Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2213–2222
Marseille, 11–16 May 2020

Domain Adapted Distant Supervision for Pedagogically Motivated Relation Extraction

Oscar Sainz, Oier Lopez de Lacalle, Itziar Aldabe, Montse Maritxalar
University of the Basque Country (UPV/EHU)
{osainz006, oier.lopezdelacalle,itziar.aldabe,montse.maritxalar}@ehu.eus

Abstract

In this paper we present a relation extraction system that given a text extracts pedagogically motivated relation types, as a previous step to obtaining a semantic representation of the text which will make possible to automatically generate questions for reading comprehension. The system maps pedagogically motivated relations with relations from ConceptNet and deploys Distant Supervision for relation extraction. We run a study on a subset of those relationships in order to analyse the viability of our approach. For that, we build a domain-specific relation extraction system and explore two relation extraction models: a state-of-the-art model based on transfer learning and a discrete feature based machine learning model. Experiments show that the neural model obtains better results in terms of F-score and we yield promising results on the subset of relations suitable for pedagogical purposes. We thus consider that distant supervision for relation extraction is a valid approach in our target domain, i.e. biology.

Keywords: distant supervision, relation extraction, domain adaptation, pedagogically motivated relations

1. Introduction

Question generation (QG) has been traditionally linked to the education and psychology domains. However, during the last decade there has also been a growth of interest on QG in the area of computing (Rus and Graesser, 2009) thanks to advances in Artificial Intelligence and Natural Language Processing (NLP). The first published systems on QG (Heilman and Smith, 2010) limited the generation of questions to the sentence level and the generation was mainly based on shallow linguistic information at text level. Then, some QG systems based the generation process on concept maps (Olney et al., 2012; Jouault et al., 2016) and aimed at building a semantic representation of text in order to generate relevant questions. For example, Olney et al. (2012) build a concept map from a biology textbook based on NLP techniques and heuristics, and Jouault et al. (2016) used linked data technologies to extract the semantic information from Wikipedia, Freebase and DBpedia in order to generate meaningful questions for history learning. Both approaches generate questions following Graesser and Person's question taxonomy (Graesser and Person, 1994), a taxonomy that describes domain-independent meaningful question types to support learning.

Interestingly, the concept map extracted in (Olney et al., 2012) contains a set of edges that facilitates linkages between the graph representations and question asking/answering. The main goal of our work is to continue the process made by (Olney et al., 2012), but building the semantic representation based on a Relation Extraction (RE) system.

In this paper we focus on the process to automatically build the semantic representation. For that, we present a RE system that given a text extracts different relation types following (Olney et al., 2012) work. We map these relation types with the relation types from ConceptNet (Speer and Havasi, 2012) and so we deploy Distant Supervision (DS) for RE with pedagogically motivated relations.

It is well-known that the generation of an annotated dataset is expensive and domain related. Thus, in this work, we define a scenario in which having thousands of unstructured and unlabeled documents, and some related knowledge bases (KB), we extract the occurring relations types in the given texts. We avoid the prohibitive cost of manual annotation through DS and automatically generate training data by aligning the information stored in the KBs with the occurrences founded in the corpus (Mintz et al., 2009; Hoffmann et al., 2011).

More specifically, RE under DS aims to predict the relation type of a pair of entities occurring in text according to a KB. DS annotation heuristically aligns occurring entities to a given KB and uses this alignment to learn the relation types. The training data are labelled automatically as follows: for a triplet $r(e_1, e_2)$ in the KB, all sentences that mention both entities e_1 and e_2 are regarded as the training instances of relation r.

Recent works on DS have shown the usefulness of such approaches when building a RE system for which non manual annotations are required. Most of the current RE systems focus on extracting triplets that express the common relation between two given entities, which are inferred from multiple relation mentions (e.g. also known as *slot filling* in knowledge based population (Ji et al., 2010)). Instead, we focus on labeling the relation mentions of a given sentence, which is the previous step to generate a semantic graph that will facilitate the generation of pedagogical questions.

In this paper we automatically build a domain specific corpus directly from Wikipedia. We annotate it using domain adapted DS techniques and we apply heuristically defined filters to remove noisy examples. Finally, we learn and evaluate two paradigms of RE approaches. Current DS systems are evaluated in a set of predefined relation types (e.g. Wikipedia, DBpedia), and are not tested how well these approaches are ported to different set of relations. Thus, we evaluate state-of-the-art approaches on a pedagogically motivated inventory of relations.

Thus, the contributions of the paper are the following:

- We define a set of relationships suitable for pedagogical purposes. For that, we propose a mapping between ConceptNet and the edges from (Olney et al., 2012), and we run a pilot study on a subset of these relationships in order to verify that the selected categories are properly extracted and therefore useful for pedagogical applications.
- We develop a framework that do not need human intervention and gives room to build a domain-specific relation extraction system based on Distant Supervision. The dataset is publicly available under a free license[1].
- We explore relation extraction models based on distant supervision, including DISTRE, a model based on transfer learning (Alt et al., 2019), and a discrete feature based machine learning model. Experiments show the superiority of the neural model over the classic pipeline model in terms of F-score.

2. Previous Work

Distant Supervision Distant Supervision was originally proposed by (Craven and Kumlien, 1999) and focused on extracting binary relations between biomedical entities, e.g. proteins, cells or diseases. Later, distant supervision was improved by (Mintz et al., 2009) and made the approach available for different and more general domains entities, such as people, locations and organizations, among others. Nevertheless, state-of-the-art approaches of relation extraction were still based on lexical and syntactic features (Mintz et al., 2009; GuoDong et al., 2005), and, as it is well-known, the assumption that all sentences containing two related entities express a relation is not always true. Therefore, current state-of-the-art in DS has merely focused on modeling noise to obtain more reliable inference results. According to (Intxaurrondo et al., 2013) noisy labels have three main sources: 1) generation of false negatives due to context which do not express an actual relationship of the given two entities, 2) generation of false negatives due to an incomplete knowledge base , and 3) when the relation of two entities is multi-label, providing noisy label in some cases. Multi-labeling problem was first tackled by (Riedel et al., 2010), and other, with *at-least-one* or multi-instance learning approaches. In a similar vein, to handle the problem where some sentences overlap different relations *multi-instance multi-label* learning (Hoffmann et al., 2011; Surdeanu et al., 2012) approaches were proposed, in which more than one label are modeled per instance and argument pair.

Deep Learning Recently, neural network approaches like PCNN (Zeng et al., 2014) have become more popular and have been extended and improved with multi-instance learning (Zeng et al., 2015) and selective attention (Lin et al., 2016a). Other learning strategies like adversarial-learning (Wu et al., 2017), capsule network (Zhang et al., 2019b) and reinforcement learning (Feng et al., 2018; Takanobu et al., 2019) have been also applied successfully. Moreover, pretraining language models and fine-tuning on specific tasks have shown several improvements on NLP tasks and they have become the most common state-of-the-art pipeline nowadays. For example, DISTRE (Alt et al., 2019) obtains state-of-the-art performance for relation extraction. On the contrary, other methods rely on approaches that make explicit the use of linguistic and semantic information. For instance, (Ji et al., 2017) includes the entity descriptions extracted from Wikipedia into the model in order to extend the background knowledge, (Yaghoobzadeh et al., 2017) proposes a model that learns the entities and the relation extraction jointly, and (Vashishth et al., 2018) uses existing side information in KBs such as relation alias to improve the quality of the weak supervision. In this paper, we evaluate DISTRE and compare it to a classic feature based approach (see Section 5. for further details).

Transfer learning as domain adaptation Transfer learning has been shown as an alternative to domain adaptation when (almost) no annotated data is available in the target domain. For example, Legrand et al. (2018) used syntax based transfer learning to the biomedical domain successfully adapting a tree long short-term memory (LSTM). And, they also showed that transfer learning can be harmful even when the source and the target are very dissimilar, and thus other ways of domain adaptation can be more effective as a first try. Zhang et al. (2019a) try to solve this type of domain-shift problem using relation-gated adversarial learning that deploys transferable features that explain linguistic variations. Di et al. (2019) propose a domain aware transfer learning in which a generic weakly supervised relation extractor is fine-tuned on the knowledge existing in an adapted KB. This work is close to ours as they keep in the KB the information related to the target domain.

3. Relation Types

ConceptNet ConceptNet is a knowledge base that intents to describe using natural language expressions the general human knowledge (Speer and Havasi, 2012). It was initially created in the Open Mind Common Sense project (Singh et al., 2002), and interestingly contains a great variety of relation types that can be easily mapped to the edges proposed by (Olney et al., 2012). It integrates knowledge from different sources with varying levels of granularity and varying registers of formality.

Concepts are connected to natural language words and phrases that can also be found in free text. This property makes ConceptNet suitable for distant supervision. The multilingual nature of ConceptNet makes it interesting as a resource for minority languages. ConceptNet aims to contain both specific facts and the messy, inconsistent world of common sense knowledge. For example, WordNet can tell you that a dog is a *type of* carnivore, but not that it is a type of pet, or WordNet can tell you that a fork *is an* utensil, but has no link between fork and eat to tell you that a fork is *used for* eating. However, this type of knowledge can be both good for DS as we can have wider coverage for label free text and bad, as we might be labeling noisy examples.

Olney et al. (2012) set of edges Olney et al. (2012) manually clustered 4371 biology triples available on the Internet to analyse the relations appearing in them. This analysis resulted in 20 relations. Additional edge relations were

[1] https://osainz59.github.io/BioPMDS/

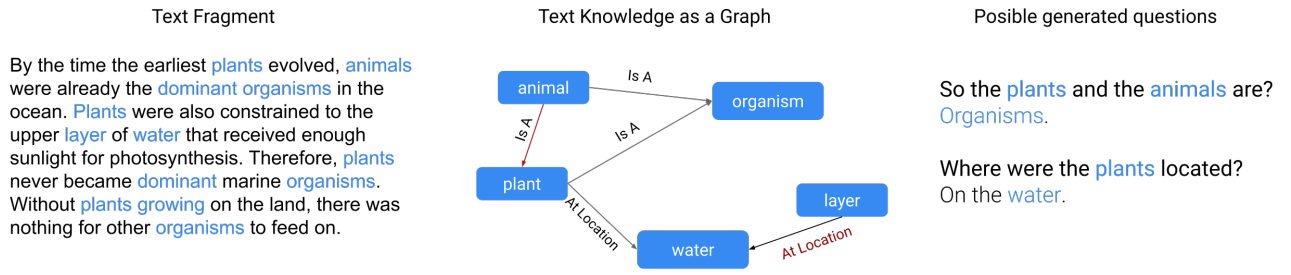

Figure 1: On the left, a **text fragment** with highlighted (on blue) entities related to biology extracted from WordNet Domains. On the middle, a handcrafted **knowledge graph** build from the triples proposed by the RE system (red edges denote incorrect triples). On the right, manually **generated questions** based on Olney et al. (2012) templates.

added based on the psychology literature as well as adjunct information gleaned from a SRL parser, raising the total number of edge relations to 30. As previously reported in concept map research in biology (Fisher et al., 2002), their cluster analysis revealed that 50% of all relations were either *is-a*, *has-part*, or *has-property*. So, we decided to focus on two of the most common relations (*is-a, has-part*) and on two more additional ones: *location* and *function*.

Olney et al. (2012)	ConceptNet
IsA	InstanceOf IsA∗
Location	AtLocation∗ LocatedNear LocationOfAction
HasPart	MadeOf MemberOf PartOf∗
Function	UsedFor∗

Table 1: Mapping of relations between Olney et al. (2012) taxonomy and ConceptNet. ∗ denotes the selected relation type in ConceptNet.

Relation type mapping Table 1 shows the mapping between the relations in ConceptNet and the edges. Note that each relation in Olney et al. (2012) can be mapped to multiple relations in ConceptNet. In the experiments we restricted the mapping to only one relation type in ConceptNet, as we wanted to validate the idea of building a pedagogical relation extractor, and thus control the noise produced by the mapping. We select the ConceptNet relations that are the most semantically close ones: *IsA*, *AtLocation*, *PartOf*, and *UsedFor*. Table 2 shows a set of triplets extracted from ConceptNet and used in our experiments.

Question templates Although the question generation step is out of scope of this work, Figure 1 shows an example of the whole intended process. The construction of the knowledge graph has been extracted using the output of the feature based RE (cf. Section 5.), and the questions were manually generated using the knowledge graph and the templates proposed by Olney et al. (2012).

1st Argument	2nd Argument	Relation
Algae	Water	AtLocation
Cell	Organism	IsA
Adaptation	Evolution	PartOf
Alveolus	Lung	PartOf
Muscle	Move	UsedFor

Table 2: Examples of relations extracted from ConceptNet

4. Corpus Annotation

In this section we describe the process we followed to generate our automatic annotated corpus. The process consist of 4 principal steps: corpus collection, domain adaptation, application of distant supervision and finally the noise removing step.

4.1. Corpus Collection

We use the hierarchy of Wikipedia categories to retrieve the documents related to the biology domain. More specifically, we start from a few category seeds, such as *Science*, *Natural Science*, *Computing*, and *Human Genetics*, 30 categories in total, and collect the categories lying two steps below in the taxonomy. We did not go deeper in the category since the flat nature of the hierarchy can gather many unrelated documents to the biology domain. We create the biology corpus with all the Wikipedia articles linked to the final set of the category set. In total, we use 9477 categories and collect 250,873 articles. Finally, we preprocess the corpus with sentence splitting, tokenization, lemmatization and POS tagging.

4.2. Domain Adaptation

ConceptNet contains a very diverse set of domains and terms, which can be problematic when applying distant supervision heuristics to annotate an unlabeled corpus. Note that we want to train a relation extraction model that is able to extract relations specific to biology, and thus we think that it is not sufficient to provide only a domain-specific corpus into the approach. We think it is key to adapt the knowledge based to the domain in order to minimize the noise produced by the distant supervision approach. For

this, we make the following steps to adapt ConcepNet to biology domain.

Term extraction We define the domain of interest with a set of terms which are related to biology. Biology terms are obtained from WordNet Domains (WNDomains) (Bentivogli et al., 2004). This knowledge base is an extension of WordNet which stores domain information about the synsets. The WNDomains uses a hierarchical structure to define these domains, for example, the term *biology* is a child of *pure_science* and the parent of *anatomy*. So in order to get terms that conform the domain, we just get all of synsets (actually, variants in the synset) which are labeled with *biology* or any domain below biology (e.g. *anatomy)*. In total we collected 41034 terms.

Knowledge-base filtering Once the domain is defined, we apply two filters to adapt the knowledge-base to our domain and relation set. In order to ensure that the triplets are from the biology domain, the first filter consists on keeping those triplets in which both arguments are from biology. The second filter is related to the relations we have decided to use.

Relation	#triplets	Proportions	Term Coverage
IsA	17105	0.868	0.989
PartOf	1881	0.095	0.151
AtLocation	543	0.027	0.031
UsedFor	159	0.008	0.016

Table 3: Statistics about the filtered ConceptNet knowledge-base.

After applying both filters the resultant knowledge-base is described on Table 3. The table shows the number of triplets, the normalize frequency (proportions) and the term coverage for each relation. The term coverage refers to the proportion between the terms that take part in that relation and the terms that take part in at least 1 of the mentioned relations. As expected *IsA* is the most productive relation type when applying distant supervision over unlabeled corpus, and presumably obtain very skewed distribution of labels. We foresee that obtained distribution of triplets across relation types can be challenging for training a relation extraction system that effectively covers all the relation types.

4.3. Application of Distant Supervision

In distant supervision, we make use of a KB and a set of documents to automatically generate our training data. For that, we take into account the entity-pairs and their relation from the KB and label each pair of entities that appear in the same document as a positive example for the given relation. For example, ConcepNet contains the fact that "cell" *is a* "organism". We take this triplet and label each pair of "cell" and "organism" that appear in the same document as a positive example for the *IsA* relation. But, it is also important to generate negative examples to predict that there is no relation between an entity pair.

Negative examples In order to collect negative examples to train our model, we consider that an entity pair which is not related in the KB shows no relation. So, following this hypothesis we label all documents where both entities appear together in the same sentence with the *Nil* relation and considered as a negative example. We extract the arbitrary number of 10,000 negative examples.

Relation	#Examples	Frequency
IsA	172,125	0.660
PartOf	51,821	0.198
AtLocation	13,988	0.053
UsedFor	12,558	0.048
Nil	10,000	0.083

Table 4: Statistics about the distantly labeled corpus before noise filtering.

Table 4 shows the number of examples grouped by the relation types and their frequency. The extracted corpus is clearly unbalanced with the *IsA* relation having 66% of the corpus examples which is again in line with previous findings.

4.4. Removing Noise

Applying DS, we can easily generate a large amount of training data, saving time and money compared to human annotated data. However, we might easily generate noisy training data like most automatic labeling systems. That is why previous works working on DS include the task of removing noise. For instance, Intxaurrondo et al. (2013) and Min et al. (2012) successfully implement and apply several heuristics to automatically detect and remove noisy mentions and so, we also apply some of their methods in our approach. A manual examination showed that many noisy examples were removed, and helped improving the confusion between relation types.

Mention Frequency This first heuristic takes into account the mention frequencies of the triplets to remove some noisy examples. The intuition behind this heuristic is that those triplets with a high frequency tend to have a bigger amount of noisy examples compared to the ones with fewer frequencies. Thus, the idea is to set a threshold so that those triplets with higher frequencies than the threshold are removed. We set the threshold empirically at 100 mentions. Figure 2 shows the mention frequencies of all the triplets in the training dataset before applying this filter. It can be seen that most of the triplets have few mentions while there are few triplets with lots of mentions.

Tuple PMI This filter takes into account the Pointwise Mutual Information (PMI) value of the triple to remove noisy examples. Thus, the filter decides if two entities are related or not based on the mention frequency of the triples. This frequency is measured as follows:

$$PMI(e_1, e_2) = \log \frac{f_{e_1 e_2}}{f_{e_1} \times f_{e_2}} \quad (1)$$

As in the previous filter, it is necessary to define a threshold to decide which triplets to keep and which examples to remove. The threshold was empirically set and consider to maintain those positive triplets with a PMI value bigger than 1.5 as well as to keep all the negative examples.

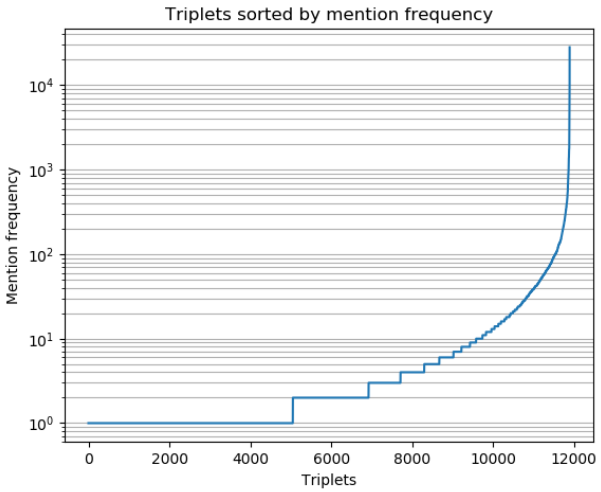

Figure 2: Triplets mention frequency sorted from lowest to highest.

Redundant Triplets ConceptNet contains triplets such as (bird, bird, IsA) and (cell membrane, membrane, IsA) that can be considered redundant. We hypothesise that the context of these type of triplets are not very informative and could be discarded. This filter has been implemented to remove these redundant triplets. Most of the examples refer to the *IsA* relation, so somehow applying this filter helps in balancing the dataset.

Relation	#Examples	Frequency
IsA	34,959	0.597
PartOf	8,044	0.137
AtLocation	4,655	0.079
UsedFor	858	0.014
Nil	10,000	0.170

Table 5: Statistics about the distantly labeled corpus after noise filtering.

Table 5 shows the distribution of the relations after applying the three noise removing filters. The removing was done applying first the redundant triplets filter, then the tuple PMI filter and finally the mention frequency filter. Compared to Table 4 the filtered corpus is much smaller, mainly due to the Mention Frequency filter which removes the triplets with more than thousands mentions. Although this filter removes just 1% of the triplets, this means to remove 46.79% of the examples. Although the corpus is still unbalanced, having the *IsA* relation 59% of the corpus examples, the number of examples has decreased substantially (from 172,125 to 34,959). A similar behaviour can be seen in the rest of the positive examples.
After applying the filters we manually analysed a small set of examples and observed overall better examples. In addition, the results obtained from the automatic evaluation are better. Of course, we can not guarantee that the automatic results are better just as a consequence of applying the filters.

5. Relation Extraction Systems

In the experiments we explore relation extraction models that follow two different paradigms: discrete feature based machine learning models and DISTRE (Alt et al., 2019), which transfers knowledge based on pre-trained language models.

5.1. Feature based Relation Extraction

This extraction machine learning approach makes use of discrete features that model the context of the occurring two arguments. Actually, the feature set used in our approach is a simplified version of the features described in (GuoDong et al., 2005). After some preliminary experiments, we decided to use a Support Vector Machine classifier with a linear kernel.

Feature	Meaning
PF	0 if the first argument comes first in the sentence, 1 in the other case.
WM1	Bag of words of first mention.
HM1	Head of the first mention.
WM2	Bag of words of the second mention.
HM2	Head of the second mention.
HM12	Heads of first and second mentions.
WBNULL	When no words in between both mentions.
WBFL	The only word in between when only one word is in between.
WBF	The first word in between where more than one word is in between.
WBL	The last word in between where more than one word is in between.
WBO	The words in between except the first and the last.
BM1F	The first word before first mention.
BM1L	The second word before first mention.
AM2F	The first word after second mention.
AM2L	The second word after second mention
ET12	Combinations of entity types of both mentions.

Table 6: Description of features used by the system.

Table 6 shows the feature set of the system, mainly composed of lexical features. On the contrary, the feature ET12 encodes some semantic information of the entity type (e.g. animal, artifact or person) of each argument, a feature extracted from the Lexicographer files in WordNet (Miller, 1995).
As an example, Table 7 shows the features extracted for the triplet (cell, organism, IsA) occurring in the following context:

Some organisms are made up of only one cell and are known as unicellular organisms.

Training-set balance As said before, the corpus we have created for the task is unbalanced. This can be a problem for predicting the long tail relations. In order to help the classifier we apply two sampling techniques: SMOTE

Feature	Value
PF:	1
WM1:	organisms
HM1:	organisms
WM2:	cell
HM2:	cell
HM12:	organisms cell
WBF:	are
WBL:	one
WBO:	made up of only
BM1F:	some
AM2F:	and
AM2L:	are
ET12:	Tops Tops

Table 7: Example of the features extracted.

(Bowyer et al., 2011) and TomekLinks (Tomek, 1976). Both algorithms use the nearest-neighbour strategy on the feature-space. On the one hand, the SMOTE algorithm generates new examples for the minority classes by sampling (linearly) interpolated data points between the existing training point and its nearest-neighbours. On the other hand, TomekLinks removes those sample pairs which comes from different classes and are nearest-neighbours between them.

5.2. Deep Learning based Relation Extraction

Deep learning approaches have become very popular in the NLP community due to their impact in the improvement of most NLP tasks. In this work, we decided to compare and evaluate DISTRE (Alt et al., 2019), which is one of the state-of-the-art models on distantly-supervised relation extraction.

Architecture and Pretraining DISTRE is based on the GPT (Radford et al., 2018) model, a pre-trained language-model of the Transformer (Vaswani et al., 2017) architecture. More specifically, it uses the *Transformer Decoder* architecture. The model consists of 12 transformer-decoder blocks with 12 attention heads and 768 dimensional states, and at the top, a feed-forward layer of 3072 dimensional states. This model was pre-trained on BookCorpus (Zhu et al., 2015) using the next token prediction strategy. That is, given a document $D = \{t_1, \ldots, t_n\}$ of tokens t_i, the model has to maximize the following likelihood:

$$\mathcal{L}_{NTP}(D) = \sum_i \log P(t_i | t_{i-1}, \ldots, t_{i-k}, \theta) \qquad (2)$$

where k is the context window considered for predicting the next token t_i using the conditional probability P. For the input representation they reused the original GPT byte-pair vocabulary, but extended with task specific tokens.

Fine-tuning On the fine-tuning process an additional head with the purpose of the relation classification task is added. This head follows the bag-level multi-instance strategy. The mentioned strategy is broadly used on distantly-supervised KB Slot-Filling tasks. Furthermore, they combine the bag-level classification with the Selective Attention (Lin et al., 2016b) in order to model the noise that comes from distant supervision. Finally, they keep the next token prediction objective, according to Radford et al. (2018) it helps the model to generalize better and converging faster at fine-tuning process.

6. Experimental Setup

In the following section we describe our experimental setup. We run the feature based and DISTRE models on our own dataset. We evaluate the models at mention level as our final goal is to make predictions at sentence level[2]. The input of the models are a single sentence and the pair of arguments, whereas the output is the relation type, including *Nil*. Due to the task formulation, we are going to ignore the bag-level classification of DISTRE and evaluate at mention level relation extraction. For that purpose, we generate a bag containing one mention, and thus obtain a prediction for every mention of a triplet.

Held-out data Following previous work (e.g. (Riedel et al., 2010)) we split the dataset into train, dev and test subsets. We keep 70% for training, 20% for the development and the remaining 10% for testing. The splitting process is done triplets-wise. That is, triplets in development and test sets are not seen in the training set. We believe that this way of splitting shows more realistic performance results.

Evaluation metrics For an automatic quantitative evaluation, we use the standard metrics in the Relation Extraction task. That is, we measure Macro-averaged Precision, Recall and F-score of the predictions. We also plotted the Micro-averaged Precision-Recall curves, and calculated the Area Under the Curve (AUC). For further analysis, we carried out an analysis of the confusion matrices and relation-level Precision/Recall curves in order to understand whether the set of relations can be properly extracted. Finally, we also present an error analysis over the predicted relations.

Hyperparameters For the feature based model, we only explore a unique parameter in the development set. We optimized the C error penalty which for this case the most suitable value was around 1.3×10^{-3}.
However, for the case of the DISTRE model we used the hyperparameter set proposed by Alt et al. (2019) and did not perform any exploration of the hyperparameters. For the experiments, the ADAM optimization scheme was used with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, a batch size of 8, a learning-rate of 6.25×10^{-5}, and a linear rate decay schedule with warm-up over 0.2% of training updates. During the training a residual dropout of 0.1 and a classifier layer dropout of 0.2 was applied. The model was trained over 10 epochs and we kept the 7th epoch checkpoint based on the results obtained in the development set.

7. Experiment Results

In this section we report the results of the performed experiments in our own dataset. We first compare DIS-

[2]Note that the evaluation differs significantly from tasks like slot-filling, in which predictions at mention level are aggregated to generate a triplet as candidate

TRE and feature based model on the held-out test dataset, then we provide an analysis of relation type confusion and precision-recall curves.

7.1. Held-out Evaluation

The Table 8 shows the performance of the models in the test set. The table reports the obtained AUC, and the macro-averaged precision, recall and Fscore. In the case of the feature based ML (FBML) model we also include an ablation study to see how each training condition affects to the results. This way, we report the results of FBML model trained with the default hyperparameters and unbalanced dataset (first row in the table), optimized hyperparameters in unbalanced dataset (+opt), default classifier in balanced dataset (+sampling), and optimized model in balanced dataset (+sampling & opt).

	AUC	Precision	Recall	F-score
FBML	0.70	0.560	0.460	0.471
+ opt	0.69	0.564	0.462	0.475
+ sampling	0.69	0.566	0.466	0.477
+ sampling & opt	0.723	0.533	0.548	0.532
DISTRE	0.726	0.676	0.554	0.573

Table 8: Precision/Recall AUC and Macro-averaged Precision, Recall and F-score of RE systems in the test set.

The DISTRE approach gets the best results in all cases. Although, it is interesting to see that the best feature based model (FBML+sampling&opt) is very close in terms of recall, and therefore, reported F1-score for both approaches only show a difference of 4 points. On the contrary, we have a very different behavior for precision. DISTRE outperforms the best FBML in 14 points in terms of precision. Regarding the training configurations of the feature based models, results show that there is a significant improvement when over and down-sampling techniques are applied to address the class imbalance problem. In general the precision keeps equal across all training configuration, but we get a substantial improvement of 8 points on recall when we optimize the C error penalty in a balanced dataset.

It is worth to note that although there are differences in Fscore between DISTRE and FBML models, the AUC values remain similar with only a difference of 0.003 between the best two models. The precision-recall curves in Figure 3 show that both cover more or less the same area but with some differences, while the feature based model achieves better precision on extremes of the recall, the DISTRE model shows a more stable curve, giving around 80% of precision for up to almost 70% of recall.

7.2. Analysis

Relation type confusion In order to know whether the selected relation types are properly discriminated, we looked at confusion matrices shown in the Figure 4.

Both matrices show similar patterns, and it seems both models are highly correlated with the original class distribution shown in the distantly annotated corpus. Without taking into account the confusion induced by the majority

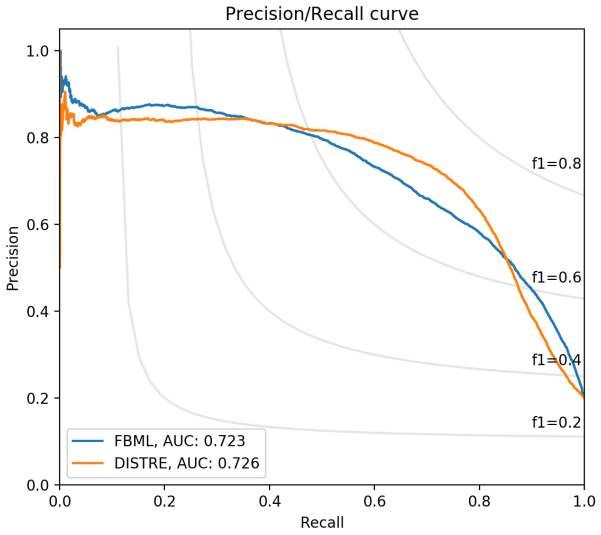

Figure 3: Comparison of Precision/Recall curves of Feature Based and DISTRE models.

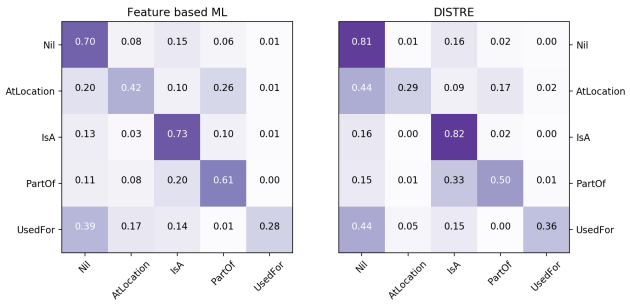

Figure 4: Confusion matrices for both relation extraction systems.

class, both models show that there is a considerable amount of miss-classifications between *AtLocation* and *PartOf* relation types, in this case in favour of *PartOf*.

On the other hand, the difference between both systems seems to be also related to class frequencies. While the DISTRE model performs significantly better on the very frequent and very infrequent relation types, the feature based model performs better on the middle range frequency relations. This can be partially explained with the ability of deep learning models to learn over large amount of data, while the classical ML based approaches can perform better in situations there is no huge amount of data to learn. This is no true for *UsedFor* relation type, where DISTRE attains better results than the feature based model. Nevertheless, this is in accordance with experiments shown in (Alt et al., 2019), in which DISTRE outperforms the rest of the models in the long tail relation types.

For a more exhaustive analysis of the precision-recall tradeoff, we provide a dedicated precision/recall curve of each relation type in Figure 5. The figure shows that the unique relation which maintains stable across all the curve is the *IsA* relation. This could be expected as the relation type

Sentence	1st Arg.	2nd Arg.	Relation	Confidence
By the time the earliest **plants** evolved, **animals** were already the **dominant organisms** in the ocean.	animal organism organism	plant plant animal	IsA IsA IsA	0.03 0.22 0.48
Plants were also constrained to the upper **layer** of **water** that received enough sunlight for photosynthesis.	plant layer	water water	AtLocation AtLocation	0.87 0.58
Therefore, **plants** never became **dominant marine organisms**.	organism	plant	IsA	0.13
Without **plants growing** on the land, there was nothing for other **organisms** to feed on.	organisms	plant	IsA	0.44
Land could not be colonized by any other **organisms** until land **plants** became established.	plants	organisms	IsA	0.24

Table 9: Examples of the output of feature based model. On the left, the sentence where domain terms appears in bold. On the right, for each entity pair the most probable relation with the confidence level. Also, if the relation was predicted in both directions, the direction with highest confidence it is returned.

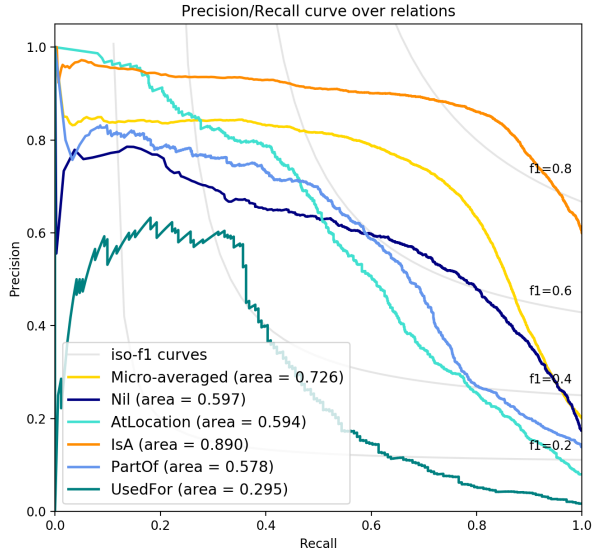

Figure 5: Precision/Recall curves obtained from DISTRE for each relation type in the subset.

is the most frequent class by far. The rest of the relation types show a curve where the precision decreases constantly while recall increases. Except for the *UsedFor* relation type, which also in the range of low recall attains a very low precision too. In the scenario where high precision is preferred we could use all the relation types except *UsedFor*, which shows more problematic performance.

Qualitative analysis We perform a qualitative analysis using some texts fragments from CK-12 Biology (Brainard et al., 2012) book. The CK-12 Foundation provides free K-12 open educational resources aligned to curriculum standards of US. The CK-12 Biology book is part of the FlexBook collection that the foundation provides. We found interesting the use of examples from this particular source to find out if the system is able to produce useful responses in a more realistic scenario.

The Table 9 shows the output of the feature based model for some sentences extracted from the book. Domain terms appears in bold in the table, and for each pair of entities the most probable relation type is shown on the right side of the table. For convenience we remove the argument pairs that show *Nil* relation types. Also, if a given pair of arguments exhibit a relation type in both directions, we select the direction with highest confidence value.

The direction of the relation seems to be one of main problems in systems. Many examples in Table 9 show that problem. For instance, (organism, plant, IsA) should be (plant, organism, IsA) in both cases. In general terms, if we ignore the problem with directionality, the system is able to identify pedagogical relations between entities. There are some miss-classified examples like (animal, plant, IsA), but they have low confidence value and could be discarded applying some confidence thresholds.

8. Conclusions and Future Work

In this paper we have focused on the process to automatically build the semantic representation of a text. For that, we have presented a domain-specific relation extraction system based on distant supervision to extract pedagogically motivated relation types described in (Olney et al., 2012). We mapped these relations with relations from ConceptNet to deploy distant supervision and we have explored two relation extraction models: a deep learning model based on transfer learning and a discrete feature based machine learning model. We run a pilot study on a subset of these relations and obtained better F-score results applying the neural model. Additionally, we obtain promising results on the automatic extraction of the selected categories suggesting that distant supervision for relation extraction is an interesting approach in such pedagogically motivated domain.

We consider that these promising results need to be corroborated in a more realistic scenario. For that, we plan to use all the mappings between the relations in order to build a more complete relation extraction system. Similarly, we consider interesting to carry out further research on transfer-learning models. Finally, we want to intrinsi-

cally evaluate the approach by first creating the semantic graph using the automatic relation extraction system, and then generating pedagogically useful questions. This will give us the real pedagogical value of our approach.

9. Acknowledgements

This work has been partly supported by the Spanish Ministry of Economy and Competitivenes under the deepReading Project RTI2018-096846-B-C21 (MCIU/AEI/FEDER,UE).

10. Bibliographical References

Alt, C., Hübner, M., and Hennig, L. (2019). 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*, pages 1388–1398, Florence, Italy, July. Association for Computational Linguistics.

Bentivogli, L., Forner, P., Magnini, B., and Pianta, E. (2004). Revising the wordnet domains hierarchy: Semantics, coverage and balancing. In *Proceedings of the Workshop on Multilingual Linguistic Ressources*, MLR '04, pages 101–108, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bowyer, K. W., Chawla, N. V., Hall, L. O., and Kegelmeyer, W. P. (2011). SMOTE: synthetic minority oversampling technique. *CoRR*, abs/1106.1813.

Brainard, J., Akre, B., Blanchette, J., Gray-Wilson, N., and Wilkin, D. (2012). *CK-12 Biology*. CK-12.

Craven, M. and Kumlien, J. (1999). Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 77–86. AAAI Press.

Di, S., Shen, Y., and Chen, L. (2019). Relation extraction via domain-aware transfer learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 1348–1357, New York, NY, USA. ACM.

Feng, J., Huang, M., Zhao, L., Yang, Y., and Zhu, X. (2018). Reinforcement learning for relation classification from noisy data. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Fisher, K., Wandersee, J., and Moody, D. (2002). *Mapping Biology Knowledge*, volume 11. 01.

Graesser, A. C. and Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal*, 31(1):104–137.

GuoDong, Z., Jian, S., Jie, Z., and Min, Z. (2005). Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 427–434, Stroudsburg, PA, USA. Association for Computational Linguistics.

Heilman, M. and Smith, N. A. (2010). Good question! statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 609–617, Los Angeles, California, June. Association for Computational Linguistics.

Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., and Weld, D. S. (2011). Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 541–550, Portland, Oregon, USA, June. Association for Computational Linguistics.

Intxaurrondo, A., Surdeanu, M., Lopez de Lacalle, O., and Agirre, E. (2013). Removing noisy mentions for distant supervision. *SEPLN*, 09.

Ji, H., Grishman, R., Dang, H. T., Griffitt, K., and Ellis, J. (2010). Overview of the tac 2010 knowledge base population track. In *In Third Text Analysis Conference (TAC*.

Ji, G., Liu, K., He, S., and Zhao, J. (2017). Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Jouault, C., Seta, K., and Hayashi, Y. (2016). Content-dependent question generation using lod for history learning in open learning space. *New Generation Computing*, 34(4):367–394, Oct.

Legrand, J., Toussaint, Y., Raïssi, C., and Coulet, A. (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, October. Association for Computational Linguistics.

Lin, Y., Shen, S., Liu, Z., Luan, H., and Sun, M. (2016a). Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133.

Lin, Y., Shen, S., Liu, Z., Luan, H., and Sun, M. (2016b). Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November.

Min, B., Li, X., Grishman, R., and Sun, A. (2012). New york university 2012 system for kbp slot filling.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (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, August. Association for Computational Linguistics.

Olney, A., Graesser, A., and Person, N. (2012). Question generation from concept maps. *Dialogue & Discourse*, 3, 03.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.

Riedel, S., Yao, L., and McCallum, A. (2010). Modeling relations and their mentions without labeled text. In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part*

III, ECML PKDD'10, pages 148–163, Berlin, Heidelberg. Springer-Verlag.

Rus, V. and Graesser, A. (2009). Workshop report: The question generation task and evaluation challenge. *Institute for Intelligent Systems, Memphis, TN, ISBN: 978-0-615-27428-7*.

Singh, P., Lin, T., Mueller, E., Lim, G., Perkins, T., and Zhu, W. (2002). Open mind common sense: Knowledge acquisition from the general public. *Lecture Notes in Computer Science*, 2519:1223–1237, 01.

Speer, R. and Havasi, C. (2012). Representing general relational knowledge in ConceptNet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3679–3686, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Surdeanu, M., Tibshirani, J., Nallapati, R., and Manning, C. D. (2012). Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 455–465, Jeju Island, Korea, July. Association for Computational Linguistics.

Takanobu, R., Zhang, T., Liu, J., and Huang, M. (2019). A hierarchical framework for relation extraction with reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7072–7079.

Tomek, I. (1976). Two Modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics*, 7(2):679–772.

Vashishth, S., Joshi, R., Prayaga, S. S., Bhattacharyya, C., and Talukdar, P. (2018). RESIDE: Improving distantly-supervised neural relation extraction using side information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1257–1266, Brussels, Belgium, October-November. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Wu, Y., Bamman, D., and Russell, S. (2017). Adversarial training for relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1778–1783.

Yaghoobzadeh, Y., Adel, H., and Schütze, H. (2017). Noise mitigation for neural entity typing and relation extraction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1183–1194, Valencia, Spain, April. Association for Computational Linguistics.

Zeng, D., Liu, K., Lai, S., Zhou, G., Zhao, J., et al. (2014). Relation classification via convolutional deep neural network.

Zeng, D., Liu, K., Chen, Y., and Zhao, J. (2015). Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762, Lisbon, Portugal, September. Association for Computational Linguistics.

Zhang, N., Deng, S., Sun, Z., Chen, J., Zhang, W., and Chen, H. (2019a). Transfer learning for relation extraction via relation-gated adversarial learning.

Zhang, X., Li, P., Jia, W., and Zhao, H. (2019b). Multi-labeled relation extraction with attentive capsule network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7484–7491.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.