Making Efficient Use of a Domain Expert's Time in Relation Extraction

Linara Adilova, Sven Giesselbach, and Stefan Rüping

Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS
Schloss Birlinghoven, 53757 Sankt Augustin, Germany
{linara.adilova,sven.giesselbach,stefan.rueping}@iais.fraunhofer.de

Abstract. Scarcity of labeled data is one of the most frequent problems faced in machine learning. This is particularly true in relation extraction in text mining, where large corpora of texts exists in many application domains, while labeling of text data requires an expert to invest much time to read the documents. Overall, state-of-the art models, like the convolutional neural network used in this paper, achieve great results when trained on large enough amounts of labeled data. However, from a practical point of view the question arises whether this is the most efficient approach when one takes the manual effort of the expert into account. In this paper, we report on an alternative approach where we first construct a relation extraction model using distant supervision, and only later make use of a domain expert to refine the results. Distant supervision provides a mean of labeling data given known relations in a knowledge base, but it suffers from noisy labeling. We introduce an active learning based extension, that allows our neural network to incorporate expert feedback and report on first results on a complex data set.

Keywords: relation extraction, convolutional neural networks, distant supervision, multi instance learning, interpretability, expert feedback

1 Introduction

Nowadays, huge collections of textual data exist that do not only include interesting documents for humans to read, but can also be mined for interesting knowledge, which can be further stored in structured form. Examples include extracting general world knowledge from Wikipedia [Vrandečić and Krötzsch, 2014], extracting knowledge about interactions of drugs, genes, and diseases from PubMed [Craven et al., 1999, Herrero-Zazo et al., 2013], or definitions from arbitrary scientific publications [Augenstein et al., 2017]. Currently, machine learning methods based on deep neural networks play an important role in the extraction of knowledge from texts, achieving top results on many benchmark data sets [dos Santos et al., 2015, Lee et al., 2017]. However, experience shows that deep learning works best in a supervised setting with a massive amount of labeled data. In practical applications, the effort to manually curate a large enough labeled data set is often prohibitively high, in particular in more specialized domains where

highly trained domain experts are required. Even in cases where one is willing to invest a high manual effort, it may make more sense to extract the required knowledge completely manually because of the unfavorable ratio between effort and precision/recall for a supervised machine learning approach.

In this paper, we address the problem of extracting relations from a large collection of documents with only negligible manual effort on the side of a domain expert. We target situations where currently knowledge extraction cannot be applied economically with respect to the manual effort required. Our approach is based on the idea of integrating the expert into the knowledge extraction process not before a deep learning method (as a mere labeling device), but instead by enabling the experts understanding of the extracted model and enabling him give high-level feedback on the results, that will be used to optimize the model. A way of making use of knowledge - in the form of knowledge graphs - in relation extraction is distant supervision. In distant supervision knowledge that is stored in a knowledge graph is aligned with textual corpora. This yields a cheap way of automatically labeling new training data.

This paper is organized as follows: in the next section, we discuss background and related work on text mining methods with a focus on deep learning. Section 3 introduces our approach, both giving a detailed description of the interactive knowledge extraction process and the distantly supervised deep network that is applied in the intermediate steps. Section 4 gives first empirical results on the proposed approach. Section 5 concludes and gives an outlook to future research.

2   Background and Related Work

In this section we shortly describe the theoretical background behind this paper and set it into the context of related work.

2.1   Relation Extraction

The task of Relation Extraction is about getting semantic meaning out of the sentences and texts that contain two entities mentions. This semantic meaning is later aligned with one of the pre-defined relations (so-called "fixed schema") or it can also be taken in its natural form as a new relation [Riedel et al., 2013]. Classical and the most used method for Relation Extraction is to get all possible linguistic characteristics of the raw textual data and then apply different kinds of classifiers on top of these constructed feature vectors [Zhang et al., 2006, Culotta and Sorensen, 2004]. With the development of Deep Learning this approach was also applied to Relation Extraction [Zeng et al., 2014, Nguyen and Grishman, 2015, dos Santos et al., 2015]. Also, the question of finding ways to perform Relation Extraction without costly construction of training datasets was always getting a lot of attention. For example, Open Information Extraction [Banko et al., 2007] performs the extraction without any human input. Furthermore, Distant Supervision was introduced [Mintz et al., 2009] as a way of utilizing existing structured data for obtaining training dataset without manual labeling of the examples.

2.2   Ranking Convolutional Neural Networks for Relation Extraction

In [dos Santos et al., 2015], a convolutional neural network for relation extraction is introduced. The model consists of multiple layers, which we will quickly describe in this subsection:

1. **Word embedding layer:** Transforming words of the input sentence to the embeddings. Every word w_i of the sentence transformed to r^{w_i} that is a row of an embedding matrix W^{wrd} for some fixed-size vocabulary.
2. **Distance embedding layer:** Transforming distances between the words in the sentence and marked named entities to the embedding vectors wp_1 and wp_2. This approach was introduced by [Zeng et al., 2014].
3. **Embedding merge layer:** Concatenates the word embedding r^w and corresponding distance embeddings (to the first and to the second named entity) wp_1 and wp_2 for every word w in the input sentence into one vector.
4. **Convolutional layer:** Convolution is applied to windows of three embedding vectors with zero padding, so the size of the input is not changed after the layer application. The number of filters d^c is 1000, where each value in one vector is a feature value for a specific triplet of words.
5. **Global maxpooling:** The maximal value is found for each filter.
6. **Scoring dense layer** In order to classify relations the closeness of a sentence representation to real valued vectors representing each of the relations that are learned during the training process is estimated. The scoring procedure is implemented as a dense layer without bias with weights matrix $W^{classes}$ consisting of the relations embeddings.

The objective function uses two of the resulting scores - one is the score, that was obtained for the correct relation according to the label of the example and the second score is one of the wrong relations scores. Thus, the objective function is calculated as follows:

– Get the score to the correct relation;
– Get the maximal score from the remaining wrong relations;
– Calculate the value of the loss according to the formula:

$$L = log(1 + exp(\gamma(m^+ - s_\theta(x)_y^+))) + log(1 + exp(\gamma(m^- + s_\theta(x)_c^-)))$$

where m^+ (m^-) is a margin for the right (wrong) answer; γ is a scaling factor; $s_\theta(x)_y^+$ is a score for the right class; $s_\theta(x)_c^-$ is a score for the wrong class.

2.3   Distant Supervision

While deep learning architectures, such as the one discussed in the previous section, show excellent performance given enough labeled examples, in practice the problem arises that also abundant sentences can usually be found, labeling

enough sentences is usually hard. Labeling is usually done either be crowd-sourcing [Angeli et al., 2014] by non-experts - which negatively influences the quality of the labels - or by experts, which in practice very much limits the amount of labeled examples one can generate.

In order to alleviate this problem, the approach of distant supervision [Mintz et al., 2009] has been proposed. In order to apply this concept, along with the text corpus a structured knowledge base that contains examples of the desired relation ht base will be used to automatically generate examples of the relation by aligning the entities from the knowledge base with the text cee string matching or more complex entity recognition solutions can be used. Hence, distant supervision relies on the following two assumptions:

1. For every triple (e_1, e_2, r) in a knowledge base every sentence containing mentions for e_1 and e_2 expresses the relation r
2. Every triple that is not in the knowledge base is assumed to be a false example for a relation (even though the reason might be in the incompleteness of the knowledge base)

Evidently, the better the knowledge base and text corpus fulfill these assumptions, the better one can expect the approach of distant supervision to work. In practice, it must be assumed that in addition to correct example sentences for the relation, additional noise is introduced.

2.4   Multi-instance Learning

For coping with the noise introduced by distant supervision, we apply multi instance learning as described in [Zeng et al., 2015]. Multi-instance learning was first introduced for drug classification [Dietterich et al., 1997].

In application for relation extraction, multi-instance learning will mean that we assume existence of at least one sentence containing the description of the relation from the knowledge base. So the set of all sentences that mention the same entity pair is considered as one bag and it has a corresponding label of the relation from the knowledge base. In order to apply neural networks for this bag-based training the maximal score example is chosen from the bag every time to fit the model, while all bags are shuffled from epoch to epoch. This approach still looses a lot of possibly useful information obtained by distant supervision, but it serves as an initial step for possible improvement of the approach.

2.5   Interpretability of a Deep Neural Network

The interpretability of machine learning models, and in particular of deep networks, is currently receiving much attention. Several different directions for making a deep neural network model understandable to a domain expert have been proposed in the literature

Rule extraction: Early approaches often focused on the extraction of rules or other understandable representations from neural networks, e.g. [Thrun, 1995]. However, for complex data such as texts, and complex models, there is rarely a concise understandable model that summarizes the whole network, hence these approaches have fallen out of interest.

Relevance propagation and feature weights: for each individual prediction, it is possible to trace which part of the model in each layer was how relevant for taking the final decision [Bach et al., 2015, Binder et al., 2016]. Taking the relevance propagation back to the level of the input features, this gives an importance for each input feature.

Local Approximations: each prediction of the classifier can be approximated locally by simpler model [Ribeiro et al., 2016]. The understandability of the approximating model can be guaranteed by using a simple model class, e.g. a linear model.

Joint Models: in certain cases, it is also possible to construct models that both give a highly accurate prediction plus a reason for the prediction. E.g. in [Lei et al., 2016], together with the model an explanatory phrase is extracted.

Instance-based models: these methods explain a classifier by means of representative instances, such as prototypes (typical well-classified instances), or critics (typical mis-classified instances) [Kim et al., 2016].

In general, generic method for classifier interpretation are hard to adapt to models working on text data, since it is a complex data type of low structure. In the case of the approach of [dos Santos et al., 2015], which is used in this paper, the authors suggest to extract representative trigrams from the text. The idea is based on relevance propagation, however make use of the property that in this model the convolutional layer is applied to three embeddings of consecutive words each. Therefore, for each trigram in the original sentence, its relevance for the predicted relation can easily be traced back. More precise, representative trigrams can be obtained from the sentences of the dataset by measuring the value that each of the trigrams in a sentence contribute to the correct class score. The value is simply a sum of all score positions that are traced back to that specific trigram. This method is very similar to the one mentioned in [Craven et al., 1999] when the most valuable words were extracted in order to have an insight into the concept learned by the model.

3 Model Description

For our experiments we first implemented the ranking convolutional neural network as described in [dos Santos et al., 2015]. We added multi instance learning to cope with noise from distant supervision as well as a feedback loop for domain experts, that lets them improve the performance by evaluating the most representative trigrams for each relation type.

3.1 Expert Feedback for Training Data Curation

An obvious problem with data set creation via distant supervision is that it adds training sentences that do not represent the relation they were sampled for. Instead of letting experts review all of those samples we propose an approach in which the expert reviews not each sentences but rather the concepts that the neural network learned for reach relation.

The representative trigrams that our model learns for each relation class can be regarded as its concepts for each relation. If the concepts make sense, our model most likely has achieved a good understanding of the relation from the distant supervised data. If not, assuming that our model is appropriate, the training data is probably not representative for the relations.

We propose the following workflow for dataset creation and model improvement, displayed in Figure 1:

1. Acquire/Construct a knowledge base with representative facts for relations of interest and text corpora that contain information about the entities and relations of interest.
2. Align the knowledge of the knowledge base with the text corpora and train a Ranking CNN with multi instance learning on it.
3. Extract representative trigrams for each relation class.
4. Show the representative trigrams to experts. Use the trigrams for performance evaluation of the model. Let the expert analyze what mistakes happened and let them filter out non-representative trigrams.
5. Filter out the training sentences for the relations which contain non-representative trigrams and start the process again with the redefined training set.
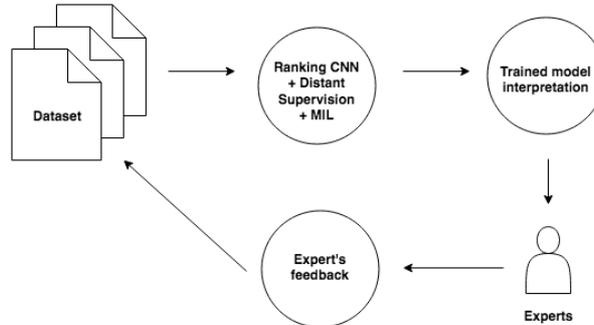

Fig. 1. Active learning approach diagram.

Our assumption here is that the sentences that contain non-representative trigrams are the ones that confuse the network the most. Removing them should at least lead to better precision of the models. The analysis of the trigrams that

the network deems as most important can yield many insights on the causes of bad accuracies.
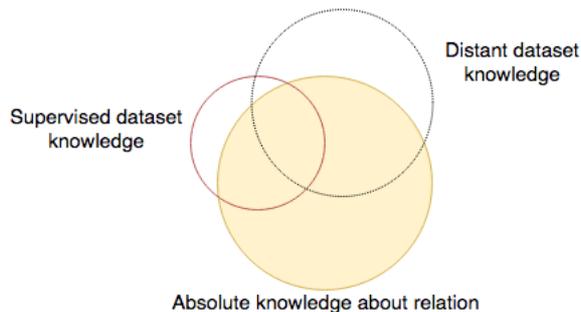

Fig. 2. Concept of knowledge about a specific relation contained in supervised and distantly supervised datasets.

If we imagine the concept of a relation as a set of knowledge, than ideally a supervised datasets capture the whole knowledge. In practice this is hardly feasible. An expert would have to label as many relevant sentences as possible, to capture the variance of the whole knowledge about the relation. Real supervised datasets rather represent a subset of the knowledge about a relation and also some noise e.g. because of wrong labeling. If we add a distantly supervised dataset we will most likely capture 3 different subsets of the overall knowledge: (1) knowledge or noise that is already included in the supervised dataset, (2) new knowledge and (3) new noise. With our proposed workflow we hope to reduce the size of the third subset, namely the newly introduced noise. This idea is reflected in Figure 2.

It is important to note that knowledge that is not reflected in the supervised training set will most likely also not be reflected in the supervised testing set. This leads to the assumption that we will underestimate the performance of our distantly supervised models when evaluating on test sets of supervised data sets.

4   Evaluation

We now evaluate the model architecture. We compare supervised training against distantly supervised training and highlight benefits and downsides of both approaches. We investigate the influence of multi-instance-learning and joint supervised and distant supervised learning. Lastly we evaluate the effect of the expert feedback on the quality of the resulting model.

4.1   Data Sets

SemEval Task8 We first evaluate our dataset on the SemEval task8 dataset. This dataset was originally used by [dos Santos et al., 2015] and we use it for model validation and comparison. The dataset contains nine bidirectional relation types and the "Other" class, that includes different relations, not included in the main ones. Hence there are 19 different relation classes. The sample sentences were manually collected from the web and annotated in three rounds, ensuring that all annotators agree on the label given to the sentence.

The KBP37 Dataset The KBP37 dataset[1], as it was called in the paper [Zhang and Wang, 2015], is a revision of MIML-RE annotation dataset from [Angeli et al., 2014], that was build from a subset of Wikipedia articles by manual annotation. The benefit of KBP37 is that it is alignable with the Wikidata[2] and KBP-slot-filling datasets[3]. The following changes were made to the KBP37 datasets by the authors of [Zhang and Wang, 2015] to adapt it to the description of the SemEval task8:

– Added direction to the relations, i.e. 'per:employee-of(e1,e2)' and 'per:employee-of(e2,e1)' instead of simply 'per:employee-of'. This is done for all the relations except for 'no-relation'
– Balance the dataset, to exclude the relations that have less than 100 examples for each of the directions. Also 80% of 'no-relation' examples are discarded
– After that examples are shuffled and split into three parts, 70% for train, 10% for development and the rest for testing.

After all modifications the dataset consists of 18 directional relations and one "no-relation" class, that will result in 37 classes for recognition. The dataset is more complex than the SemEval task8 dataset. It contains longer sentences (almost twice as long as the longest in SemEval) and it also has multi-relational pairs, making it closer to the real world problem of relation extraction but also more difficult to solve. Also it can be observed that the relations and entities in this dataset are more specific. Most of the entities in the dataset are either names of persons or companies. The relations are very specific, e.g. there are three different classes for placement of headquarters of a company. One for city, state and country.

One more important aspect of the dataset is that human labeling is error prone. Thus there are also very imprecise examples. Here are two examples for the alternate names class:

It was because of < e1 >Abu Talib< /e1 > 's (a.s.) good fortune that apart from < e2 >his< /e2 > ancestral services and prestige he also inherited from sons of Ismail (a.s.) high status and courage. **per:alternate-names(e2,e1)**

---

[1] https://github.com/zhangdongxu/kbp37
[2] https://query.wikidata.org/
[3] https://nlp.stanford.edu/software/mimlre.shtml

The discography of $< e1 >$ Billie Piper $< /e1 >$ (as known as $< e2 >$ Billie $< /e2 >$) an English pop music singer consists of two studio albums two compilation albums and nine singles. **per:alternate-names(e2,e1)**

In the second sentence we have a well labeled example for the class. The first sentence though is hardly an alternate name. It is rather an example for an anaphora resolution task. Such ambiguous labeling will make the classification task even more difficult as it is not obvious even for human annotators why both examples should belong to the same class.

Knowledge Bases for Distant Supervision As knowledge bases for distant supervision we used both relational pairs from MIML-RE[4], i.e. from TAC KBP, and Wikidata[5]. Wikidata [Vrandečić and Krötzsch, 2014] is a crowd-sourced knowledge base. Its users collaborate on filling it with facts, but they also collaborate on validating the data and updating the scheme of the knowledge base. The TAC KBP data is from a knowledge base population task by the Text Analysis Conference, with the goal of discovering information about entities and incorporate it into a Knowledge Base. For relational facts alignment the knowledge base of the Stanford Natural Language Processing group was used[6].

Knowledge base relations were aligned with the New York Times corpus[7]. The amount of entity pairs for each of the relations varies a lot - from less than 1000 to more than 50000. In order to create an artificial "Other" class we chose the relations "per:religion", "per:children" and "org:political/religious-affiliation". When investigating the entity pairs from MIML-RE we found them to be not very accurate. An example being an entity of the type "person-name" that contains only a single letter. In order to minimize noise effect of these pairs, entity pairs from Wikidata were added to the knowledge base. Wikidata contains less matching data for the corresponding relations, but the relations are more precise. We additionally cleaned entity pairs by removing the ones containing one-letter entities or names consisting only of capital letters with dots.

To align the knowledge bases with our textual corpus, we simply matched the strings of the entity names with the texts. If a sentence includes both entities which are part of a relation, we used it as sample for the relation.

4.2 Supervised Training Evaluation

To validate the correctness of the implementation of the ranking convolutional neural network described in Section 2.2, it was tested on the test set of SemEval2010 Task8 dataset and KBP37 test dataset. The scores we achieved is compared to other scores in Table 1. We can conclude, that the model achieves

---

[4] https://nlp.stanford.edu/software/mimlre.shtml
[5] https://query.wikidata.org/
[6] https://nlp.stanford.edu/software/mimlre.shtml
[7] https://catalog.ldc.upenn.edu/ldc2008t19

comparable quality to the reference model and our implementation seems to be correct. We also notice, that the results achieved with our CNN are higher than with the Recurrent Neural Network from [Zhang and Wang, 2015].

Classifier	SemEval2010	KBP37
CR-CNN [dos Santos et al., 2015]	84.1	-
RNN [Zhang and Wang, 2015]	79.6	58.8
Supervised Ranking CNN	**84.39**	**61.26**

Table 1. F1-scores for testing datasets.

4.3   Distant Supervision Evaluation

The results of training the network in various ways with distant supervision can be seen in Table 2. For comparison we also add the results of supervised training.

Experiment	P	R	F1	Manual Effort
Supervised training	**67.74**	**57.88**	**61.26**	17638
Distantly supervised training	50.71	45.24	43.81	**0**
Distantly supervised + MIL	51.82	46.61	45.40	**0**

Table 2. Precision, Recall, F1-scores and manual effort (number of sentences the expert has to label) for distantly supervised training evaluation.

While distant supervision performs worse than supervised training - which was to be expected - the results are usable in practice. In particular, they are significantly higher results than random assignment (with 37 classes, the F1-score for random assignment would be around 0.2%). A publicly available knowledge base and appropriate text corpora can hence serve for the automatic creation of a training set for a neural network tackling the task of relation extraction.

Moreover, in the context of the task to continuously extract new knowledge from newly published texts under a constraint budget of manual intervention, this approach is more appealing than both manual extraction and supervised training. We can quantify the savings on the side of the expert by evaluating how many sentences an expert would have to read in each of the settings: for the KBP37 dataset, the number of sentences in the testing set is 3403 and in order to get relations from them the experts should fully comprehend all the information. Moreover, with the manual approach this should be done for all new texts again. The manually supervised approach would require full comprehension for creating the training dataset that is 17638 sentences and later the experts would check the obtained results (1969 sentences). For the distant supervision on the other

hand, all that is required is a result check that is around 1586 sentences and it can be repeated continuously to get all the relations.

The second observation we can make is that multi-instance learning has a slight positive effect on the performance. Multi-instance learning improved the results in every experiment by almost 2%. Multi-instance learning did improve precision and recall simultaneously.

It is also important to notice, that the supervised training and testing datasets are tightly coupled and they will have common context and common biases. Thus evaluating the distantly supervised model on the existing testing dataset might not be an objective choice. There exist other ways to evaluate the results of distant supervision, for example, as done in [Mintz et al., 2009], but they would not show a realistic comparison to the supervised results.

Furthermore, we investigate the dependency between the performance of the approach and the complexity of the sentences. The dependency can be seen in the Figure 3. Spikes around the large values of length are not representative, as the number of the examples there much smaller (3-5 sentences). For all other values, with higher length the number of errors grows and the number of right answers drops. Any distantly supervised dataset will always be characterized by longer sentences on average, so this aspect should be taken into account when the dataset is constructed. For example, sentences longer than some limit can simply not be included in the final set of training examples.
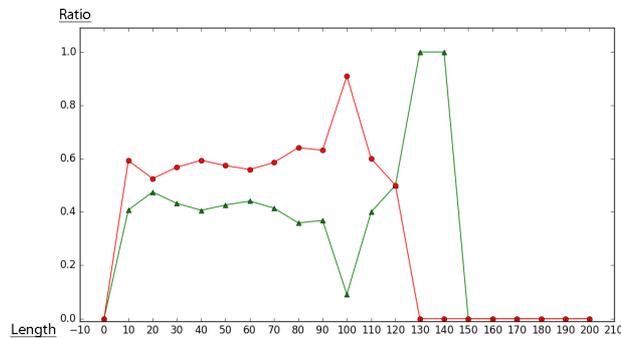
Fig. 3. Correlation of amount of correct and wrong answers with sentence length. Number of correct answers (green) and wrong answers (red) is normalized by the overall amount of examples of specific length.

To inspect the model in more detail, we extracted the representative trigrams for each class, see Table 3. A first immediate finding of looking at the trigrams was that many of them make sense but tend to include the names of entities and might hence even overfit to the names in the training set. For example, for the

org:founded-by	*founder of the; open society institute; fox broadcasting company; ethical treatment of*
per:alternate-names	*known as dwight; known as dj; known as milli; known as matthew; real name is; real name was; , mimi smith; name was selena*
org:members	*soccer league milwaukee; american league boston; national league colorado; football league saskatchewan; midwest league burlington; hockey league .; football league and; basketball league ,; football league 's; soccer league ,*
org:top-members/employees	*said gene russianoff; chief operating officer; , managing director; , chief executive; chief executive of; sony pictures entertainment; executive vice president; , vice president*
per:countries-of-residence	*england .; france .; states .; australia .; united states ,; philharmonic .;) italy ::; the like ,*
org:founded	*, 2000 .; the 1980 's; , 2001 ,; in 1997 ,; in 1996 ,; , 2001 ,; , 2000 ,; in 1997 ,; in 1999 ,; in 1998 ,*
org:subsidiaries	*, a subsidiary; high school in; the walt disney; a division iii; the university of; high school ,; department stores company; general motors corporation*
per:employee-of	*(columbia); secretary of state; senator daniel inouye; senator sam brownback; (columbia); (interscope); (atlantic); blue note label*
per:country-of-birth	*england .; states .; france .; africa .; united states ,; united states in; , england ,; united states attorney*
per:cities-of-residence	*los angeles ,; los angeles band; revved-up vancouver outfit; in london ,; city .; paris .; los angeles ,; angeles .*
org:alternate-names	*states department of; california , los; and municipal employees; the university of; known as dwight; known as dj; known as milli; known as matthew*
org:country-of-headquarters	*the university of; states .; japan .; germany .; in london ,; york city ,; arbor , mich; cambridge , mass*
org:stateorprovince-of-headquarters	*university school of; the university of; , ohio ,; university .; the university of; university in tokyo; life insurance company; institute of technology*
per:spouse	*benazir bhutto ,; brad pitt and; david lynch ; starring david arquette; and her husband;by richard gere; director herbert ross; starring michael douglas*
org:city-of-headquarters	*in london ,; york city ,; arbor , mich; cambridge , mass; the university of; hill , n; arlington , va; city .;*
per:stateorprovinces-of-residence	*california .; york .; of california at; florida .; new york ,; new york city; in california ,; new york times*
per:title	*) film review; director of;) television review; the actor who; the director of; this film is; director of; prime minister ,; the director ,*
per:origin	*of american art; the american artist; the american painter; 20th-century american art; american art ,; american art .; american academy of; american art at; french mathematician ,*

Table 3. Representative trigrams.

relation *org:founded*, it is obvious that the concrete years should be replaced by a placeholder.

4.4   Using Expert Feedback

We have seen that we can construct a useful data set for relation extraction using distant supervision and multi instance learning. Now we want to evaluate whether feedback from experts about the concepts learned by our model can be used to improve the quality of our dataset and the model. For this experiment we used the model trained on the distantly supervised data set with the relations from KBP37 and sentences from the New York Times corpus.

To evaluate the approach of integrating expert feedback to improve the model, we conduct the following experiment: from the representative trigrams of Table 3, we select nonsensical trigrams plus trigrams that are too specific, e.g. overfit on specific names. Sentences matching those trigrams are removed from the training set, as they are suspected to introduce too much noise, and the model is trained again on the filtered data set. Table 2 shows the results, with classes where the F1-score changes by less than 1.00 between the initial and filtered run are excluded because of space constraints.

Class	F1-score initial	F1-score filtered	New sensible trigrams
org:founded-by	**19.30**	13.33	2
org:members	**14.86**	13.64	0
org:top-members/employees	**44.77**	40.49	1
per:alternate-names	24.72	**31.17**	1
per:cities-of-residence	52.82	**54.73**	0
per:countries-of-residence	9.33	**13.20**	0
per:country-of-birth	25.83	**26.86**	0
per:employee-of	**43.39**	39.91	3
per:spouse	**43.56**	36.00	1
per:stateorprovinces-of-residence	43.95	**45.49**	1
per:title	87.45	**87.55**	2

Table 4. Changes in F1-score after filtering examples with non representative trigrams in them. Training is performed without Multiple Instance Learning. Classes with difference in F1 smaller that 1.00 excluded due to space constraints.

In detail, the following effects can be seen in the trigrams:

"per:cities-of-residence": all the trigrams contained names of the cities. Here, even after filtering all the new trigrams contain only city names.

"per:countries-of-residence": a lot of non representative trigrams were filtered. As a result, network started to concentrate more on the persons names in the form of "johan anderson of", "vanessa gusmeroli of".

"org:founded-by": most of the trigrams included companies names, so they were filtered out. This allowed to obtain other trigrams such as "clifford noble opened" or "dick clark productions" but it worsened the overall score as previously learned company names were not taken into account anymore.

"org:members": the data to the information about sport leagues and the trigrams contained only leagues names both before and after filtering.

"per:stateorprovince-of-residence": performance improved as it started to learn also constructions like "pete domenici of".

"org:top-members/employees": had a lot of persons names in its trigrams. So, filtering them out again affected overfitting of the network, but it allowed to get such trigrams as "editor of the" for example.

"per:alternate-names": filtering out trigrams with names allowed to get "real name was" for example without loosing "real name is" and "known as dwight". In this case the network started to see really good constructions.

"per:country-of-birth": after filtering started learning persons names more, that helped it to give better results.

"per:employee-of": overfits to companies names. Filtering trigrams allowed to get "of state" and "former defense secretary" but worsened the result because it does not make conclusion by the company name anymore.

"per:spouse": the relation has a lot of training examples with celebrity names. All of such trigrams were filtered out. It allowed to learn at least "her husband," leaving all the others names again.

At first glance, the results may look unconvincing: results improve for 5 relations, but get worse for 5 relations. However, looking at the trigrams before and after the filtering the following two observations can be made:

1. **Performance is mostly influenced by overfitting on entities:** it is clear from looking at the trigrams, that very often concrete names of cities, persons, or organizations are learned, which is not a desired behaviour. Because of the random training and test split, very often these entities occur in both training and test data, such that good results are obtained still. Removing these trigrams has a negative effect in most cases, as no more general relations can be learned. Interestingly, in some cases removing non-sensical trigrams allows the network to identify even more concrete entities which improves the results, e.g. in the case of "per:country-of-birth".

2. **More sensible trigrams improve the results:** some examples, e.g. the relation "per:alternate-names" or "per:stateorprovince-of-residence" show improved results with more sensical trigrams.

In summary, it might be meaningful for the expert to make the decision on which trigrams to include based on a comparision of the trigrams both before and after the filtering: in the case where no more meaningful trigrams are found, it might make sense to conclude that no general model can be found and not filter the overfitted trigrams afterall.

5   Conclusion and Outlook

Despite the many successes of deep learning in relation extraction, for many practical problems, the availability of labeled data is the main limiting factor. Due to the complexity of the knowledge that is to be extracted from the texts, supervised approaches need many more examples than what usually is available in practical applications. In this paper we explored possibilities to make use of a domain experts knowledge in a more efficient way than using him as a labeling device.

It has been shown that distant supervision, in combination with multi-instance learning, is a meaningful method for relation extraction and well surpasses both manual information extraction and state-of-the-art supervised approaches when performance in relation to manual effort is concerned. The necessary effort by the domain experts can in this case be constrained to the identification of a meaningful structured database for generating distantly supervised examples.

An analysis of distant supervision and multi-instance learning in the specific case of the KBP37 dataset showed that the quality of the attainable results can be limited effects of overfitting on specific entities. We have shown that in this case the domain expert can contribute by inspecting the predictions made by the deep model on the level of representative trigrams. With the insight gained, the expert can contribute to improve the quality of the model by removing examples that were wrongly labeled by distant supervision, or giving input on pre-processing steps that may help the generalization ability of the model.

Future work will aim at a more in-depth evaluation of the approach. Our hypothesis is that the presented approach will be more effective in the case of more specialized relations and in-depth knowledge, for example in the case of medical texts. Finally, obviously representative trigrams are only a very coarse tool for making the model more understandable.

References

[Angeli et al., 2014] Angeli, G., Tibshirani, J., Wu, J., and Manning, C. D. (2014). Combining distant and partial supervision for relation extraction. In *EMNLP*, pages 1556–1567.

[Augenstein et al., 2017] Augenstein, I., Das, M., Riedel, S., Vikraman, L., and McCallum, A. (2017). Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications. *arXiv preprint arXiv:1704.02853*.

[Bach et al., 2015] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

[Banko et al., 2007] Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open information extraction from the web. In *IJCAI*, volume 7, pages 2670–2676.

[Binder et al., 2016] Binder, A., Bach, S., Montavon, G., Müller, K.-R., and Samek, W. (2016). Layer-wise relevance propagation for deep neural network architectures. In *Information Science and Applications (ICISA) 2016*, pages 913–922. Springer.

[Craven et al., 1999] Craven, M., Kumlien, J., et al. (1999). Constructing biological knowledge bases by extracting information from text sources. In *ISMB*, volume 1999, pages 77–86.

[Culotta and Sorensen, 2004] Culotta, A. and Sorensen, J. (2004). Dependency tree kernels for relation extraction. In *Proc. ACL'04*, page 423. Association for Computational Linguistics.

[Dietterich et al., 1997] Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1):31–71.

[dos Santos et al., 2015] dos Santos, C. N., Xiang, B., and Zhou, B. (2015). Classifying relations by ranking with convolutional neural networks. *CoRR*, abs/1504.06580.

[Herrero-Zazo et al., 2013] Herrero-Zazo, M., Segura-Bedmar, I., Martínez, P., and Declerck, T. (2013). The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920.

[Kim et al., 2016] Kim, B., Khanna, R., and Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems*, pages 2280–2288.

[Lee et al., 2017] Lee, J. Y., Dernoncourt, F., and Szolovits, P. (2017). Mit at semeval-2017 task 10: Relation extraction with convolutional neural networks. *arXiv preprint arXiv:1704.01523*.

[Lei et al., 2016] Lei, T., Barzilay, R., and Jaakkola, T. (2016). Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*.

[Mintz et al., 2009] Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proc. ACL'09*, ACL '09, pages 1003–1011, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Nguyen and Grishman, 2015] Nguyen, T. and Grishman, R. (2015). *Relation extraction: Perspective from convolutional neural networks*.

[Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proc. KDD 2016*, pages 1135–1144. ACM.

[Riedel et al., 2013] Riedel, S., Yao, L., McCallum, A., and Marlin, B. M. (2013). Relation extraction with matrix factorization and universal schemas.

[Thrun, 1995] Thrun, S. (1995). Extracting rules from artificial neural networks with distributed representations. In *Proc. NIPS'95*, pages 505–512.

[Vrandečić and Krötzsch, 2014] Vrandečić, D. and Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

[Zeng et al., 2015] Zeng, D., Liu, K., Chen, Y., and Zhao, J. (2015). Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proc. EMNLP 2015, Lisbon, Portugal*, pages 17–21.

[Zeng et al., 2014] Zeng, D., Liu, K., Lai, S., Zhou, G., Zhao, J., et al. (2014). Relation classification via convolutional deep neural network. In *COLING*, pages 2335–2344.

[Zhang and Wang, 2015] Zhang, D. and Wang, D. (2015). Relation classification via recurrent neural network. *CoRR*, abs/1508.01006.

[Zhang et al., 2006] Zhang, M., Zhang, J., Su, J., and Zhou, G. (2006). A composite kernel to extract relations between entities with both flat and structured features. In *Proc. ACL'06*, pages 825–832. Association for Computational Linguistics.