SAFFRON: tranSfer leArning For Food-Disease RelatiOn extractioN

Gjorgjina Cenikj
Jožef Stefan International
Postgraduate School
Ljubljana, Slovenia
Computer Systems Department
Jožef Stefan Institute
Ljubljana, Slovenia
gjorgjina.cenikj@ijs.si

Tome Eftimov
Computer Systems
Department
Jožef Stefan Institute
Ljubljana, Slovenia
tome.eftimov@ijs.si

Barbara Koroušić Seljak
Computer Systems
Department
Jožef Stefan Institute
Ljubljana, Slovenia
barbara.korousic@ijs.si

Abstract

The accelerating growth of big data in the biomedical domain, with an endless amount of electronic health records and more than 30 million citations and abstracts in PubMed, introduces the need for automatic structuring of textual biomedical data. In this paper, we develop a method for detecting relations between food and disease entities from raw text. Due to the lack of annotated data on food with respect to health, we explore the feasibility of transfer learning by training BERT-based models on existing datasets annotated for the presence of *cause* and *treat* relations among different types of biomedical entities, and using them to recognize the same relations between food and disease entities in a dataset created for the purposes of this study. The best models achieve macro averaged F1 scores of 0.847 and 0.900 for the *cause* and *treat* relations, respectively.

1 Introduction

The ongoing prevalence of malnutrition, the rising incidence of chronic diseases affected by diet, and the fact that even food that is generally considered to be healthy can be harmful to patients suffering from certain diseases or when ingested in combination with specific drugs, require a profound understanding of the role of nutrition in the complex environmental interactions that contribute to the development or treatment of different ailments. The effect of food on human health is the subject of numerous biomedical studies, however, the sheer volume and the predominantly unstructured form of newly published articles prevents medical professionals from keeping up with recent discoveries, and impedes the development of systems for knowledge-base construction, Decision Support, and Question-Answering (QA), which brings about the need for information extraction (IE) methods for structuring the newly published knowledge.

Knowledge graphs (KGs) are specialized data representation structures that store information as a collection of interlinked descriptions of entities. The development of Relation Extraction (RE) methods is necessary for automatic linking of the nodes in KGs and reducing the amount of work required by the experts in order to create and extend these resources.

A lot of research effort has been dedicated to extracting relations between different biomedical entities, however, the lack of annotated data impedes the development of food-disease RE methods, which are necessary for linking food entities to concepts from the biomedical domain, and understanding the impact of nutrition on human health.

Transfer learning (TL) (Weiss et al., 2016; Zhuang et al., 2019) is a potential solution for this problem, which involves improving a learner from one domain by transferring information from a related domain. The use of TL in this paper is twofold. On the one hand, we use models that are pre-trained on large amounts of data, and fine-tune them for the RE task. On the other hand, we investigate the feasibility of re-purposing existing annotated IE resources in the biomedical domain as a potential strategy for making up for the deficit of such resources in the food domain.

We focus on the detection of *cause* and *treat* relations among food and disease entities, and represent the RE task as a binary classification problem, meaning that we train separate classifiers that detect the presence of each relation type. We perform fine-tuning of BERT (Devlin et al., 2018), BioBERT (Lee et al., 2019) and RoBERTa (Liu et al., 2019) models, which have achieved state of the art results in several Natural Language Processing (NLP) tasks.

To train the classifiers, we use the CrowdTruth (Dumitrache et al., 2017, 2015b,a) and Adverse Drug Events (ADE) (Gurulingappa et al., 2012) datasets, which contain sentences annotated

Proceedings of the BioNLP 2021 workshop, pages 30–40
June 11, 2021. ©2021 Association for Computational Linguistics

for the existence of relations between different types of biomedical entities. We then apply TL in order to use the classifiers trained on the source datasets to directly predict relations among food and disease entities. The reasoning behind the use of TL in this setting is that even though the sentences contain entities of different types, by masking the entity occurrences in the sentence, the models could use the context words around the entities and pick up on linguistic features such as keywords or sentence structure to detect the presence of a particular relation. Even though our goal is focused on detecting the relations between food and disease entities, we believe the method to be general enough to be applicable for entities of any type, as long as the relation is the same as the one the model was trained to recognize.

To evaluate the proposed models, we introduce a dataset of 608 sentences, which are extracted from abstracts of scientific articles from PubMed and are manually annotated for the presence of *cause* and *treat* relations between food and disease entities. To the best of our knowledge, this is the first English RE dataset in the food domain, and it is publicly available on GitHub [1], as an open-source resource that can be reused in future studies.

The rest of the paper is organized as follows. In the next section, we give an overview of the RE work in the domains of biomedicine, and food and nutrition. The data sources used for the experiments are described in Section 3. The text representation and classification models are presented in Section 4, while their evaluation is discussed in Section 5.

2 Related work

In the past decade, numerous methods have been developed for extracting biomedical relations, such as drug-drug (Dewi et al., 2017; Liu et al., 2016; Kim et al., 2015; Sahu and Anand, 2018), protein-protein (Koyabu et al., 2015; Fan et al., 2018; Zhou et al., 2018), drug-disease (Wang et al., 2017; Bchir and Karaa, 2013), chemical-gene (Lim and Kang, 2018) and chemical-protein (Lung et al., 2019) interactions.

In the domain of food and nutrition, the efforts directed at creating RE systems have been quite more limited in comparison. A gold standard for food RE has been generated for the German language (Wiegand et al., 2012b), and different methods such as distant supervision (DS), pattern-matching, and the use of co-occurrence measures have been investigated for the detection of food relations for customer advice (Wiegand et al., 2012a; Reiplinger et al., 2014). A Chinese food RE system (Miao et al., 2012) has also been developed, which treats RE as a sequence labeling task and adopts Conditional Random Fields (CRFs) models to extract relations between food and disease entities from Chinese biomedical data. However, resources in other languages are not easily re-purposed for the English language.

A related resource in the English language which contains extracted relations of food and disease entities is the NutriChem database (Jensen et al., 2014; Ni et al., 2017), which links plant-based foods with their small molecule components, drugs and disease phenotypes. A critical difference between NutriChem and the method we aim to develop in this work is the fact that NutriChem limits its scope to plant-based foods, while we do not pose a limitation on the type of foods or diseases between which the relations occur, and aim to extract relations from a broader range of food categories.

The benefits of TL have previously been investigated for the purposes of biomedical NER (Sun and Yang, 2019; Francis et al., 2019) and RE (Zhang et al., 2019; Peng et al., 2019; Hafiane et al., 2020). Recent work has been aimed at solving the challenges of imbalanced relation distribution, linguistic variation and partial transfer using relation-gated adversarial learning (Zhang et al., 2019), and capturing dependency tree information using TreeLSTM models (Legrand et al., 2018).

BERT has achieved state-of-the-art results on natural language processing (NLP) tasks, including RE between several types of biomedical entities, which is one of the tasks in the Biomedical Language Understanding Evaluation (BLUE) benchmark (Peng et al., 2019). A comparison of the performance of BERT models for detecting relations between proteins and chemicals, and genomic factors and drugs or drug responses (Hafiane et al., 2020), finds that depending on the target corpus, different variants of BERT may achieve the best results, and that fine-tuning the models is preferable over freezing the layers of the original model and only updating the weights of new layers added on top of the original ones. Guided by these findings, we perform fine-tuning of several BERT variants

[1] https://github.com/gjorgjinac/food-disease-dataset

for the RE task.

The Adverse Drug Events (ADE) corpus (Gurulingappa et al., 2012), which is one of the source datasets in our experiments, has been extensively used for training RE models, and more recently, for the exploitation of inter-task correlations for joint entity and relation extraction using different approaches, such as adversarial training (Bekoulis et al., 2018), Cross-Modal Attention Networks (Zhao et al., 2020) and BERT models (Eberts and Ulges, 2019). However, unlike the previous work done with this corpus, our goal is not to predict relations between the annotated entities, but to learn the context words used for expressing causal relations, so they can be recognized regardless of the entities between which they occur.

3 Data

TL usually involves the use of two types of datasets: source datasets and target datasets, where models are trained on the source datasets, and adapted to make predictions on the target datasets. We are specifically interested in extracting relations between food and disease entities, and we use the help of two existing source datasets, the CrowdTruth (Dumitrache et al., 2017) and the ADE dataset (Gurulingappa et al., 2012), in order to extract relations in the target FoodDisease dataset, which was created for the purposes of this study.

3.1 The CrowdTruth dataset

The CrowdTruth dataset (Dumitrache et al., 2017) for medical RE contains annotated data for *cause* and *treat* relations in sentences from abstracts of PubMed articles.

The dataset contains 4028 sentences annotated for the existence of a *cause* relation and 3983 sentences annotated for the existence of a *treat* relation. Every sample of the dataset contains the name of a relation, and a sentence containing two entities between which the relation may or may not occur. Each entity is further described by its UMLS name, its starting and ending position in the sentence, and the exact textual form in which it appears in the sentence. Apart from this, each sample is assigned several labels which indicate whether the relation is observed between the two terms.

The initial data (Wang and Fan, 2014) were collected using Distant supervision (DS) (Mintz et al., 2009), which is a inexpensive and straightforward way of labeling training data, but is also prone to producing noisy, low quality labels (Dumitrache et al., 2015b; Ji et al., 2017; Chen et al., 2021). Because of that, the annotations for the *cause* and *treat* relations collected using DS were further refined using the CrowdFlower platform where a multi-label annotation task was executed through crowdsourcing (Dumitrache et al., 2017, 2015b,a). Additionally, experts annotated sentences with binary labels, based on whether a specified seed relation discovered by DS is present between two given biomedical entities that occur in the sentence.

The *sentence relation* score given for each sample is computed as the cosine similarity between the vector containing the sum of the annotations of the non-expert workers, and the unit vector for the relation. Here, the unit vector refers to a one-hot vector where the value corresponding to the relation is equal to 1, and all other components are equal to 0. This score is in the range [0, 1]. The *crowd* score is calculated using the *sentence relation* score, by applying a threshold of 0.5 to separate positive from negative examples, and rescaling the obtained positive and negative samples in the ranges [0.5, 1], and [-1, -0.5], respectively.

The *expert* label is based on the experts' annotations and it takes values of either 1 or -1, indicating the presence or absence of the relation, respectively. However, due to the cost, limited time and availability of experts, the expert annotations were limited to 975 samples in the *cause* dataset and 621 samples in the *treat* dataset.

3.1.1 Target variable construction in the CrowdTruth dataset

The target variable is a binary indicator of the existence of the *cause* or *treat* relationship in the respective dataset. As the CrowdTruth dataset contains multiple indicators of these relations, we choose to rely on the labels assigned by experts, but since these are not available for all samples, we also use the *crowd* score, which has been shown to give reliable results in the original studies (Dumitrache et al., 2017, 2015b,a). To be more precise, if the sentence has been labeled by an expert, the target label is assigned a value of 1, if the score given by the expert is 1, or 0, if the score given by the expert is -1. If the sentence has not been labeled by an expert, the target label is assigned a value of 1, if the *crowd* score is positive, or 0, if the *crowd* score is negative.

3.2 The Adverse Drug Events (ADE) dataset

The ADE dataset (Gurulingappa et al., 2012) contains 6821 sentences expressing truthful relations between drugs and adverse effects they have been shown to cause, and 279 sentences with relations between drugs and dosages. Each sample consists of a sentence, the name of a drug, the name of a condition (if the relation expressed is *adverse effect*) or a dosage term (if the relation expressed is *dose*), and their starting and ending position in the sentence. The sentences were extracted from MEDLINE case reports, and were manually annotated by three annotators. There are 1319 unique drugs, 3341 unique conditions, and 130 unique dosage terms. In order to be consistent with the nomenclature in the other datasets, we refer to the *adverse effect* relation in the ADE dataset as a *cause* relation, and to the *condition* entities as *diseases*. The intuition behind using relations annotated as *adverse effect* to detect *cause* relations between food and disease entities is that one would use similar sentence structures to describe a disease occurring as a result of the ingestion of a particular drug or food.

3.3 The FoodDisease dataset

Since there was no data labeled for the existence of *cause* and *treat* relations between food and disease entities, for the purposes of this research we constructed a dataset containing 608 sentences from abstracts of PubMed articles. Fig. 1 depicts the steps taken in order to generate the dataset.

BuTTER (Cenikj et al., 2020) and SABER (Giorgi and Bader, 2019) were used for finding the food and disease entities in each abstract. Both are Named Entity Recognition (NER) methods based on Bidirectional Long Short-Term Memory and Conditional Random Fields. BuTTER extracts food entities from raw text, and is trained on the golden version of the FoodBase corpus (Popovski et al., 2019), which contains 1000 recipes annotated with food entities. In particular, we used the lexical lemmatized BuTTER model introduced in (Cenikj et al., 2020), which achieves a macro averaged F1 score of 0.946.

SABER is a biomedical NER tool, which provides several pre-trained NER models, from which we use the *DISO* model [2] to extract disease entities.

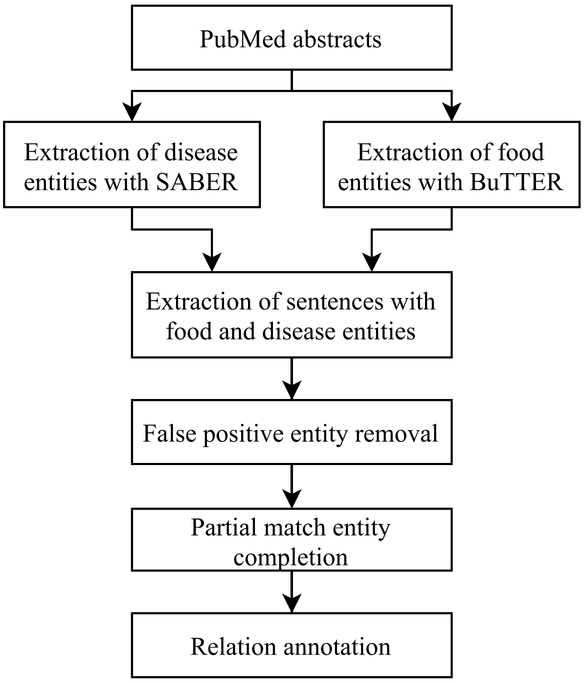

PubMed abstracts
Extraction of disease entities with SABER
Extraction of food entities with BuTTER
Extraction of sentences with food and disease entities
False positive entity removal
Partial match entity completion
Relation annotation

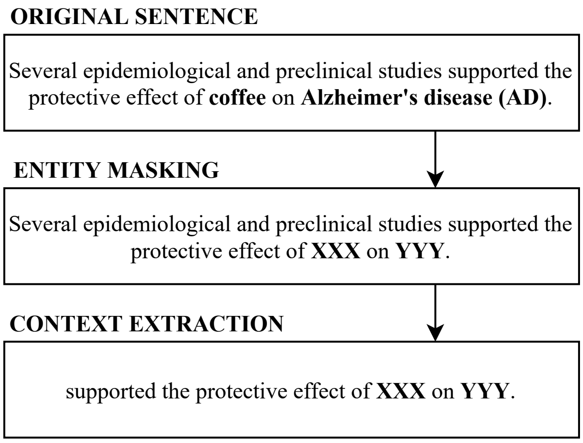
Figure 1: Steps taken to generate the FoodDisease dataset

The abstracts were filtered so that only sentences which contain at least one food and one disease entity were kept. The entities in each sentence were then manually checked and corrected in order to remove false positives and complete partial matches. Removing the false positive entities means that the tokens that were incorrectly extracted as food or disease entities by the BuTTER and SABER methods were discarded. Completing partial matches entails adding the missing words in entities which should contain multiple words, but some of them were not captured by the automatic annotators. Each sample contains a single food and a single disease entity, even if multiple such entities are mentioned in the sentence. Finally, each sentence was assigned binary labels to indicate if the *cause* and *treat* relations are present.

4 Methodology

In this section, we describe the proposed RE method, starting with the preprocessing applied to accomplish the generalization of the models trained on the source datasets to the target dataset. We then introduce the pre-trained transformer models used for text representation, and their fine-tuning for the RE task.

[2] https://baderlab.github.io/saber/resources



Several epidemiological and preclinical studies supported the protective effect of **coffee** on **Alzheimer's disease (AD)**.

Several epidemiological and preclinical studies supported the protective effect of **XXX** on **YYY**.

supported the protective effect of **XXX** on **YYY**.

Figure 2: Application of the preprocessing steps on a sentence from the FoodDisease dataset

4.1 Data preprocessing

Since the datasets we are using are annotated with relations between different types of biomedical entities, and we would like the developed models to generalize, and be able to extract the same relations between different types of entities, we mask out the entity mentions in each sentence. The reasoning behind this is that the model would not learn to detect relations between the concrete entities, but instead, use the surrounding words to determine whether they express the particular relation.

Since there could be several relations present in one sentence, we use a context window of length 5, i.e. use the words whose positions in the sentence are in the range (i-5,j+5), where i is the word index of the first occurring entity in the sentence, and j is the word index of the second occurring entity in the sentence.

Fig. 2 shows an example of the preprocessing steps being applied on a sentence from the FoodDisease dataset. The bolded words in the original sentence are the food and disease entities, which get masked out in the *Entity Masking* step, where they are replaced by *XXX* and *YYY*, respectively. These masking tokens are chosen arbitrarily, since their only use is for the model to distinguish between the subject and object entity. In the *Context Extraction* step, the final preprocessed version of the sentence is generated by keeping only the words in between the entities, and the 5 words that precede the first entity, *coffee*. Had there been additional words after the second entity, *Alzheimer's disease (AD)*, the first 5 of them would also be included in the context.

4.2 Text representation

In order to represent the textual data in numerical format, we use 3 pre-trained transformed-based models: BERT, RoBERTa and BioBERT.

4.2.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) is a bidirectional, contextual representation model that achieves state-of-the-art results in several natural language processing tasks. Following the principles of transductive TL, BERT is pre-trained on an unsupervised Mask Language Modeling (MLM) or Next Sentence Prediction (NSP) task, and then fine-tuned on another downstream task, such as NER, Natural Language Inference or Question Answering. The pre-trained BERT models can be finetuned without substantial modifications in their architecture. In the simplest case, only the output layer needs to be replaced, depending on the task that the model is intended to perform. We use the original BERT model, which is pre-trained on the BooksCorpus (Zhu et al., 2015) and English Wikipedia, and fine tune it for relation classification.

4.2.2 RoBERTa

RoBERTa (Robustly Optimized BERT Approach) (Liu et al., 2019) is a text representation model based on the original BERT architecture, with a number of improvements introduced in the pre-training phase, some of which include training on a larger amount of data, longer training, removal of the NSP task, and introduction of dynamic masking. Apart from the BooksCorpus and Wikipedia, which are used for the pretraining of BERT, RoBERTa is trained on data from 3 additional sources: the CommonCrawl News dataset (Nagel, 2016), the OpenWebText corpus (Gokaslan and Cohen, 2019) and Stories subset from the Common Crawl dataset (Trinh and Le, 2018).

4.2.3 BioBERT

BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) (Lee et al., 2019) is a domain-specific version of the BERT model. Due to the fact that biomedical texts contain a considerable amount of domain-specific proper nouns and terms that do not appear in more general texts and would hence be unfamiliar to the original BERT, the data on which

BioBERT is trained is supplemented by PubMed abstracts and full-text articles from PubMed Central. As a result, BioBERT has been shown to outperform BERT in biomedical NER, RE, and QA (Lee et al., 2019).

4.3 Models

We perform end-to-end fine-tuning of the pre-trained BERT, RoBERTa and BioBERT models for the RE task. In order to adapt the original architecture to perform binary classification, the last layer of the models is replaced with a dropout and a linear layer which performs binary classification. During fine-tuning, the model parameters are initialized with the values from the pre-training step, and are fine-tuned using the labeled data from the source datasets. The input of a BERT model can unambiguously represent both a single sequence and a pair of text sequences (for example, a question and an answer) in one token sequence, by using a separator token [SEP] to mark the end of each sequence. We explore both types of inputs and construct two different models:

- Single Sequence Classifier (SSC) - The model takes a single sequence as an input and performs simple binary classification.
- Sequence Pair Classifier (SPC) - The model takes as input two sequences. The first sequence is the sequence that we want to classify (the one that is used on its own in the SSC), while the second sequence is a concatenation of 10 randomly sampled sequences which have positive labels for the relation we are aiming to detect. We refer to the first sequence as the *sequence of interest*, while we call the concatenation of 10 sequences a *ground truth* for the relation in question. The sentences used in the ground truth sequences are not used as sequences of interest.

  The intuition behind this approach is that we can reformulate the task *Does sequence X express relation Y?* as *Is sequence X similar to other sequences that contain relation Y?*. The task is still a binary classification, and the label remains the same as for the SSC.

  We construct 10 ground truth sequences for each relation, and pair each sequence of interest with each ground truth. The same generated ground truths are used at training and prediction time. For each sequence of interest in the test set, we generate 10 predictions (one for each ground truth) and assign the average of the predicted probabilities as the probability of the sequence of interest belonging to the positive class.

Table 1: Examples of inputs given to the SSC and SPC models when identifying the *treat* relation

Inputs given to the SSC model

Input	Label
supported the protective effect of XXX on YYY	1
XXX is known to cause YYY	0

Inputs given to the SPC model

Input	Label
Sequence of interest: supported the protective effect of XXX on YYY *Ground truth*: XXX has been used in the treatment of YYY; XXX is known to cure YYY; XXX is associated with a reduced incidence of YYY	1
Sequence of interest: XXX is known to cause YYY *Ground truth*: XXX has been used in the treatment of YYY; XXX is known to cure YYY; XXX is associated with a reduced incidence of YYY	0

Table 1 features examples of the inputs given to the SSC and SPC models that identify the *treat* relation. The first input sample expresses a *treat* relation, so the label is one, while the second input sample expresses a *cause* relation, so the label is zero. The inputs of the SSC model are the same as for the *sequences of interest* of the SPC model. For the sake of simplicity, for the SPC model in the examples, we demonstrate one ground truth, which is a concatenation of 3 sequences that represent a *treat* relation. In our experiments, we use 10 such *ground truths*, each being a concatenation of 10 sequences.

During the fine-tuning, the AdamW optimizer is used with a learning rate of $4 * 10^{-5}$. An early stopping strategy is applied to prevent overfitting. The models are trained for a maximum of 10 epochs, or until the improvement in validation loss of 2 consecutive epochs does not surpass $5 * 10^{-3}$.

The source codes for fine-tuning the SSC mod-

Table 2: Number of samples from the positive and negative class in each dataset

Dataset	CrowdTruth		ADE	FoodDisease	
Relation	Cause	Treat	Cause	Cause	Treat
Class					
Positive	1429	1406	6821	142	323
Negative	2555	2578	1685	466	285

els[3] and the SPC[4] models are publicly available on the Colab platform.

5 Evaluation

5.1 Evaluation on the source datasets

When applying TL, a model trained on a source dataset can experience some degradation in performance when evaluated on the target dataset. In order to get an idea about the upper bound of the performance expected on the target dataset, the models' performance is first evaluated on the same, source datasets they were trained on using 10-fold cross validation.

All 3 of the datasets are unbalanced, and the class distribution of each dataset is given in Table 2. For the ADE dataset, we only train classifiers for the detection of the *cause* relation, since that dataset does not contain annotations for the *treat* relations. We consider the sentences annotated with the *dose* relation in the ADE dataset to be negative samples for the *cause* relation. However, since there are only 279 such sentences, in order to avoid extreme class unbalance, we supplement the negative samples in the train portion of the ADE dataset by adding the samples that are annotated as positive for the *treat* relation in the CrowdTruth dataset. 10% of the training portion of each fold is removed and used for validation, preserving the ratio of the positive and negative samples.

Table 3: Macro averaged F1 scores obtained from the evaluation on the source datasets when the proposed preprocessing is applied, averaged from 10 folds

Dataset	CrowdTruth		ADE	FoodDisease	
Relation	Cause	Treat	Cause	Cause	Treat
Model: SSC					
BERT	0.753	0.880	0.871	0.744	0.886
RoBERTa	0.740	0.879	0.866	0.711	0.884
BioBERT	0.750	0.890	0.894	0.847	0.871
Model: SPC					
BERT	0.745	0.873	0.822	0.478	0.835
RoBERTa	0.752	0.880	0.743	0.433	0.835
BioBERT	0.771	0.884	0.873	0.545	0.900

Because of the unbalanced class distribution in all three datasets, we evaluate the models in terms of the macro averaged f1 scores, averaged from all folds, and these are depicted in Table 3. The models are both trained and evaluated on the datasets indicated in the table header. The SSC and SPC models combined with 3 different pretrained BERT models (BERT, RoBERTa and BioBERT) result in 6 models, which are evaluated on the 3 datasets. The first group of three rows of scores refers to the performance of the SSC model, while the second group refers to the SPC model. The underlined values refer to the highest f1 macro score in each column, and we can note that the BioBERT models give the best performance. The SSC models generally outperform the SPC models.

The performance of the SPC models which detect the *cause* relation in the FoodDisease dataset is notably lower than the rest of the models. Looking into the models' raw predictions, it is obvious that the models predict the negative class too often, which results in high recall for the negative class, but very low recall for the positive class. This can be attributed to the fact that from the 114 positive samples in the training portion of each fold, 100 are used for constructing the ground truth sequences used by the SPC models, leaving only 14 positive samples for training. Annotating more data, decreasing the number of ground truth sequences or the number of sentences in each ground truth sequence, and balancing the data are possible strategies which are expected to remedy this anomaly.

5.2 Transfer learning evaluation

In this subsection, we report the performance reached by the models trained on the CrowdTruth and ADE source datasets, when evaluated on the target FoodDisease dataset. In this case, the models are trained on balanced data, since the class distribution in the source datasets does not reflect the distribution in the target dataset, and are evaluated on the whole FoodDisease dataset.

[3] https://colab.research.google.com/drive/1UOFuk6-_6c-za6P54SiIhdY9ZH6T5Xpo?usp=sharing

[4] https://colab.research.google.com/drive/1HA78g3YG90UuYT9SZPxw5brE6LqPKC2s?usp=sharing

Table 4: Macro averaged F1 scores obtained from the evaluation on the target FoodDisease dataset, when the proposed preprocessing is applied

Dataset	CrowdTruth		ADE
Relation	Cause	Treat	Cause
Model: SSC			
BERT	0.727	0.841	0.750
RoBERTa	0.805	0.883	0.710
BioBERT	0.805	0.878	0.750
Model: SPC			
BERT	0.585	0.689	0.619
RoBERTa	0.701	0.838	0.648
BioBERT	0.636	0.872	0.639

Table 5: Macro averaged F1 scores obtained from the evaluation on the target FoodDisease dataset, when the entire sentence is being used as input

Dataset	CrowdTruth		ADE
Relation	Cause	Treat	Cause
Model: SSC			
BERT	0.595	0.828	0.568
RoBERTa	0.659	0.759	0.228
BioBERT	0.610	0.900	0.633
Model: SPC			
BERT	0.557	0.837	0.608
RoBERTa	0.594	0.844	0.587
BioBERT	0.657	0.881	0.625

Table 4 features the macro averaged F1 scores that the models achieve when the preprocessing introduced in subsection 4.1 is applied on the input.

When comparing the results in Table 3 and 4, we can observe that the SPC models and the models trained on the ADE dataset experience performance deterioration when they are evaluated on the target dataset, but the SSC models trained on the CrowdTruth dataset have a similar performance in both evaluations. This is expected to some extent, since the relations in the ADE dataset are originally annotated as *adverse effect*, which we loosely interpret as a *cause* relation, while the sentences in the CrowdTruth dataset are annotated for precisely *cause* and *treat* relations.

Additionally, we conduct experiments to evaluate the proposed preprocessing technique, which we compare to the scenario when no preprocessing is applied (neither the *Entity Masking* nor the *Context Extraction* step) and the entire sentences are given to the model. The macro averaged F1 scores obtained in such a setting are featured in Table 5. The best results are achieved by the RoBERTa and BioBERT models. Most of the models benefit from the preprocessing, which is especially noticable in the SSC models that identify the *cause* relation, where the proposed preprocessing leads to an improvement of the averaged macro f1 scores of at least 0.100. Looking into the metrics for the positive and negative class separately reveals that the lower performance of the models which do not use the proposed processing is due to their lower precision in identifying the positive class.

Interestingly, the SPC models that identify the *treat* relation seem to perform better without the preprocessing, even though only one the performance of the BERT model differs by a large margin, while the performances of the BioBERT and RoBERTa models differ by less than 0.010.

It is important to note that the evaluation on these models on the FoodDisease dataset may be somewhat flawed, since it may hide the possible disadvantage of using entire sentences as input, because all of the sentences in the FoodDisease dataset are unique. This would mean that if a sentence contains both relations, as for example *Nuts are known to reduce the risk of heart disease, but can also cause allergies*, the dataset would either contain the *(food, relation, disease)* triple *(nuts, treat, heart disease)* or the triple *(nuts, cause, allergies)*, but not both. The models that do not use the proposed preprocessing and get the entire sentence as input, would in this case produce an identical output for both triples, but when evaluated on the FoodDisease dataset, they would not be penalized for doing so.

Overall, the best models trained on the source datasets achieve a macro F1 scores of 0.805 and 0.900, for the detection of *cause* and *treat* relations, respectively, between food and disease entities in the target dataset. In comparison, the performance of the best models trained on the target FoodDisease dataset (the SSC-BioBERT and SPC-BioBERT in Table 3) is 0.847 and 0.900. This indicates that the application of TL using pretrained transformer models enables us to train models using a small amount of annotated data, but we can also obtain satisfactory results with no annotated data for the specific RE task, by repurposing annotations for the same relations between different entities.

6 Conclusion

In this paper, we propose Relation Extraction (RE) models for the detection of *cause* and *treat* relations between food and disease entities from raw text. To make up for the absence of annotated data for this task, we explore the feasibility of Transfer Learning (TL) by using the transformer models BERT, RoBERTa, and BioBERT, which are pretrained on large amounts of data, and fine-tuned for performing RE between various types of biomedical entities. The models are trained to recognize relations based on the context words used to express each relation, rather than the entities themselves, so they can successfully generalize to the task of recognizing the relations between food and disease entities, and likely, other types of entities, though this is not evaluated in the scope of this paper.

In order to evaluate the proposed approach, we introduce the FoodDisease dataset, which consists of 608 sentences annotated for the existence of the *cause* and *treat* relations between food and disease entities in sentences of PubMed abstracts. The dataset is released as an open-source resource, and is, to the best of our knowledge, the first annotated English RE dataset of such kind in the food domain.

The best models that are fine-tuned on this dataset achieve macro averaged F1 scores of 0.847 and 0.900 for the *cause* and *treat* relations, respectively. The best models which are fine-tuned using the data where the entities are not food-disease pairs, but other biomedical entities of various types, achieve macro averaged F1 score of 0.805 for the *cause* relation and 0.900 for the *treat* relation. This indicates that in the event where no experts are available to annotate data, the proposed method enables the repurposing of existing RE datasets for the training of models that can recognize the relation that the dataset is annotated for, between different types of entities.

The developed models will be used as part of an information extraction pipeline which will structure the findings of experts in biomedical scientific literature, with the aim of alleviating the process of linking knowledge graphs from the domain of biomedicine to the domain of food and nutrition.

Acknowledgements

This work was supported by the Ad Futura grant for postgraduate study; the Slovenian Research Agency (research core funding programmes P2-0098); the European Union's Horizon 2020 research and innovation programme under grant agreement 863059 (FNS-Cloud, Food Nutrition Security) and under grant agreement 101005259 (COMFOCUS).

References

Aida Bchir and Wahiba Ben Abdessalem Karaa. 2013. Extraction of drug-disease relations from medline abstracts. In *2013 World Congress on Computer and Information Technology (WCCIT)*, pages 1–3.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Adversarial training for multi-context joint entity and relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2830–2836, Brussels, Belgium. Association for Computational Linguistics.

Gjorgjina Cenikj, Gorjan Popovski, Riste Stojanov, Barbara Koroušić Seljak, and Tome Eftimov. 2020. BuTTER: BidirecTional LSTM for Food Named-Entity Recognition. In *Proc. Big Food and Nutrition Data Management and Analysis at IEEE BigData 2020*, pages 3550–3556.

Tiantian Chen, Nianbin Wang, Hongbin Wang, and Haomin Zhan. 2021. Distant supervision for relation extraction with sentence selection and interaction representation. *Wireless Communications and Mobile Computing*, 2021:1–16.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Ika Novita Dewi, Shoubin Dong, and Jinlong Hu. 2017. Drug-drug interaction relation extraction with deep convolutional neural networks. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1795–1802.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2015a. Achieving expert-level annotation quality with crowdtruth: The case of medical relation extraction. In *BDM2I@ISWC*.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2015b. Crowdtruth measures for language ambiguity: The case of medical relation extraction. In *LD4IE@ISWC*.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2017. Crowdsourcing ground truth for medical relation extraction. *CoRR*, abs/1701.02185.

Markus Eberts and Adrian Ulges. 2019. Span-based joint entity and relation extraction with transformer pre-training. *CoRR*, abs/1909.07755.

Ziling Fan, Luca Soldaini, Arman Cohan, and Nazli Goharian. 2018. Relation extraction for protein-protein interactions affected by mutations. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '18, page 506–507, New York, NY, USA. Association for Computing Machinery.

Sumam Francis, Jordy Van Landeghem, and Marie-Francine Moens. 2019. Transfer learning for named entity recognition in financial and biomedical documents. *Information*, 10(8).

John Giorgi and Gary Bader. 2019. Towards reliable named entity recognition in the biomedical domain. *bioRxiv*.

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus. `http://Skylion007.github.io/OpenWebTextCorpus`.

Harsha Gurulingappa, Abdul Mateen-Rajput, and Luca Toldo. 2012. Extraction of potential adverse drug events from medical case reports. *Journal of biomedical semantics*, 3(1):15–15.

Walid Hafiane, Joel Legrand, Yannick Toussaint, and Adrien Coulet. 2020. Experiments on transfer learning architectures for biomedical relation extraction. ArXiv:2011.12380.

Kasper Jensen, Gianni Panagiotou, and Irene Kouskoumvekaki. 2014. NutriChem: a systems chemical biology resource to explore the medicinal value of plant-based foods. *Nucleic Acids Research*, 43(D1):D940–D945.

Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*.

Sun Kim, Haibin Liu, Lana Yeganova, and W. John Wilbur. 2015. Extracting drug–drug interactions from literature using a rich feature-based linear kernel approach. *Journal of Biomedical Informatics*, 55:23–30.

Shun Koyabu, Thi Thanh Thuy Phan, and Takenao Ohkawa. 2015. Extraction of protein-protein interaction from scientific articles by predicting dominant keywords. *BioMed Research International*, 2015:928531.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Joël Legrand, Yannick Toussaint, Chedy Raïssi, and Adrien Coulet. 2018. Syntax-based transfer learning for the task of biomedical relation extraction. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 149–159, Brussels, Belgium. Association for Computational Linguistics.

Sangrak Lim and Jaewoo Kang. 2018. Chemical–gene relation extraction using recursive neural network. *Database*, 2018. Bay060.

Shengyu Liu, Buzhou Tang, Qingcai Chen, and Xiaolong Wang. 2016. Drug-drug interaction extraction via convolutional neural networks. *Computational and Mathematical Methods in Medicine*, 2016:6918381.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Pei-Yau Lung, Zhe He, Tingting Zhao, Disa Yu, and Jinfeng Zhang. 2019. Extracting chemical–protein interactions from literature using sentence structure analysis and feature engineering. *Database*, 2019. Bay138.

Qingliang Miao, Shu Zhang, Bo Zhang, and Hao Yu. 2012. Extracting and visualizing semantic relationships from Chinese biomedical text. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 99–107, Bali, Indonesia. Faculty of Computer Science, Universitas Indonesia.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. 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*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Sebastian Nagel. 2016. Cc news. `http://commoncrawl.org/2016/10/news-dataset-available/`. Accessed: 2021-03-10.

Yueqiong Ni, Kasper Jensen, Eirini Kouskoumvekaki, and Gianni Panagiotou. 2017. Nutrichem 2.0: exploring the effect of plant-based foods on human health and drug efficacy. *Database: The Journal of Biological Databases and Curation*, 2017(1).

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and elmo on ten benchmarking datasets. *CoRR*, abs/1906.05474.

Gorjan Popovski, Barbara Koroušić Seljak, and Tome Eftimov. 2019. FoodBase corpus: a new resource of annotated food entities. *Database*, 2019. Baz121.

Melanie Reiplinger, Michael Wiegand, and Dietrich Klakow. 2014. Relation extraction for the food domain without labeled training data – is distant supervision the best solution? In *Advances in Natural Language Processing*, pages 345–357, Cham. Springer International Publishing.

Sunil Kumar Sahu and Ashish Anand. 2018. Drug-drug interaction extraction from biomedical texts using long short-term memory network. *Journal of Biomedical Informatics*, 86:15–24.

Cong Sun and Zhihao Yang. 2019. Transfer learning in biomedical named entity recognition: An evaluation of BERT in the PharmaCoNER task. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 100–104, Hong Kong, China. Association for Computational Linguistics.

Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning. *CoRR*, abs/1806.02847.

Chang Wang and James Fan. 2014. Medical relation extraction with manifold models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 828–838, Baltimore, Maryland. Association for Computational Linguistics.

Pengwei Wang, Tianyong Hao, Jun Yan, and Lianwen Jin. 2017. Large-scale extraction of drug-disease pairs from the medical literature. *J. Assoc. Inf. Sci. Technol.*, 68(11):2649–2661.

Karl Weiss, Taghi Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big Data*, 3(1):9.

Michael Wiegand, Benjamin Roth, and Dietrich Klakow. 2012a. Data-driven knowledge extraction for the food domain. In *Proceedings of KONVENS 2012*, pages 21–29. ÖGAI.

Michael Wiegand, Benjamin Roth, Eva Lasarcyk, Stephanie Köser, and Dietrich Klakow. 2012b. A gold standard for relation extraction in the food domain. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 507–514, Istanbul, Turkey. European Language Resources Association (ELRA).

Ningyu Zhang, Shumin Deng, Zhanlin Sun, Jiaoyan Chen, Wei Zhang, and Huajun Chen. 2019. Transfer learning for relation extraction via relation-gated adversarial learning. *CoRR*, abs/1908.08507.

Shan Zhao, Minghao Hu, Zhiping Cai, and Fang Liu. 2020. Modeling dense cross-modal interactions for joint entity-relation extraction. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4032–4038. International Joint Conferences on Artificial Intelligence Organization.

Huiwei Zhou, Zhuang Liu, Shixian Ning, Yunlong Yang, Chengkun Lang, Yingyu Lin, and Kun Ma. 2018. Leveraging prior knowledge for protein–protein interaction extraction with memory network. *Database*, 2018. Bay071.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2019. A comprehensive survey on transfer learning. *CoRR*, abs/1911.02685.