep learning approaches have shown encouraging results in the classification of discourse relations and particularly for implicit relations. However, most of the research on discourse relation classification has been carried on the same standard English corpora. Indeed, the training of discourse relation classifiers require large amounts of high quality annotated data that is currently only available in English. French, despite being a widely-spoken language, does not benefit from such resources.

This thesis has shown that machine-translated training resources from English to French can constitute a low-cost, yet efficient alternative to the lack of annotated training data in French. Two French causal relation classifiers were trained with translated data and obtained results significantly above random on original French texts. The first model, based on recurrent neural networks (RNN), was only able to identify explicit causal relations. The second model, a fine-tuned pre-trained French BERT model, obtained the best results and displayed high performances on explicit and implicit relations. However, poor and shallow translations have introduced a significant amount of noise in the training data which has negatively impacted both models' performances when transferring to original French data.

The alternative offered by machine-translated data should be tested on other languages. French and English are very close languages in which causality is expressed in similar ways. At the opposite, for distant languages like Chinese or Arabic, a simple machine-translation from English might fail to properly convey the different aspects of causality in the target language. This method should also be evaluated on rare languages with small speaker populations for which machine-translation systems are under-developed. Multi-lingual models constitute another solution to the lack of training data in "rarer" languages. Such models can be fine-tuned on original English data and can then be used on other languages without

further training. Hence, future research should compare this "zero-shot learning" approach to the machine-translation approach presented in this thesis.

REFERENCES

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber, "The Penn Discourse TreeBank 2.0." p. 8, Jan. 2008.

C. Roze, C. Braud, and P. Muller, "Which aspects of discourse relations are hard to learn? Primitive decomposition for discourse relation classification," in *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. Stockholm, Sweden: Association for Computational Linguistics, 2019, pp. 432–441. [Online]. Available: https://www.aclweb.org/anthology/W19-5950

J. Yang, S. C. Han, and J. Poon, "A Survey on Extraction of Causal Relations from Natural Language Text," *arXiv:2101.06426 [cs]*, Oct. 2021, arXiv: 2101.06426. [Online]. Available: http://arxiv.org/abs/2101.06426

M. Kurfalı and R. Östling, "Probing Multilingual Language Models for Discourse," in *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*. Online: Association for Computational Linguistics, 2021, pp. 8–19. [Online]. Available: https://aclanthology.org/2021.repl4nlp-1.2

T. Isbister, F. Carlsson, and M. Sahlgren, "Should we Stop Training More Monolingual Models, and Simply Use Machine Translation Instead?" *arXiv:2104.10441 [cs]*, Apr. 2021, arXiv: 2104.10441. [Online]. Available: http://arxiv.org/abs/2104.10441

T. Vu and A. Moschitti, "Multilingual Answer Sentence Reranking via Automatically Translated Data," *arXiv:2102.10250 [cs]*, Feb. 2021, arXiv: 2102.10250. [Online]. Available: http://arxiv.org/abs/2102.10250

E. Pitler, M. Raghupathy, H. Mehta, A. Nenkova, A. Lee, and A. Joshi, "Easily Identifiable Discourse Relations," p. 4, 2008.

E. Pitler, A. Louis, and A. Nenkova, "Automatic sense prediction for implicit discourse relations in text," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - ACL-IJCNLP '09*, vol. 2. Suntec, Singapore: Association for Computational Linguistics, 2009, p. 683. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1690219.1690241

O. Biran and K. McKeown, "Aggregated Word Pair Features for Implicit Discourse Relation Disambiguation," p. 5, 2013.

B. Zhang, J. Su, D. Xiong, Y. Lu, H. Duan, and J. Yao, "Shallow Convolutional Neural Network for Implicit Discourse Relation Recognition," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, 2015, pp. 2230–2235. [Online]. Available: http://aclweb.org/anthology/D15-1266

D. Cai and H. Zhao, "Pair-Aware Neural Sentence Modeling for Implicit Discourse Relation Classification," in *Advances in Artificial Intelligence: From Theory to Practice*, S. Benferhat, K. Tabia, and M. Ali, Eds. Cham: Springer International Publishing, 2017, vol. 10351, pp. 458–466, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-319-60045-1_47

H. Bai and H. Zhao, "Deep Enhanced Representation for Implicit Discourse Relation Recognition," *arXiv:1807.05154 [cs]*, Jul. 2018, arXiv: 1807.05154. [Online]. Available: http://arxiv.org/abs/1807.05154

K. Sun, Y. Li, D. Deng, and Y. Li, "Multi-Channel CNN Based Inner-Attention for Compound Sentence Relation Classification," *IEEE Access*, vol. 7, pp. 141 801–141 809, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8847427/

W. Shi and V. Demberg, "Next Sentence Prediction helps Implicit Discourse Relation Classification within and across Domains," 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)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 5789–5795. [Online]. Available: https://www.aclweb.org/anthology/D19-1586

Y. Kishimoto, Y. Murawaki, and S. Kurohashi, "Adapting BERT to Implicit Discourse Relation Classification with a Focus on Discourse Connectives," p. 7, 2020.

Y. Ma, Y. Yan, and J. Liu, "Implicit Discourse Relation Classification Based on Semantic Graph Attention Networks," in *The 5th International Conference on Computer Science and Application Engineering*. Sanya China: ACM, Oct. 2021, pp. 1–5. [Online]. Available: https://dl.acm.org/doi/10.1145/3487075.3487156

D. Garcia, "COATIS, an NLP system to locate expressions of actions connected by causality links," in *Knowledge Acquisition,*

Modeling and Management, J. G. Carbonell, J. Siekmann, G. Goos, J. Hartmanis, and J. van Leeuwen, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997, vol. 1319, pp. 347–352, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/BFb0026799

C. Roze, L. Danlos, and P. Muller, "LEXCONN: A French Lexicon of Discourse Connectives," *Discours*, no. 10, Jul. 2012. [Online]. Available: http://journals.openedition.org/discours/8645

D. Marcu and A. Echihabi, "An unsupervised approach to recognizing discourse relations," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*. Philadelphia, Pennsylvania: Association for Computational Linguistics, 2001, p. 368. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1073083.1073145

C. Sporleder and A. Lascarides, "Using automatically labelled examples to classify rhetorical relations: an assessment," *Natural Language Engineering*, vol. 14, no. 03, Jul. 2008. [Online]. Available: http://www.journals.cambridge.org/abstract_S1351324906004451

A. Nie, E. Bennett, and N. Goodman, "DisSent: Learning Sentence Representations from Explicit Discourse Relations," p. 14, 2019.

W. Shi, F. Yung, and V. Demberg, "Acquiring Annotated Data with Cross-lingual Explicitation for Implicit Discourse Relation Classification," p. 10, 2019.

C. Braud, "Identification automatique des relations discursives implicites à partir de corpus annotés et de données brutes," p. 239, 2016.

G. Lample and A. Conneau, "Cross-lingual Language Model Pretraining," *arXiv:1901.07291 [cs]*, Jan. 2019, arXiv: 1901.07291. [Online]. Available: http://arxiv.org/abs/1901.07291

T. Isbister, F. Carlsson, and M. Sahlgren, "Should we Stop Training More Monolingual Models, and Simply Use Machine Translation Instead?" *arXiv:2104.10441 [cs]*, Apr. 2021, arXiv: 2104.10441. [Online]. Available: http://arxiv.org/abs/2104.10441

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv:1810.04805 [cs]*, May 2019, arXiv: 1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *arXiv:1706.03762 [cs]*, Dec. 2017, arXiv: 1706.03762. [Online]. Available: http://arxiv.org/abs/1706.03762

L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, de la Clergerie, D. Seddah, and B. Sagot, "CamemBERT: a Tasty French Language Model," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020, pp. 7203–7219. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.645

L.-M. Ho-Dac and M.-P. Péry-Woodley, "Annotation des structures discursives : l'expérience ANNODIS," *SHS Web of Conferences*, vol. 8, pp. 2647–2661, 2014. [Online]. Available: http://www.shs-conferences.org/10.1051/shsconf/20140801286

G. Klein, F. Hernandez, V. Nguyen, and J. Senellart, "The OpenNMT Neural Machine Translation Toolkit: 2020 Edition," vol. 1, p. 8, 2020.

C. Atallah, "La ressource EXPLICADIS, un corpus annoté spécifiquement pour l'étude des relations de discours causales," p. 7, 2015.

S. Zufferey and B. Cartoni, "English and French causal connectives in contrast," *Languages in Contrast*, vol. 12, no. 2, pp. 232–250, Nov. 2012. [Online]. Available: http://www.jbe-platform.com/content/journals/10.1075/lic.12.2.06zuf

M. C. J. Scholman, T. J. M. Sanders, and Hoek, Jet, "Is there less annotator agreement when the discourse relation is underspecified?" *Proceedings of the Integrating Perspectives on Discourse Annotation (DiscAnn) Workshop*, p. 6, 2021.

S. Karimi, A. Metke-Jimenez, M. Kemp, and C. Wang, "Cadec: A corpus of adverse drug event annotations," *Journal of Biomedical Informatics*, vol. 55, pp. 73–81, Jun. 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1532046415000532

Appendices

A FINE-TUNING TABLES

Table 6: Validation loss during grid search (RNN model, task 1)

	Learning rate			
Weight decay		**0.0001**	**0.0005**	**0.001**
	0.00001	0.565	0.520	0.558
	0.00005	0.580	0.554	**0.357**
	0.0001	0.632	0.639	0.638

Table 7: Validation F1-score during grid search (RNN model, task 1)

	Learning rate			
Weight decay		**0.0001**	**0.0005**	**0.001**
	0.00001	0.411	0.457	0.448
	0.00005	0.394	0.436	**0.458**
	0.0001	0.383	0.391	0.192

Table 8: Validation loss during grid search (RNN model, task 2)

	Learning rate			
Weight decay		**0.0001**	**0.0005**	**0.001**
	0.00001	0.679	0.660	0.688
	0.00005	0.676	0.638	**0.616**
	0.0001	0.757	0.758	0.884

Table 9: Validation F1-score during grid search (RNN model, task 2)

	Learning rate			
Weight decay		**0.0001**	**0.0005**	**0.001**
	0.00001	0.311	0.320	0.318
	0.00005	0.287	**0.342**	0.211
	0.0001	0.283	0.305	0.088

Table 10: Validation loss during grid search (BERT model, task 1)

	Learning rate			
Weight decay		0.000001	0.000005	0.00001
	0.001	0.614	0.649	0.640
	0.01	0.606	0.634	0.607
	0.1	0.608	**0.597**	0.631

Table 11: Validation F1-score during grid search (BERT model, task 1)

	Learning rate			
Weight decay		0.000001	0.000005	0.00001
	0.001	0.401	0.546	0.577
	0.01	0.409	0.545	**0.582**
	0.1	0.398	0.550	0.579

Table 12: Validation loss during grid search (BERT model, task 2)

	Learning rate		
Weight decay	0.000001	0.000005	0.00001
0.001	0.721	0.581	0.696
0.01	0.742	**0.575**	0.690
0.1	0.735	0.575	0.688

Table 13: Validation F1-score during grid search (BERT model, task 2)

	Learning rate		
Weight decay	0.000001	0.000005	0.00001
0.001	0.157	0.484	0.499
0.01	0.148	0.483	**0.505**
0.1	0.142	0.488	0.501

B RECALL AND PRECISION

Table 14: Recall and precision (task 1)

	MT-PDTB		Explicadis	
	Recall	Precision	Recall	Precision
Random baseline	0.501	0.166	0.489	0.130
RNN model	0.587	0.411	0.490	0.234
BERT model	0.663	0.715	0.519	0.625

Table 15: Recall and precision for the reason class (task 2)

	MT-PDTB		Explicadis	
	Recall	Precision	Recall	Precision
Random baseline	0.323	0.097	0.317	0.059
RNN model	0.373	0.471	0.219	0.292
BERT model	0.633	0.694	0.563	0.691

Table 16: Recall and precision for the result class (task 2)

	MT-PDTB		Explicadis	
	Recall	Precision	Recall	Precision
Random baseline	0.313	0.061	0.373	0.080
RNN model	0.525	0.158	0.653	0.136
BERT model	0.529	0.641	0.487	0.599

C PDTB EXCERPTS AND THEIR MACHINE-TRANSLATIONS IN FRENCH

Explicit reason relation:
Arg1:[Longer maturities are thought to indicate declining interest rates] because **Arg2**:[they permit portfolio managers to retain relatively higher rates for a longer period]

Machine translation: **Arg1**:[On pense que les échéances plus longues indiquent une baisse des taux d'intérêt] parce qu' **Arg2**:[ils permettent aux gestionnaires de portefeuille de conserver des taux relativement plus élevés pour une période plus longue]

Implicit reason relation:
Arg1:[The agents will make more than routine inquiries about such items as marital status and dependents]; **Arg2**:[they want to look at living standards and business assets]

Machine translation: **Arg1**:[Les agents feront plus que des enqupêtes de routine sur des éléments tels que l'état matrimonial et les personnes à charge]; **Arg2**:[ils veulent regarder les niveaux de vie et les actifs commerciaux]

Explicit result relation:
Arg1:[The figures exclude lump-sum payments and cost-of-living adjustments,] so **Arg2**:[the actual wage increases may have been bigger]

Machine translation: **Arg1**:[Les chiffres excluent les paiements forfaitaires et les ajustements de coût de la vie,] donc **Arg2**:[les augmentations salariales réelles peuvent avoir été plus importantes]

Implicit result relation:
Arg1:[parents are returning to the cloth diaper.] **Arg2**:[Business is up 35% in the past year]

Machine translation: **Arg1**:[les parents retournent à la couche de tissu.] **Arg2**:[L'activité est en hausse de 35% l'année dernière]

D EXPLICADIS DATASET EXCERPTS

Explicit explanation (reason) relation: **Arg1**:[La tour 7 du WTC s'est effondrée dans l'après-midi] en raison **Arg2**:[d'incendies et des dégats occasionnés par la chute des Twin Towers.]
7 WTC collapsed in the afternoon, due to fires and damages caused by the fall of the Twin Towers.

Implicit explanation (reason) relation:
Arg1:[Le désenchantement est à la mesure de ces attentes.] **Arg2**:[Le roi ne semble se préoccuper que de réformes fiscales.]
Disenchantment was up to expectations. The king did not seem to bother with fiscal reforms.

Explicit result relation:
Arg1:[Elle n'y était pas parvenue à cette date.] Alors **Arg2**:[le tribunal lui avait accordé un ultime délai.]
She could not make it on this date. So, the tribunal granted her a final deadline.

Implicit result relation:
Arg1:[Ces attentats ont été vécus presque en temps réel par des centaines de millions de téléspectateurs à travers le monde,] **Arg2**:[le choc psychologique a été considérable au plan international.]
These terrorist attacks were experienced almost in real time by hundreds of millions of viewers around the World, the psychological shock was considerable at the international level.