sustainability

Article

Extracting Semantic Relationships in Greek Literary Texts

Despina Christou * and **Grigorios Tsoumakas**

School of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece; greg@csd.auth.gr
* Correspondence: christoud@csd.auth.gr

Abstract: In the era of Big Data, the digitization of texts and the advancements in Artificial Intelligence (AI) and Natural Language Processing (NLP) are enabling the automatic analysis of literary works, allowing us to delve into the structure of artifacts and to compare, explore, manage and preserve the richness of our written heritage. This paper proposes a deep-learning-based approach to discovering semantic relationships in literary texts (19th century Greek Literature) facilitating the analysis, organization and management of collections through the automation of metadata extraction. Moreover, we provide a new annotated dataset used to train our model. Our proposed model, REDSandT_Lit, recognizes six distinct relationships, extracting the richest set of relations up to now from literary texts. It efficiently captures the semantic characteristics of the investigating time-period by finetuning the state-of-the-art transformer-based Language Model (LM) for Modern Greek in our corpora. Extensive experiments and comparisons with existing models on our dataset reveal that REDSandT_Lit has superior performance (90% accuracy), manages to capture infrequent relations (100%F in long-tail relations) and can also correct mislabelled sentences. Our results suggest that our approach efficiently handles the peculiarities of literary texts, and it is a promising tool for managing and preserving cultural information in various settings.

Keywords: relation extraction; distant supervision; deep neural networks; Transformers; Greek NLP; literary fiction; heritage management; metadata extraction; Katharevousa

Citation: Christou, D.; Tsoumakas, G. Extracting Semantic Relationships in Greek Literary Texts. *Sustainability* **2021**, *13*, 9391. https://doi.org/10.3390/su13169391

Academic Editor: Charalampos Dimoulas and Amir Mosavi

Received: 30 June 2021
Accepted: 16 August 2021
Published: 21 August 2021

1. Introduction

An important part of humanity's cultural heritage resides in its literature [1], a rich body of interconnected works revealing the history and workings of human civilization across the eras. Major novelists have produced their works by engaging with the spirit of their time [2] and capturing the essence of society, human thought and accomplishment.

Cultural Heritage (CH) in its entirety constitutes a "cultural capital" for contemporary societies because it contributes to the constant valorization of cultures and identities. Moreover, it is also an important tool for the transmission of expertise, skills and knowledge across generations and is closely related to the promotion of cultural diversity, creativity and innovation [3]. For this reason, proper management of the development potential of CH requires a sustainability-oriented approach, i.e., one that ensures both the preservation of the heritage from loss and its connection to the present and the future. Proper management of literary cultural heritage, therefore, requires extensive digitization of collections and procedures that allow for the automatic extraction of semantic information and metadata to ensure the organization of past collections and their linkage with present and future documents.

Until recently, engaging with a large body of literature and discovering insights and links between storytellers and cultures was a painstaking process which relied mainly on close reading [4]. Nowadays, however, the large-scale digitization of texts as well as developments in Artificial Intelligence (AI) and Natural Language Processing (NLP) are making it possible to explore the richness of our written heritage with methods that were not possible before at an unprecedented scale, while facilitating the management and preservation of texts [5].

One of the opportunities afforded by digitization is *relation extraction* (RE): the automatic discovery of relations between entities in a document. This task plays a central role in NLP because relations can be used to populate knowledge bases (KB), to index corpora in search engines, to answer questions related to the text, to assist in the comparative analysis of texts and to understand/analyze the narration of a story. In this paper, we present a novel deep-learning model for RE that enables applications in all of the above domains by automatically identifying relations between entities in 19th century Greek literary texts. Although there are several RE approaches in the literature, the particular texts we are interested in (fiction), the language and the specific period all present significant challenges. We will have more to say about these shortly.

Most RE methods follow a supervised approach; thus, the required large amount of labeled training data constitutes perhaps the greatest barrier for real-world applications. In order to overcome this challenge, RE research has adopted distantly supervised approaches that are based upon automatically constructed datasets. Towards that end, Reference [6] proposed to use distant supervision (DS) from a KB, assuming that if two entities in a KB exhibit a relation, then all sentences mentioning those entities express that relation. This assumption inevitably results in false positives and to remotely generated records containing incorrect labels. In order to mitigate the problem of *wrong labeling*, Reference [7] relaxed the assumption so that it does not apply to all instances and, together with [8,9], proposed multi-instance learning. In that setting, classification shifts from instance-level to bag-level, with current state-of-the-art RE methods focusing on reducing the effect of noisy instances.

At the same time, extracting relations from literary texts has been undertaken only in the broader context of people in dialogue [10–13], people in the same place [14] and event extraction [14,15] and not, thus far, in the context of predefined relations among named entities other than person and place. We also emphasize the fact that state-of-the-art RE approaches are evaluated mostly on news corpora. The reason is that literary texts put emphasis on the narrative craft and exhibit characteristics that go beyond journalistic, academic, technical or more structured forms of literature. Moreover, literary texts are characterized by creative writing peculiarities that can vary significantly from author to author and time to time. Moreover, as most works of literature have been digitized through OCR systems, the digitized versions can also suffer from character or word misspellings. All these make it extremely challenging to discover entity relations in literary texts.

In order to address these challenges, we propose REDSandT_Lit (Relation Extraction with Distant Supervision and Transformers for Literature), a novel distantly supervised transformer-based RE model that can efficiently identify six distinct relationships from Greek literary texts of the 19th century, the period that "contains" the largest part of digitized Modern Greek literature. Since no related dataset exists, we undertook the construction of a new dataset including 3649 samples annotated through distant supervision with seven semantic relationships, including 'NoRel' for instances with non-labelled relation. Our dataset is in the *Katharevousa* variant of Greek, an older, more formal and more complex form of the Modern Greek language in which a great part of Modern Greek literature is written in. In order to capture the semantic and syntactic characteristics of the language, we exploited the state-of-the-art transformer-based Language Model (LM) for Modern Greek (GREEK-BERT [16]), which we fine-tuned on our specific task and language. In order to handle the problem of noisy instances as well as the long sentences which are typical in literary writing, we guided REDSandT_Lit to focus solely on a compressed form of the sentence that includes only the surrounding text of the entity pair together with their entity types. Finally, our model encodes sentences by concatenating the entity-pair type embeddings, with relation extraction to occur at bag-level as a weighted sum over the bag's sentences predictions. Regarding the selected transformer-based model, the reasons for choosing BERT [17] are twofold: (i) BERT is the only transformer-based model pre-trained in Modern Greek corpora [16], and (ii) BERT considers bidirectionality while training with [18], showing BERT to capture a wider set of relations compared to

GPT [19] under a DS setting. Extensive experimentation and comparison of our model to several existing models for RE reveals REDSandT_Lit's superiority. Our model captures with great precision (75–100% P) all relations, including the infrequent ones that other models failed to capture. Moreover, we will observe that fine-tuning a transformer-based model under a DS setting and incorporating entity-type side information highly boosts RE performance, especially for the relations in the long-tail of the distribution. Finally, REDSandT_Lit manages to find additional relations that were missed during annotation.

Our proposed model is the first to extract semantic relationships from 19th century Greek literary texts, and the first, to our knowledge, to extract relationships between entities other than person and place; thus, we provide a broader and more diverse set of semantic information on literary texts. More precisely, we expand the boundaries of current research from narration understanding to extended metadata extraction. Even though online repositories provide several metadata that accompany digitized books to facilitate search and indexing, digitized literary texts contain rich semantic and cultural information that often goes unused. The six relationships identified by our model can further boost the books' metadata, preserve more information and facilitate search and comparisons. Moreover, having access to a broader set of relations can boost downstream tasks, such as recommending similar books based on hidden relations. Finally, distant reading [4] goes one step further with readers and storytellers in terms of understanding the story set more quickly and easily.

The remainder of this paper is organized as follows: Section 2 contains a brief literature review, Section 3 discusses our dataset and proposed methodology. Sections 4 and 5 contain our results and discussion, respectively.

2. Related Work

Our work is related to distantly supervised relation extraction, information extraction from literary texts and metadata enhancement.

2.1. Distantly-Supervised Relation Extraction

Distant supervision [20,21] plays a key role in RE meeting its need for a plethora of training data in a simple and cost-effective manner. Mintz et al. [6] were the first to propose DS to automatically construct corpora for RE, assuming that all sentences that include an entity pair that has a relation in a KB express the same relation. Of course, this assumption is very loose and is accompanied by noisy labels. Multi-instance learning methods were proposed to alleviate the problem by performing relationship classification at the bag level, where a bag contains instances that mention the same entity pair [7,8].

With the training framework being typically the aforementioned, research focused on features and models that better suppress noise. Until the advent of neural networks (NNs), researchers used simple models heavily relying on handcrafted features (part-of-speech tags, named entity tags, morphological features, etc.) [7,8]. Later on the focus turned to model architecture. Initially, a method based on a convolutional neural network (CNN) was proposed by [22] to automatically capture the semantics of sentences, while piecewise-CNN (PCNN) [23] became the common architecture for embedding sentences and handling DS noise [24–29]. Moreover, Graph-CNNs (GCNN) proved an effective method for encoding syntactic information from text [30].

The development of pre-trained language models (LMs) that rely on transformer architecture [31] and enable to transfer common knowledge in downstream tasks has been shown to capture semantic and syntactic features better [32]. In particular, it has been shown that pre-trained LMs significantly improve the performance in text classification tasks, prevent overfitting and increasing sample efficiency [33]. Moreover, methods in [34,35] that fine-tune the pre-trained LM models, as also observed in [18,19] who extended GPT [32] and BERT [17] models, respectively, to the DS setting by incorporating a multi-instance training mechanism, show that pre-trained LMs provide a stronger signal for DS than specific linguistic and side-information features [30].

2.2. Information Extraction from Literary Texts

While relation extraction has been extensively studied in news and biomedical corpora, extracting semantic relationships from literary texts is a much less studied area. Existing research attempts to understand narration mostly from the viewpoint of character relationships but not to augment existing KBs or enhance a story's metadata in an online repository. An explanation based on [10] is the difficulty in automatically determining meaningful interpretations (i.e., predefined relations) and the lack of semantically annotated corpora. Therefore, most research is focused on extracting a limited set of relationships among characters, such as "interaction" [10–12], "mention" [10] and "family" [13].

The key challenges in extracting relations from literary texts are listed out in [36], an excellent survey on extracting relations among fictional characters. The authors point out that there can be significant stylistic differences among authors and grammar misformats in books of different periods, while the closed-word fashion of fiction where plot involves recurring entities entails coreference resolution issues. This work aims at capturing relations not only between people or places but also between organizations, dates and work of art titles.

2.3. Metadata Enhancement

It was only two decades ago when book information was only available by accessing libraries. On the other hand, nowadays we suffer from information overload, with libraries now including their own databases to facilitate search [37].

With increasing digital content being added to the enormous collection of libraries, archives, etc., providing machine-readable structured information to facilitate information integration and presentation [38] is becoming increasingly important and challenging. Moreover, research has shown that providing metadata in fiction books highly affects the selection of a fiction book and their perception on the story [39,40]. For that reason, we believe that enhancing the metadata of literary texts is crucial.

3. Materials and Methods

As discussed in the Introduction, extracting cultural information from literary texts demands either a plethora of annotations or robust augmentation techniques that can capture a representative sample of annotations and boost machine learning techniques. Meanwhile, automatically augmented datasets are always accompanied by noise, while creative writing's characteristics set an extra challenge.

In this section, we present a new dataset for Greek literary fiction from the 19th century. The dataset was created by aligning entity pair-relation triplets to a representative sample of Greek 19th century books. Even though we efficiently manage to augment the training samples, these inevitably suffer from noise and include imbalanced labels. Moreover, the special nature of the 19th century Greek language sets an extra challenge.

We present our model as follows: a distantly supervised transformer-based RE method based on [18] that has proven to efficiently suppress noise from DS using multi-instance learning and exploiting a pre-trained transformer-based LM. Our model proposes a simpler configuration for representing the embedding of the final sentence, which manages to capture a larger number of relations by using information about the entity types and the Greek BERT's [16] pre-trained model.

3.1. Benchmark Dataset

Preserving semantic information from cultural artifacts requires either extensive annotation that is rarely available or automatically augmented datasets to sufficiently capture context. In the case of literary texts, no dataset exists to train our models. Taking into account that the greatest part of digitized Modern Greek literature refers to the 19th century, we construct our dataset by aligning relation-triples from [41] to twenty-six (26) literary Greek books of the 19th century (see Table A1). Namely, we use the provided relation triplets (i.e., head-tail-relationship triplets) as an external knowledge base (KB) to automatically extract sentences that include the entity pairs, assuming that these sentences also express the same relationship (distant supervision).

The dataset's six specific relations and their statistics can be found in Table 1. Train, validation and test datasets follow a 80%-10%-10% split. We assume that a relationship can occur within a period of three consequent sentences and only between two named entities. Sentences that include at least two named entities of different types but do not constitute a valid entity pair are annotated with a "NoRel" relation. These can either reflect sentences with no actual underlying relation or sentences for which the annotation is missed. The dataset also includes the named entity types of the sentence's entity pair. The following five entity types are utilized: person (PER), place (GPE), organization (ORG), date (DATE) and book title (TITLE). We made this dataset publicly available (Data available at: https://github.com/intelligence-csd-auth-gr/extracting-semantic-relationships-from-greek-literary-texts (accessed on 3 August 2021)) to encourage further research on 19th century Greek literary fiction.

Table 1. Dataset's Statistics.

Relations	Short Description	#Train Samples	#Val Samples	#Test Samples
NoRel	No Relation	1721	217	217
artAuthor	Book Author	737	94	92
pubDate	Book Publication Date	210	27	28
workAt	Working Relationship	141	18	17
orgPlace	Organization Place	52	7	7
orgDate	Organization Founding Date	27	3	3
artHero	Book Hero	26	3	3
TOTAL		2913	369	367

The challenges of this dataset are threefold. At first, similar to all datasets created via distant supervision, ours also suffers from noisy labels (false positives) and is imbalanced, including relations with a varying number of samples. Secondly, the dataset includes misspellings stemming from the books' digitization through OCR systems. Lastly, the documents use a conservative form of the modern Greek language, *katharevousa*, which was used between the late 18th century and 1976. Katharevousa, which covers a significant part of modern Greek literature, is more complex than modern Greek, including additional cases, compound words and other grammatical features that set an extra challenge for the algorithm.

3.2. The Proposed Model Architecture

In this section, we present our approach towards extracting semantic relationships from literary texts. We highlight that the specific challenges that we have to address are as follows: DS noise, imbalanced relations, character misspellings due to OCR, Katharevousa form of Greek language and creative writing peculiarities. Inspired by [18,19] who showed that DS and pre-trained models can suppress noise and capture a wider set of relations, we propose an approach that efficiently handles the aforementioned challenges by using multi-instance learning, exploiting a pre-trained transformer-based language model and incorporating entity type side-information.

In particular, given a bag of sentences $\{s_1, s_2, \dots, s_n\}$ that concern a specific entity pair, our model generates a probability distribution on the set of possible relations. The model utilizes the GREEK-BERT pre-trained LM [16] to capture the semantic and syntactic features of sentences by transferring pre-trained common-sense knowledge. In order to capture the specific patterns of our corpus, we fine-tuned the model using multi-instance learning; namely, we trained our model to extract the entity pairs' underlying relation given their associated sentences.

During fine-tuning, we employ a structured, RE-specific input representation to minimize architectural changes to the model [42]. Each sentence is transformed to a structured format, including a compressed form of the sentence along with the entity pair and their entity types. We transform the input into a sub-word level distributed representation using byte-pair encoding (BPE) and positional embeddings from GREEK-BERT fine-tuned on our corpus. Lastly, we concatenate the head and tail entities' types embeddings, as shaped from BERT's last layer, to form the final sentence representation that we used to classify the bag's relation.

The proposed model can be summarized in three components: the sentence encoder, the bag encoder and model training. Components are described in the following sections with the overall architecture shown in Figures 1 and 2.

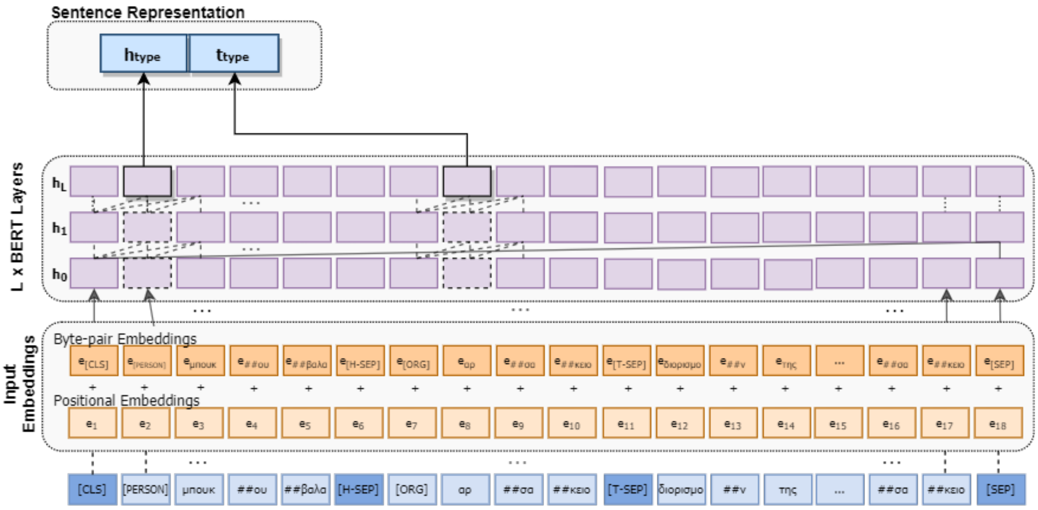

Figure 1. Sentence Representation in REDSandT_Lit. The input embedding h_0 is created by summing the positional and byte pair embeddings for each token in the structured input. States h_t are obtained by self-attending over the states of the previous layer h_{t-1}. The final sentence representation is shaped by concatenating the head entity embedding h_{L_h-type} and the tail entity embedding h_{L_t-type}. Head and tail entity type embeddings are marked with bold lines.

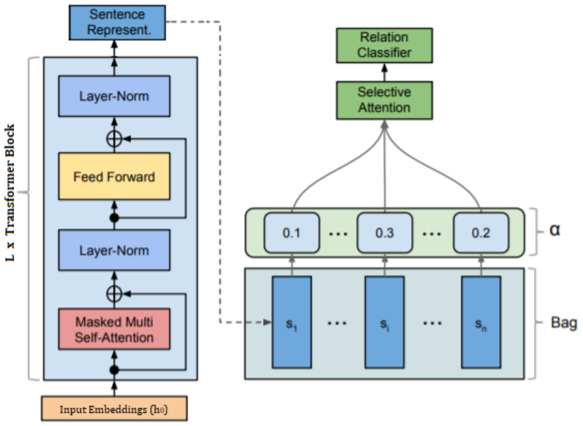

Figure 2. Transformer architecture (**left**) and training framework (**right**). We used BERT transformer architecture and precisely the *bert-base-greek-uncased-v1* GREEK-BERT LM. Sentence representation s_i is formed as shown in Figure 1. Reprinted with permission from [18]. Copyright 2021 Copyright Despina Christou.

3.2.1. Sentence Encoder

Our model encodes sentences into a distributed representation by concatenating the head (h) and tail (t) entity type embeddings. The overall sentence encoding is depicted in Figure 1, while the following sections examine in brief the parts of the sentence encoder in a bottom-up manner.

In order to capture the relation hidden between an entity pair and its surrounding context, RE requires structured input. To this end, we encoded sentences as a sequence of tokens. At the very bottom of Figure 1 is this representation, which starts with the head entity type and token(s) followed by the delimiter (H- SEP), continues with the tail entity type and token(s) followed by the delimiter [T- SEP] and ends with the token sequence of a compressed form of the sentence. The whole input starts and ends with the special delimiters [CLS] and [SEP], respectively, which are typically used in transformer models. In BERT, for example, [CLS] acts as a pooling token representing the whole sequence for downstream tasks, such as RE. We do not follow that convention. Furthermore, tokens refer to the sub-word tokens of each word, where each word is also lower-cased and normalized in terms of accents and other diacritics; for example the word "Αρσάκειο" (Arsakeio) is split into the "αρ" ("ar"), "##σα" ("##sa") and "##κειο" ("##keio") sub-word tokens.

Input Representation

As discussed in Section 3.1, samples including a relation can include up to three sentences; thus, samples generally referenced as sentences within the document can entail information which is not directly related to the underlying relation. Moreover, creative writing's focus on narration results in long secondary sentences that further disrupt the content linking the two entities. In order to focus on the important to the relation tokens, we adopt two distinct compression techniques, namely the following:

- *trim_text_1*: Given a sentence, it preserves the text starting from the three preceding words of the head entity to the three following words of the tail entity;
- *trim_text_2*: Given a sentence, it preserves only the surrounding text of the head and tail entities, with surrounding text referring to the three preceding and following words of each entity.

Our selection is based on the fact that context closer to the entities holds the most important relational information. We experimented with two compressed versions of the text, one that keeps all text between the two entities (*trim_text_1*) and one that keeps only the very close context (*trim_text_2*) assuming that the in-between text, if long enough, typically constitutes a secondary sentence, irrelevant to the underlying relation. Our assumption is reassured in our experiments (see Sections 4 and 5).

After suppressing the sentences to a more compact form, we also incorporate the head and tail entities text and types in the beginning of the structured input to bias LM focusing on the important for the entity pair features. Extensive experimentation reveals that the extracted entity type embeddings hold the most significance information for extracting the underlying relation within two entities. Entity types are considered known and are also provided in the dataset.

Input Embeddings

Input embeddings to GREEK-BERT are presented as h_0 in Figure 1. Each token's embedding results from summing the positional and byte pair embeddings for each token in the structured input.

Position embedding is an essential part of BERT's attention mechanism, while byte-pair embedding is an efficient method for encoding sub-words to account for vocabulary variability and possible new words in inference.

To make use of sub-word information, the input is tokenized using byte-pair encoding (BPE). We use the tokenizer of the pre-trained model (35,000 BPEs) to which we added seven task-specific tokens (e.g., [H-SEP], [T-SEP] and five entity type tokens). We forced the model not to decompose the added tokens into sub-words because of their special meaning in the input representation.

Sentence Representation

Input sequence is transformed into feature vectors (h_L) using GREEK-BERT's pre-trained language model fine-tuned in our task. Each sub-word token feature vector

$(h_L i \ldots D_t)$ is the result of BERT's attention mechanism over all tokens. Intuitively, we do understand that feature vectors of specific tokens are more informative and contribute more in identifying the underlying relationship.

To the extent that each relation constrains the type of the entities involved and vice versa [30,43], we represent each sentence by concatenating the head and tail entities' type embeddings:

$$s_i = [h_{L_{h-type}}; h_{L_{t-type}}] \tag{1}$$

where $s_i \in \Re^{d_h * 2}$.

While it is typical to encode sentences using the vector of the [CLS] token in h_L [11], our experiments show that representing a sentence as a function of the examining entity pair types reduces noise, improves precision and helps in capturing the infrequent relations.

Several other representation techniques were tested; i.e., we tested the method of also concatenating the [CLS] vector to embed the overall sentence's information and also using the sentence representation from [18], including relation embeddings and further attention mechanisms, with the presented method to outperform. Our intuition is that the LM was not able to efficiently capture patterns in Katharevousa since manual observation revealed most words to have split in many sub-words. This occurs because Katharevousa differs to Modern Greek, while some words/characters were also misspelled in the OCR process.

3.3. Bag Encoder

Bag encoding, i.e., aggregation of sentence representations in a bag, comes to reduce noise generated by the erroneously annotated relations accompanying DS.

Assuming that not all sentences equally contribute to bag's representation, we use selective attention [24] to highlight the sentences that better express the underlying relation.

$$B = \sum_i \alpha_i s_i, \tag{2}$$

As observed in the above equation, selective attention represents each bag as a weighted sum over its individual sentences. Attention α_i is calculated by comparing each sentence representation against a learned representation r:

$$\alpha_i = \frac{exp(s_i r)}{\sum_{j=1}^{n} exp(s_j r)} \tag{3}$$

At last, the bag representation B is fed to a softmax classifier in order to obtain the probability distribution over the relations:

$$p(r) = Softmax(W_r \cdot B + b_r), \tag{4}$$

where W_r is the relation weight matrix, and $b_r \in \Re^{d_r}$ is the bias vector.

3.4. Training

Our model utilizes a transformer model, precisely GREEK-BERT, which fine-tunes on our specific setup to capture the semantic features of relational sentences. Below, we briefly present the overall process.

Pre-training

For our experiments, we use the pre-trained *bert-base-greek-uncased-v1* language model [16], which consists of 12 layers, 12 attention heads and 110M parameters where each layer is a bidirectional Transformer encoder [31]. The model is trained on uncased Modern Greek texts of Wikipedia, European Parliament Proceedings Parallel Corpus (Europarl) and OSCAR (clean part of Common Crawl) with a total of 3.04B tokens. GREEK-BERT is pre-trained using two unsupervised tasks, masked LM and next sentence prediction, with masked LM being its core novelty as it allows the previously impossible bidirectional training.

Fine-tuning

We initialize our model' s weights with the pre-trained GREEK-BERT model and fine-tune only the last four layers under the multi-instance learning setting presented in Figure 2, using the specific input shown in Figure 1. After experimentation, only the last four layers are fine-tuned.

During fine-tuning, we optimize the following objective:

$$L(D) = \sum_{i=1}^{|B|} logP(l_i|B_i;\theta) \tag{5}$$

where for all entity pair bags $|B|$ in the dataset, we want to maximize the probability of correctly predicting the bag's relation (l_i) given its sentences' representation and parameters (ϑ).

3.5. Experimental Setup

3.5.1. Hyper-Parameter Settings

In our experiments we utilize *bert-base-greek-uncased-v1* model with hidden layer dimension $D_h = 768$, while we fine-tune the model with *max_seq_length* $D_t = 128$. We use the Adam optimization scheme [44] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a cosine learning rate decay schedule with warm-up over 0.1% of training updates. We also minimize loss using the cross entropy criterion.

Regarding dataset-specific REDSandT_Lit model's hyper-parameters, we automatically tune them on the validation set based on F1- score. Table 2 shows the applied search space and selected values for the dataset-specific hyper-parameters.

Table 2. Dataset-specific Model Hyper-parameters.

Hyper-Parameter	Search Space	Optimal Value
Batch Size	[4, 8, 16]	8
Epochs	[3, 4]	3
Dropout	[0.2, 0.4, 0.5, 0.6]	0.4
Learning Rate	$[5e^{-3}, 5e^{-4}, 6.25e^{-5}, 5.5e^{-5}]$	$5e^{-5}$
Weight Decay	[0.01, 0.001]	0.001
Fine-tuned layers	[last(2, 4, 8), all]	last 4

Experiments are conducted in Python 3.6, on a PC with 32.00 GB RAM, Intel i7-7800X CPU@ 3.5 GHz and NVIDIA's GeForce GTX 1080 with 8 GB. Fine-tuning takes about 5 min for the three epochs. The implementation of our method is based on the following code: https://github.com/DespinaChristou/REDSandT (accessed on 18 May 2021).

3.5.2. Baseline Models

In order to show the proposed method's effectiveness, we compare against three strong baselines in our dataset. More precisely, we compare REDSandT_Lit to the standard feature-based [45] and NN-based [46] approaches used in the literature while also comparing to the Greek version of BERT [16]. All models were tested on both sentence compression formats presented in Section 3.2.1 and are indicated with respective (1, 2) superscripts. For the Bi-LSTM approach we also experimented with both full-word and BPE tokenization indicated with ($\star$) and ($\star\star$) superscripts, respectively.

Feature-based Methods

- SVM^1: A Support Vector Machine classifier. Sentences are encoded using the first-presented compression format.

- SVM^2: A Support Vector Machine classifier. Sentences are encoded using the second-presented compression format.

NN-based Methods

- $BiLSTM^{1,\star}$: A Bidirectional Recurrent Neural Network (RNN) classifier. Sentences are encoded using the first-presented compression format, while full-word tokenization is used.
- $BiLSTM^{1,\star\star}$: A Bidirectional RNN classifier. Sentences are encoded using the first-presented compression format, while BPE tokenization is used.
- $BiLSTM^{2,\star}$: A Bidirectional RNN classifier. Sentences are encoded using the second-presented compression format, while full-word tokenization is used.
- $BiLSTM^{2,\star\star}$: A Bidirectional RNN classifier. Sentences are encoded using the second-presented compression format, while BPE tokenization is used.

Transformer-based Methods

- *GREEK-BERT*: BERT (*bert-base-uncased*) fine-tuned on modern Greek corpora. We fine-tune this to our specific dataset and task.
- $REDSandT^2$: The default REDSandT approach for distantly supervised RE. We use *GREEK-BERT* as base, and we fine-tune the model on our corpus and specific task. Sentences are encoded using the second-presented compression format.
- $REDSandT\_Lit^1$: The proposed variant of REDSandT fine-tuned on our corpora and specific task. Sentences are encoded using the first-presented compression format.
- $REDSandT\_Lit^2$: The proposed variant of REDSandT fine-tuned on our corpora and specific task. Sentences are encoded using the second-presented compression format.

3.5.3. Evaluation Criteria

In order to evaluate our model against baselines, we report accuracy macro-P, R, F and weighted-P, R, F for all models. For a more in-depth analysis of models' performance in each relation, we report Precision, Recall and F1-score metrics for all models and relations. Moreover, we conduct Friedman's statistical significance test to compare all presented models on our dataset, following [47,48].

4. Results

In this section, we present the results of our model against the predefined baselines both overall and for each relation, separately.

4.1. Overall Models Evaluation

Table 3 compares our model to the baseline models mentioned above. We observed the following: (1) both $REDSandT\_Lit^1$ and $REDSandT\_Lit^2$ are better overall in terms of precision, recall and F1-score, followed by SVM^2 and $BiLSTM^{2,*}$; (2) preserving the surrounding context of entity pairs (*trim_text_2*) almost always results in better results; and (3) using full-word tokenization in Bi-LSTM models shows a tremendous performance improvement over using BPE tokenization. Focusing on the $REDSandT\_Lit$ models, a detailed investigation of their performance on each separate relation showed that the high accuracy achieved by $REDSandT\_Lit^1$ was mainly due to that model being highly accurate in identifying "NoRel" relations. This explains the differences in macro vs. weighted metrics of $REDSandT\_Lit^1$.

Moreover, when it comes to training times, the SVM models are clearly the winner with training times less than a sec, with the rest models deviating from 4 min (BERT-based trained in GPU) to 20 min (BiLSTM trained in CPU). Moreover, it is worth mentioning that the extra complexity added by bag training induces only 10 s additional training time in REDSandT_Lit compared to the training time of the simple Bert models.

Table 3. Baselines Comparison. We report the overall accuracy (ACC), precision (P), recall (R) and F1-score (F1) at the Test set. For P, R and F1 we present both macro-version and weighted-version of the metrics.

Models	**ACC**	**Precision**		**Recall**		**F1-Score**	
		Macro	**Weighted**	**Macro**	**Weighted**	**Macro**	**Weighted**
$BERT^1$	0.88	0.81	0.89	0.76	0.88	0.75	0.88
$BERT^2$	0.90	0.72	0.90	0.75	0.90	0.73	0.90
$BiLSTM^{1,\star}$	0.90	0.76	0.90	0.80	0.89	0.76	0.90
$BiLSTM^{1,\star\star}$	0.84	0.67	0.84	0.59	0.84	0.61	0.84
BiLSTM^{2,*}	0.91	0.88	0.91	0.78	0.91	0.80	0.90
$BiLSTM^{2,\star\star}$	0.82	0.47	0.82	0.46	0.82	0.46	0.82
SVM^1	0.90	0.84	0.91	0.86	0.90	0.83	0.90
SVM^2	0.91	0.84	0.91	0.88	0.91	0.86	0.91
REDSandT_Lit^1	**0.93**	**0.91**	**0.93**	0.88	**0.93**	0.89	**0.93**
REDSandT_Lit^2	0.90	**0.91**	0.91	**0.92**	0.90	**0.91**	0.91

In order to validate the contribution of all presented models, we compare (i) all examined models and (ii) the best performed ones by using the Friedman's statistical test. As observed in Table 4, the p-value of both compared model sets is less than 0.05 (actually close to zero); thus, we have sufficient evidence to conclude that using different models results in statistical differences in the predicted relations and that our outcomes are statistical significant.

Table 4. Friedman's Statistical Test—We compare (i) all models and (ii) only the best performed models (those highlighted in bold in Table 3).

Compared Models	**Friedman's Statistical Test**	
	Statistic	***p*-Value**
All models	84.85	$2e^{-14}$
Best performed models	47.69	$4e^{-11}$

4.2. Models Evaluation on Each Relation

Tables 5–7 compare our models to the above-mentioned baselines across all relations, reporting precision, recall and F1-score, respectively. Overall, we observed following: (1) the $REDSandT\_Lit$ models exhibit strong performance across all relations, while $REDSandT\_Lit^2$ best captures relations in the long-tail; (2) SVM^1, SVM^2 and $BERT^2$ are generally consistent but all Bi-LSTM models exhibit significant performance variabilities; and (3) SVM models perform well regardless of chosen sentence compression.

Table 5. Baselines Comparison—We report Precision (P) (in % format) at Test set for all relations.

Models	**Precision**						
	NoRel	**ArtAuthor**	**PubDate**	**WorkAt**	**OrgPlace**	**OrgDate**	**artHero**
$BERT^1$	97	88	65	73	80	100	0
$BERT^2$	96	83	77	64	75	75	100
$BiLSTM^{1,\star}$	96	25	75	83	77	94	83
$BiLSTM^{1,\star\star}$	91	0	100	67	67	65	78
BiLSTM^{2,*}	94	100	100	80	83	68	91
$BiLSTM^{2,\star\star}$	91	0	0	29	54	73	81
SVM^1	95	89	73	74	100	60	100
SVM^2	95	91	77	73	80	75	100
REDSandT_Lit^1	98	86	88	89	78	100	100
REDSandT_Lit^2	95	85	82	75	100	100	100

Table 6. Baselines Comparison—We report Recall (R) (in % format) at Test set for all relations.

Models				Recall			
	NoRel	ArtAuthor	PubDate	WorkAt	OrgPlace	OrgDate	artHero
$BERT^{1}$	91	90	93	94	57	100	0
$BERT^{2}$	90	89	82	94	43	100	33
$BiLSTM^{1,\star}$	93	33	100	71	86	88	87
$BiLSTM^{1,\star\star}$	90	0	67	29	64	77	85
BiLSTM^{2,*}	95	33	100	57	86	88	86
$BiLSTM^{2,\star\star}$	92	0	0	29	79	47	76
SVM^{1}	94	88	79	82	57	100	100
SVM^{2}	94	86	86	94	57	100	100
REDSandT_Lit^{1}	90	96	100	100	100	67	67
REDSandT_Lit^{2}	88	95	100	88	71	100	100

Table 7. Baselines Comparison—We report F1-score (F) (in % format) at Test set for all relations.

Models				F1-score			
	NoRel	ArtAuthor	PubDate	WorkAt	OrgPlace	OrgDate	artHero
$BERT^{1}$	94	89	76	82	67	100	0
$BERT^{2}$	93	86	79	76	55	86	50
$BiLSTM^{1,\star}$	94	29	86	77	81	91	85
$BiLSTM^{1,\star\star}$	90	0	80	40	66	70	81
BiLSTM^{2,*}	94	50	100	67	85	77	88
$BiLSTM^{2,\star\star}$	91	0	0	29	64	57	79
SVM^{1}	94	89	76	78	73	75	100
SVM^{2}	94	88	81	82	67	86	100
REDSandT_Lit^{1}	94	91	93	94	88	80	80
REDSandT_Lit^{2}	92	90	90	81	83	100	100

5. Discussion

5.1. Error Analysis

Figure 3 presents the confusion matrices for *REDSandT_Lit*^{2} and SVM^{2} models. Even though the SVM model seems to slightly over-perform the *REDSandT_Lit* approach, the confusion matrices show that this superiority comes from the "NoRel" relation. Excluding the "NoRel" relation, *REDSandT_Lit*^{2} model performs much better across all relations including those in the long tail. As previously discussed, "NoRel" relation can include sentences which do not contain a relation or were not annotated. For this reason, we further analyze the performance in this class below.

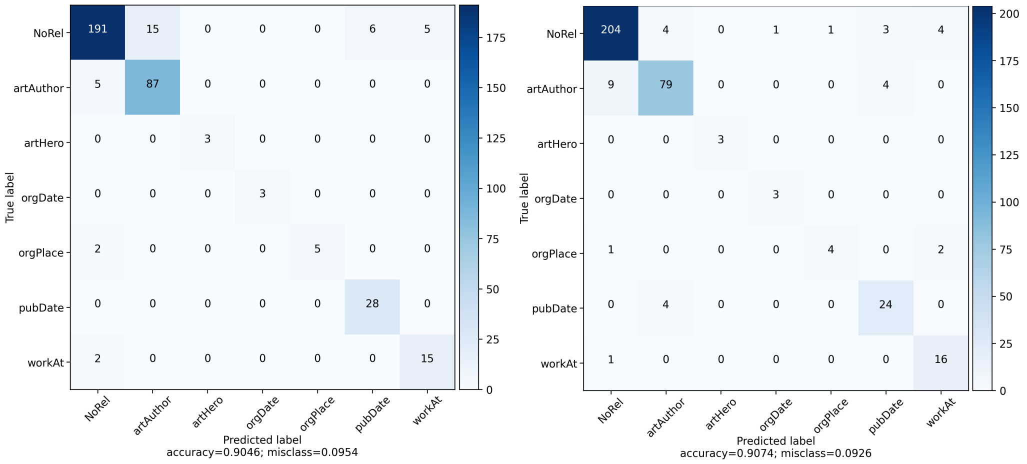

Figure 3. Confusion matrices of *REDSandT_Lit*^{2} (**left**) and SVM^{2} (**right**) models.

5.2. Effectiveness on Mislabelled Sentences

Sentences marked with “NoRel” relation correspond to sentences that include at least two recognized entities but where not annotated with a relation. This can correspond either in no underlying relation within sentence or a missed annotation. In order to examine this case, we further investigate the performance of the best performing models on the “NoRel” relation. Our goal is to reveal the model that can capture missed annotations and propose it as an efficient model that can correct mislabels and augment samples which in our case and industry-wise is of high importance.

Table 8 compares the best two models on predicting mislabelled samples. We observe that $REDSandT\_Lit^2$ is superior to SVM^2 in this task and precisely in identifying “artAuthor” relations within sentences that were not annotated.

Table 8. Comparing the two best performing models ($REDSandT\_Lit^2$, SVM^2) on predicting mislabelled samples in “NoRel” relation.

Model	**Head Entity**	**Tail Entity**	**Pred. Relation**
$REDSandT\_Lit^2$	Οιδίπους Τύραννος (Oedipus Tyrannus)	Σοφοκλής (Sophocles)	“ArtAuthor”
SVM^2			“NoRel”
$REDSandT\_Lit^2$	Πανόραμα της Ελλάδος (Panorama tis Ellados, Panorama of Greece)	1865	“PubDate”
SVM^2			“PubDate”
$REDSandT\_Lit^2$	Χριστοπούλου (Christopoulou)	Λυρικά (Lyrika)	“artAuthor”
SVM^2			“NoRel”
$REDSandT\_Lit^2$	Αγροτικαί Επιστολαί (Agrotike Epistole, Rural Letters)	Γ. Δροσίνης (G. Drosinis)	“artAuthor”
SVM^2			“NoRel”
$REDSandT\_Lit^2$	Τρεις Τάφοι (Treis Tafoi, Three Tombs)	Σοφοκλής Καρύδης (Sofoklis Karudis)	“artAuthor”
SVM^2			“NoRel”
$REDSandT\_Lit^2$	Πανεπιστήμιω του Μονάχου (Panepistimio tou Monaxou, Munich University)	Otto Bardenhewer	“workAt”
SVM^2			"NoRel"
$REDSandT\_Lit^2$	Βρετανικόν Μουσείον (Bretanikon Mouseion, British Museum)	Λονδίνο (Londino, London)	“orgPlace”
SVM^2			“orgPlace’

6. Conclusions and Future Work

We proposed a novel distantly supervised transformer-based relation extraction model, REDSandT_Lit, that can automate metadata extraction from literary texts, thus helping sustaining important cultural insights that otherwise could be lost in unindexed raw texts. Precisely, our model efficiently captures semantic relationships from Greek literary texts of the 19th century. We constructed the first dataset for this language and period, including 3649 samples annotated through distant supervision with six semantic relationships. The dataset is in the Katharevousa variant of Greek, in which a great part of Modern Greek literature is written. In order to capture the semantic and syntactic characteristics of the language, we exploited GREEK-BERT, a pre-trained language model on modern Greek, which we fine-tuned on our specific task and language. To handle the problem of noisy instances, as well as the long sentences that are typical in literary writing, we guided REDSandT_Lit to focus solely on a compressed form of the sentence and the entity types of the entity pair. Extensive experiments and comparisons with existing models on our dataset revealed that REDSandT_Lit has superior performance, manages to capture infrequent relations and can correct mislabelled sentences.

Extensions of this work could focus on augmenting our dataset to facilitate direct BERT pre-training on the Katharevousa form of the Greek language. Even though we achieve high accuracy with pre-trained models in Modern Greek and finetuned on the Katharevousa variant, this inconsistency suggests that augmenting the studied data and providing a model specific to these data can further improve results. Moreover, we would like to further investigate the effect of additional side-information such as POS info and entities description, while also an end-to-end model that is not based on pre-recognized entities and extracts both entities and relations in one pass. At last, although there is extensive research on ancient Greek philosophy, literature and culture, as well as research in modern Greek Natural Language Processing (NLP) tools, the very important (from a cultural, literary and linguistic point of view) Katharevousa form of the Greek language has not been studied in terms of automatic NLP tools. Thus, creating automated tools specific to this form is a step towards revealing important cultural insights for the early years of the modern Greek state.

Author Contributions: Conceptualization, D.C. and G.T.; methodology, D.C.; software, D.C.; validation, D.C. and G.T.; formal analysis, D.C.; investigation, D.C.; resources, G.T.; data curation, D.C.; writing—original draft preparation, D.C.; writing—review and editing, D.C. and G.T.; visualization, D.C.; supervision, G.T.; project administration, G.T.; funding acquisition, G.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work was co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH–CREATE-INNOVATE (project code:T1EDK-05580) in the context of the ECARLE (Exploitation of Cultural Assets with computer-assisted Recognition, Labeling and meta-data Enrichment) project.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available at: https://github.com/intelligence-csd-auth-gr/extracting-semantic-relationships-from-greek-literary-texts (accessed on 3 August 2021).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute;
DOAJ	Directory of open access journals;
AI	Artificial Intelligence;
NLP	Natural Language Processing;
RE	Relation Extraction;
DS	Distant Supervision;
KB	Knowledge Base;
REDSandT	Relation Extraction with Distant Supervision and Transformers;
REDSandT_Lit	Relation Extraction with Distant Supervision and Transformers for Literature;
CNN	Convolutional Neural Network;
PCNN	Piecewise Convolutional Neural Network;
LSTM	Long Short-term Memory;
Bi-LSTM	Bidirectional Long Short-term Memory;
GCNN	Graph-CNN;
LM	Language Model;
BPE	Byte-pair Encoding;

OCR	Optical Character Recognition;
POS	Part-of-Speech.

Appendix A

Table A1. Nineteenth century Greek books catalogue.

id	Book Title
1	Επτανησιώται λόγιοι του 19ου αιώνα. χ.τ., 1890.
2	Α. Πάλλη, Αλέξιος, ή, Αι τελευταίαι ημέραι των Ψαρρών: ιστορικόν διήγημα, Εκ του τυπογραφείου "Η Αυγή", Ζάκυνθος, 1860.
3	Δροσίνης Γεώργιος, Διηγήσεις αγωνιστού ήτοι σκηνογραφίαι εκ της ελληνικής επαναστάσεως: προς ανάγνωσιν εν τη Δ′ τάξει των Δημοτικών Σχολείων κατά το πρόγραμμα του επί του της Δημοσίας Εκπαιδεύσεως Υπουργείου, Κασδόνης, Εν Αθήνα, 1889.
4	Άννινος Μπάμπης, Τα πρώτα έτη του Ζαν Μωρεάς, Δ. και Π. Δημητράκου, Αθήνα, 1860.
5	Δ. Φωτιάδης, Ύμνοι του πρωτομάρτυρος Ρήγα του Φεραίου: μετά συντόμως βιογραφία αυτού, Εκ του Τυπογραφείου της Φιλοκαλίας, Αθήνα, 1878.
6	Ν. Περπινιάν, Οθων και Αμαλία, βασιλεύς και βασίλισσα της Ελλάδος: ύμνοι Β′ τη αρχαία ελληνίδι φωνί και μέτρω, Εκ της τυπογραφίας του Βυζαντοδείκτου, Κωνσταντινούπολη, 1854.
7	Βικέλας Δημήτριος, Από Νικοπόλεως εις Ολυμπίαν : επιστολαί προς φίλον, Εκ του τυπογρ. Ανδρ. Κορομηλά και Κοραή, Αθήνα, 1886.
8	Περβάνογλου Ιωάννης, Μιχαήλ ο Παλαιολόγος: ιστορικόν διήγημα, Λειψία, 1883.
9	Ι. Π. Κόκκαλης, Το φάσμα της προκυμαίας Πατρών: μυθιστορικόν δοκίμιον και διάφορα ποιήματα, Τυπογραφείον η "Αυγή", Ζάκυνθος, 1869.
10	Κοραής Αδαμάντιος, Βίος Αδαμαντίου Κοραή, 2η εκδ., Εκ της τυπογραφίας Γ. Μελισταγούς, Ερμούπολη, 1836.
11	Ξενόπουλος Γρηγόριος, Ελληνικού αγώνος το τριακοσιάδραχμον έπαθλον: διήγημα, Διονύσιος Σ. Χιώτης, Αθήνα, 1885.
12	Ξενόπουλος Γρηγόριος, Στρατιωτικά διηγήματα, Εκδότης Γεώργιος Κασδόνης, Εν Αθήνα, 1892.
13	Σακελλαρόπουλος Σπυρίδων Κ., Εκθεσις των κριτών του Λασσανείου Δραματικού Διαγωνισμού: αναγνωσθείσα υπο του εισηγητού Σ. Κ. Σακελλαροπούλου εν των Πανεπιστημίω τη 28 Μαρτίου 1899, Τύποις Π. Δ. Σακελλαρίου, Αθήνα, 1899.
14	Γ. Βερναρδάκης, Ποικίλα φιλολογικά: απόσπασμα εκ της επετηρίδος του Παρνασσού, Εκ του Τυπογραφείου της Εστίας, Αθήνα, 1900.
15	Σούτσος Αλέξανδρος, Το πανόραμα της Ελλάδος, Εκ του τυπογρ. των αδελφών Περρή, Αθήνα, 1875.
16	Η. Σ. Σταθόπουλος, Του ποιητικού διαγωνισμού του 1857 τα επεισόδια: και μιας λογικής αριθμητικής η επίκρισις, Εκ του τυπογραφείου Ιω. Αγγελόπουλου, Αθήνα, 1857.
17	Παρρέν Καλλιρρόη, Ζωή ενός έτους, Παρασκευάς Λεωνής, Αθήνα, 1897.
18	Κοραής Αδαμάντιος, Απάνθισμα επιστολών Αδαμαντίου Κοραή. Εκ της τυπογραφίας Κ. Ράλλη, Αθήνα, 1839.
19	Αρσένης Ιωάννης, Πάνθεον Ελλήνων ποιητών, 1879.
20	Κόκκινος Εμμανουήλ Χ., Εκθεσις της επί του Μεταφραστικού Αγώνος του κ. Δ. Α. Οικονόμου επιτροπής κατά την Γ′ αυτού περίοδον, Εκ του τυπογραφείου Ερμού, Αθήνα, 1876.
21	Ροΐδης Εμμανουήλ Δ., Το "ταξίδι" του Ψυχάρη, γλωσσική μελέτη, τυπ. Εστία, Αθήνα, 1888.
22	Istria Dora d′, Αι Ιόνιοι νήσοι: υπό την δεσποτείαν της Ενετίας και την Αγγλικήν προστασίαν και η εν αυταίς Ελληνική ποίησις μετα περιλήψεως τινός της αρχαίας αυτών ιστορίας, Εκ του Τυπογραφείου Δ. Ειρηνίδου, Αθήνα, 1859.
23	Δέρβος Γεώργιος Ι., Λόγος εισιτήριος εις το μάθημα της χριστιανικής γραμματολογίας, Εκ του τυπογραφείου Αποστολόπουλου, Αθήνα, 1898.
24	Κοδρικάς Παναγιώτης, Προς τους ελλογιμοτάτους νέους εκδότας του Λόγιου Ερμού: εις Βιένναν της Αουστρίας, Παρίσι, 1816.
25	Παλαμάς Κωστής, Διονύσιος Σολωμός, βιογραφία και κριτική, Ύμνος εις την ελευθερίαν, Αλεξάνδρεια, Pan-African Anglo-Hellenic Editions, Αλεξάνδρεια, 1890.
26	Ορφανίδης Θεόδωρος, Κρίσις του Βουτσιναίου Ποιητικού Αγώνος του έτους, 1876

References

1. Katsan, G. *History and National Ideology in Greek Postmodernist Fiction*; Fairleigh Dickinson: Teaneck, NJ, USA, 2013.
2. Prosser, J. *American Fiction of the 1990s: Reflections of History and Culture*; Routledge: London, UK, 2016.
3. UNESCO. Culture for Development Indicators (CIDS) Methodology Manual. Available online: https://en.unesco.org/creativity/sites/creativity/files/cdis/heritage_dimension.pdf (accessed on 12 August 2021).
4. Jänicke, S.; Franzini, G.; Cheema, M.F.; Scheuermann, G. Eurographics Conference on Visualization (EuroVis) (2015) On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. In *Proceedings of the Eurographics Conference on Visualization (EuroVis)-STARs, Cagliari, Italy, 25–29 May 2015*; Eurographics Association: Cagliari, Italy, 2015; pp. 88–103.
5. Holtorf, C.; Högberg, A. *Cultural Heritage and the Future*; Routledge: London, UK, 2020.
6. Mintz, M.; Bills, S.; Snow, R.; Jurafsky, D. 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, Suntec, Singapore, 2–7 August 2009*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2009; pp. 1003–1011.
7. Riedel, S.; Yao, L.; McCallum, A. Modeling relations and their mentions without labeled text. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Barcelona, Spain, 20–24 September 2010; Volume 6323, pp. 148–163. [CrossRef].
8. Hoffmann, R.; Zhang, C.; Ling, X.; Zettlemoyer, L.; Weld, D.S. Knowledge-based weak supervision for information extraction of overlapping relations. In Proceedings of the ACL-HLT 2011—49th Annual Meeting of the Association for Computational Linguistics-Human Language Technologies, Portland, OR, USA, 21 June 2011; Volume 1, pp. 541–550.
9. Surdeanu, M.; Tibshirani, J.; Nallapati, R.; Manning, C.D. Multi-instance multi-label learning for relation extraction. In Proceedings of the EMNLP-CoNLL 2012—2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea, 12–14 July 2012; pp. 455–465.
10. Elson, D.K.; Dames, N.; Mckeown, K.R. Extracting Social Networks from Literary Fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Uppsala, Sweden, 2010; pp. 138–147.
11. Chaturvedi, S.; Srivastava, S.; Daume, H.; Dyer, C. Modeling evolving relationships between characters in literary novels. In Proceedings of the 30th AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; pp. 2704–2710.
12. Devisree, V.; Reghu Raj, P. A Hybrid Approach to Relationship Extraction from Stories. *Procedia Technol.* **2016**, *24*, 1499–1506. [CrossRef]
13. Makazhanov, A.; Barbosa, D.; Kondrak, G. Extracting Family Relationship Networks from Novels. *arXiv* **2014**, arXiv:1405.0603.
14. Lee, J.; Yeung, C.Y. Extracting networks of people and places from literary texts. In Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation, Bali, Indonesia, 16–18 December 2012; pp. 209–218.
15. Sims, M.; Park, J.H.; Bamman, D. Literary event detection. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 3623–3634.
16. Koutsikakis, J.; Chalkidis, I.; Malakasiotis, P.; Androutsopoulos, I. GREEK-BERT: The greeks visiting sesame street. In Proceedings of the 11th Hellenic Conference on Artificial Intelligence, Athens, Greece, 2–4 September 2020; pp. 110–117. [CrossRef]
17. Devlin, J.; Chang, M.W.; Lee, K.; Google, K.T.; Language, A.I. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
18. Christou, D.; Tsoumakas, G. Improving Distantly-Supervised Relation Extraction through BERT-based Label & Instance Embeddings. *IEEE Access* **2021**, *9*, 62574–62582. [CrossRef]
19. Alt, C.; Hübner, M.; Hennig, L. Fine-tuning Pre-Trained Transformer Language Models to Distantly Supervised Relation Extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Florence, Italy, 2019; pp. 1388–1398. [CrossRef]
20. Craven, M.; Kumlien, J. Constructing biological knowledge bases by extracting information from text sources. In Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, Heidelberg, Germany, 6–10 August 1999; pp. 77–86.
21. Snow, R.; Jurafsky, D.; Ng, A.Y. Learning syntactic patterns for automatic hypernym discovery. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 5–8 December 2005; pp. 1297–1304.
22. Zeng, D.; Liu, K.; Lai, S.; Zhou, G.; Zhao, J. Relation classification via convolutional deep neural network. In Proceedings of the COLING 2014—25th International Conference on Computational Linguistics, Dublin, Ireland, 23–29 August 2014; pp. 2335–2344.
23. Zeng, D.; Liu, K.; Chen, Y.; Zhao, J. Distant supervision for relation extraction via Piecewise Convolutional Neural Networks. In Proceedings of the EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1753–1762.
24. Lin, Y.; Shen, S.; Liu, Z.; Luan, H.; Sun, M. Neural Relation Extraction with Selective Attention over Instances. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA, 7–12 August 2016; Volume 1, pp. 2124–2133. [CrossRef]
25. Liu, T.; Wang, K.; Chang, B.; Sui, Z. A Soft-label Method for Noise-tolerant Distantly Supervised Relation Extraction. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 1790–1795. [CrossRef]

26. Wang, G.; Li, C.; Wang, W.; Zhang, Y.; Shen, D.; Zhang, X.; Henao, R.; Carin, L. Joint embedding of words and labels for text classification. In Proceedings of the ACL 2018—56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; Volume 1, pp. 2321–2331. [CrossRef]
27. Ye, Z.X.; Ling, Z.H. Distant Supervision Relation Extraction with Intra-Bag and Inter-Bag Attentions. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 2810–2819. [CrossRef]
28. Wu, Y.; Bamman, D.; Russell, S. Adversarial Training for Relation Extraction. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 1778–1783. [CrossRef]
29. Qin, P.; Xu, W.; Wang, W.Y. DSGAN: Generative Adversarial Training for Distant Supervision Relation Extraction. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; Volume 1, pp. 496–505.
30. Vashishth, S.; Joshi, R.; Prayaga, S.S.; Bhattacharyya, C.; Talukdar, P. RESIDE: Improving Distantly-Supervised Neural Relation Extraction using Side Information. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 1257–1266.
31. Vaswani, A.; Brain, G.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, K.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
32. Radford, A.; Salimans, T. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed on 10 August 2020).
33. Howard, J.; Ruder, S. Universal Language Model Fine-tuning for Text Classification. In Proceedings of the ACL 2018—56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 328–339.
34. Shi, P.; Lin, J. Simple BERT Models for Relation Extraction and Semantic Role Labeling. *arXiv* **2019**, arXiv:1904.05255.
35. Han, X.; Wang, L. A novel document-level relation extraction method based on BERT and entity information. *IEEE Access* **2020**, *8*, 96912–96919. [CrossRef]
36. Labatut, V.; Bost, X. Extraction and analysis of fictional character networks: A survey. *ACM Comput. Surv.* **2019**, *52*, 89. [CrossRef]
37. Benjamins, V.R.; Contreras, J.; Blázquez, M.; Dodero, J.M.; Garcia, A.; Navas, E.; Hernández, F.; Wert, C. Cultural Heritage and the Semantic Web. In Proceedings of the European Semantic Web Symposium, Heraklion, Crete, Greece, 10–12 May 2004; pp. 433–444.
38. Heravi, B.R.; Boran, M.; Breslin, J. Towards social semantic journalism. In Proceedings of the International AAAI Conference on Web and Social Media, Dublin, Ireland, 4–7 June 2012; Volume 6.
39. Vakkari, P.; Luoma, A.; Pöntinen, J. Books' interest grading and dwell time in metadata in selecting fiction. In Proceedings of the 5th Information Interaction in Context Symposium, Regensburg, Germany, 26–30 August 2014; pp. 28–37.
40. Caswell, D. Structured journalism and the semantic units of news. *Digit. J.* **2019**, *7*, 1134–1156. [CrossRef]
41. Koidaki, F.; Tiktopoulou, K. Encoding semantic relationships in literary texts. A methodological proposal for linking networked entities into semantic relations. In Proceedings of the Ballisage: Markup Conference, Virtual, 2–6 August 2021. pp. 28–37. Available online: http://www.balisage.net/Proceedings/vol26/html/Koidaki01/BalisageVol26-Koidaki01.html (accessed on 21 August 2021).
42. Xu, Y.; Mou, L.; Li, G.; Chen, Y.; Peng, H.; Jin, Z. Classifying relations via long short term memory networks along shortest dependency paths. In Proceedings of the Conference Proceedings—EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1785–1794. [CrossRef]
43. Sennrich, R.; Haddow, B.; Birch, A. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, Berlin, Germany, 7–12 August 2016; Volume 3, pp. 1715–1725. [CrossRef]
44. Kingma, D.P.; Lei Ba, J. Adam: A method for Stochastic Optimization. In Proceedings of the 3rd International Conference for Learning Representations, San Diego, CA, USA, 7–9 May 2015.
45. Cortes, C.; Vapnik, V. Support-vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
46. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]
47. Vázquez, E.G.; Escolano, A.Y.; Riaño, P.G.; Junquera, J.P. Repeated measures multiple comparison procedures applied to model selection in neural networks. In *Bio-Inspired Applications of Connectionism, Proceedings of the 6th International Work-Conference on Artificial Neural Networks, Spain, 13–15 June 2001*; Springer: Berlin/Heidelberg, Germany, 2001; pp. 88–95.
48. Pizarro, J.; Guerrero, E.; Galindo, P.L. Multiple comparison procedures applied to model selection. *Neurocomputing* **2002**, *48*, 155–173. [CrossRef]