Active Learning for Interactive Relation Extraction in a French Newspaper's Articles

Cyrielle Mallart[1, 2] **Michel Le Nouy**[1] **Guillaume Gravier**[3] **Pascale Sébillot**[2]
cyrielle.mallart@irisa.fr, michel.lenouy@ouest-france.fr, guig@irisa.fr, pascale.sebillot@irisa.fr
(1) SIPA Ouest-France, 10 rue du Breil, 35000 Rennes, France
(2) INSA Rennes, IRISA, Campus de Beaulieu, 35042 Rennes, France
(3) CNRS, IRISA, Campus de Beaulieu, 35042 Rennes, France

Abstract

Relation extraction is a subtask of natural language processing that has seen many improvements in recent years, with the advent of complex pre-trained architectures. Many of these state-of-the-art approaches are tested against benchmarks with labelled sentences containing tagged entities, and require important pre-training and fine-tuning on task-specific data. However, in a real use-case scenario such as in a newspaper company mostly dedicated to local information, relations are of varied, highly specific type, with virtually no annotated data for such relations, and many entities co-occur in a sentence without being related. We question the use of supervised state-of-the-art models in such a context, where resources such as time, computing power and human annotators are limited. To adapt to these constraints, we experiment with an active-learning based relation extraction pipeline, consisting of a binary LSTM-based lightweight model for detecting the relations that do exist, and a state-of-the-art model for relation classification. We compare several choices for classification models in this scenario, from basic word embedding averaging, to graph neural networks and Bert-based ones, as well as several active learning acquisition strategies, in order to find the most cost-efficient yet accurate approach in our French largest daily newspaper company's use case.

1 Motivation

Relation extraction is a mature field of natural language processing that aims at finding the relations between identified entities in texts. Recent research has focused on the use of trainable models to automatically extract and classify relations between entities. The state-of-the-art research on relation extraction focuses on large, complex models, that require either a long time to train or some pre-training with a fine-tuning phase. Task-specific labelled data is needed to train the final classification model, with research relying on benchmarks such as Zhang et al. (2017) or Hendrickx et al. (2010), which are often made of single sentences with clearly tagged entities and definitive, non-ambiguous relations.

In a real-life scenario, however, relation extraction is used in a language processing pipeline, e.g., in order to confront different reports of a same event, or to grasp a general picture of a situation. In the specific case of a regional newspaper company such as ours, the intent is to create a knowledge graph from the content of the newspaper's own articles, so as to facilitate journalists' investigation with an easy access to information and to possible relations that are otherwise drowned in pages of text. In this realistic scenario, data is of a very different nature from standard corpora, exhibiting certain features that are specific to the regional ecosystem, thus challenging off-the-shelf models or learning methods.

The main feature of this data is that, while it is abundant, almost none of it is annotated, as human expert annotation is expensive. We thus turn to active learning (AL) as a means to alleviate the cost of labelling datasets. This approach, opposed to simply annotating samples, allows for a selection of the most helpful training samples, and therefore also allows a reduction of the number of annotations that have to be made by a human to reach satisfactory performance. Besides, it also allows annotation to be done in several installments of comfortable length for the annotator, reducing their fatigue and potential labelling errors.

Another issue than the scarcity of labelled data constrains the models that can be used within this active learning paradox, due to the local nature of the articles. Entities being often specific to the local context, making use of external data is not an option, also because of ownership and trust issues. Furthermore, relation types vary greatly, from one journalist's interests to the other, requiring the

Proceedings of Recent Advances in Natural Language Processing, pages 886–894
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_101

ability to quickly rebuild models from very limited data. On top of these general concerns, in articles, a large number of entities mentioned co-occur but are not actually related. To eliminate such fallacious cases is not straightforward, due to the complex language style and numerous entities mentions within articles.

We therefore develop an active learning approach to relation extraction in the newspapers' articles to deal with the scarcity and cost of labelled data. With the underlying idea that two light models are more likely to be accurately trained with a limited amount of data in an active learning scenario than a single large end-to-end model, we separate the task of detecting a relation from the classification of said relation. A first LSTM-based model specializes in detecting the fallacious candidates by outputting whether two entity mentions within a sentence are related or not, letting a second classification model focus on subtle differences between relations instead of having to clean the data at the same time. We aim to find, in this active learning on newspaper context and with the data particularities outlined above, whether a complex state-of-the-art classification model is relevant, or whether a shallow approach is better suited. Our goal is also to find an active learning strategy that reaches satisfactory results the fastest, so as to reduce human annotation.

In the following, we describe the architecture of the whole system and detail the different models used. A first experiment compares the performances of three classification models as data becomes available, therefore checking the amount of data at which a more complex model is better suited than a simpler, lighter model. In a second experiment, we compare three active learning scenarios with a fixed pipeline consisting of the LSTM-based relation detection model and a C-GCN classification model. We aim at finding a cost-effective active learning within our framework so as to minimize the amount of annotations needed. We therefore contribute to a study of several active learning scenarios to fit our newspaper use-case with specific, unbalanced local data, using machine learning models with different levels of depth. We notably study the relevance of very deep learning models in such a data-scarce scenario.

2 Related Works

All the outlined particularities of the data in our regional newspaper articles render most of the state-of-the-art work in the field of relation extraction inapplicable directly to this scenario, although relation extraction has excellent state-of-the-art models. For instance, even if Open Information Extraction (Banko et al., 2007; Mesquita et al., 2013; Del Corro and Gemulla, 2013) revolving around the identification then extraction of potential entities (noun groups) and potential relations between them, is a breakthrough in treating large-scale corpora, this method relies on the extraction of nominal groups, which does not link to entities of the newspaper's knowledge graph, and may lead to the identification of some relations that make sense grammatically, but have no meaning semantically.

In unsupervised relation extraction (Hasegawa et al., 2004; Takase et al., 2015), most works cluster linguistic patterns if two given entities co-occur a sufficient number of times. Used directly, these clusters' usability is limited, as they have to be studied and labelled by hand. Preemptive Information Extraction (Rosenfeld and Feldman, 2007; Shinyama and Sekine, 2006) uses such clusters of candidate relations as high-precision seeds that feed a second, semi-supervised model. Those methods require numerous documents with redundant entities pairs and linguistic patterns, otherwise seeds might be corrupted and the semi-supervised model experiences semantic drift. Finally, state-of-the-art supervised models range from learning on syntax trees and hand-crafted features based on dependency parsing (Zelenko et al., 2003; Kambhatla, 2004; Xu et al., 2015; Liu et al., 2016; Cai et al., 2016) to deep learning (Wang et al., 2016; Lin et al., 2016). Most of the latest approaches are building on the computationally heavy and data-intensive transformer model (Devlin et al., 2019), such as Yamada et al. (2020); Baldini Soares et al. (2019); Wu and He (2019). All the aforementioned approaches thus fall within one of two categories: either they are heuristic, count-based approaches, that do not work on our very local data as journalists tend to avoid redundancy in their writing; or they are learning-based, which requires large amount of labeled data or pre-existing knowledge which we do not have.

The active learning paradigm is a way to address the shortage of annotated data as well as the evolution of the needs, such as the addition of new

relation types, by selecting most helpful samples to train the model: the main goal is to find the best strategy for the selection of said samples. In active learning, a small number of samples is chosen, labelled by an oracle and used to optimally train the model, the process repeating until a stopping criterion is met.Schröder and Niekler (2020) point out the conflicting paradigms of deep learning and active learning: deep neural networks excel, but under the strict requirement that abundant data be available, which defeats active learning's frugality objective. Uncertainty sampling has been shown to be adaptable to deep classifiers in Prabhu et al. (2019) for NLP, and we select this approach, as it only relies on the distribution of probabilities of the sample once it has run through the model. Therefore, there is little additional cost to creating an uncertainty-based sample to label, and many models can be used in these framework, as long as they output such probability distribution. For deep learning, Sener and Savarese (2017) or Siméoni et al. (2021) report no improvement of uncertainty sampling in image classification scenario, while Siddhant and Lipton (2018) find that both uncertainty-based sampling and Bayesian approaches outperform random across 3 NLP tasks. Our aim is to find how these results transfer on real-life data, and whether deep learning is truly an improvement over shallower approaches in this context.

3 Methodology and Protocol

We present the global architecture of the active learning relation detection system that we study before providing details on the different classifiers.

3.1 Architecture of the System

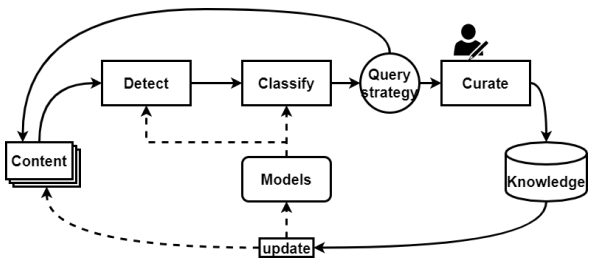


Figure 1: System architecture diagram

In the following, several working hypotheses have been made to reduce noise on our particular data. We justify them, and present the specifics of the use-case, as well as the structure of the iterative system that we implemented.

Firstly, only the relations explained within a sentence are considered as potential relations. Furthermore, in our data, sentences are considered to be independent. This assumption comes from considering only the couples of entities that appear in the same sentence to reduce fallacious co-occurrences and avoid error propagation when using coreference resolution systems. This assumption can be challenged, and we leave binary relation detection across sentence boundaries to be treated in future work.

The proposed architecture is presented in Figure 1. This iterative system is a pool-based active learning architecture, revolving around an expert oracle and a learner, described in Sec. 3.2. The pool of unlabeled data consists of samples, each containing the sentence s, the surface form of the entities e_1 and e_2 as well as their positions in the sentence and their types. On top of this, each sample goes through NLP pre-treatment and therefore contains the dependency parse of the sentence, the part-of-speech and NER tag of words in the sentence as well as the part-of-speech, NER tags and dependency parse tags along the shortest dependency path between the two entities. During one iteration, our learner predicts relations R from the 13 identified relation classes for the unlabelled data in the pool of content. Some examples from this pool are selected according to a query strategy based on the prediction probability. These chosen samples are presented to a human annotator, or oracle, who annotates them, and these annotated samples join a pool of labeled data. This new knowledge is used to train the learner, that in the next iteration predicts the relation class of the rest of the data in the unlabeled pool.

In this work, we use a pool-based approach, with a selected sample of articles on the same subject, namely the local enterprise landscape, so as to stay within one topic of interest and to be able to compare methodologies. This is also in phase with the scenario where a journalist starts exploring a specific topic from selected content. Stream-based solutions to adapt to the constant stream of information created by journalists every day are not considered here.

Our query strategy revolves around uncertainty-based sampling, where examples that the model is least certain about are selected and presented to the oracle for correction. Here, we have three propositions for the choice of query strategy.

- Random: take k samples randomly from the entire pool of annotations not used for training yet.

- Least likely: take the k samples less likely in their prediction, i.e., that have the lowest prediction probability.

- Mixture: take, for each of the l predicted relations, the sample with the lowest probability to belong to this class, plus $k - l$ of the most likely samples overall. This allows to control at the same time the border cases where the model does not distinguish very well, and to catch cases where the model is very confident of a wrong relation.

3.2 Models Description

As outlined above, active learning requires a learner. Our proposed learner consists of two models. A first detection model specializes in verifying that given a sentence and two entities, the sentence actually expresses some relation between the two given entities. Then, a classification model is applied on the samples that are predicted as being related so as to predict the type of relation. We tested three classification models: C-GCN, a basic model based on word embedding averages and a BERT-based model, described in the following.

3.2.1 Detection Model

Figure 2 presents the overall architecture of our detection network applied to an example sentence. In each sentence, the couples of entities are extracted and the model learns whether the words that are on the shortest dependency path between the two entities of a couple depict a relation or not. Only information about the type of entities and the word features along the shortest dependency path between the two entities are used in the model. The model features two branches, one modeling the entities through their types with a fully-connected layer (left branch), the other the syntax information between the two entities with a LSTM model (right branch). LSTM networks' ability to deal with sequential data to retain or forget relevant information makes us confident that this architecture, despite being simple, still can retain enough information to correctly detect relations. The two branches are finally merged with a fully connected layer with sigmoid activation to predict the probability of a relation between the two entities.

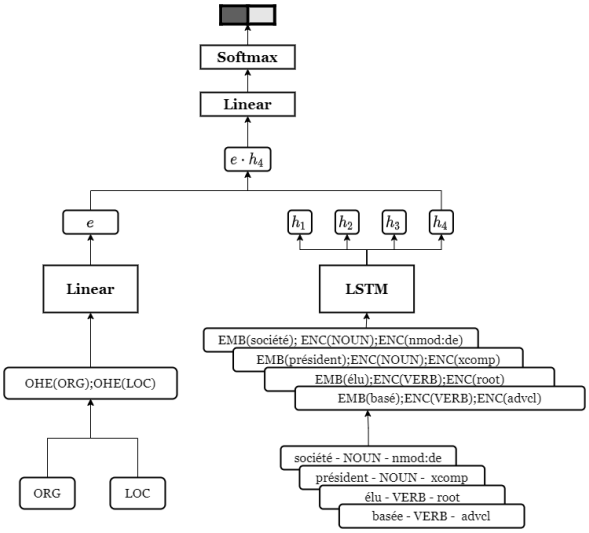


Figure 2: Binary detection model, applied to the sentence "Le président du département Olivier Richefou a été élu président de la société publique locale Espace Mayenne, basée près du 42e RT à Laval."

The use of the shortest dependency path between the entities, along with word and part-of-speech features of the words along the path, accounts for the syntax that can relate entities. All these features were extracted with the StanfordNLP library (Manning et al., 2014). The word features include notably the embedding of each word as obtained from a pre-trained skip-gram model (Mikolov et al., 2013) obtained from Fauconnier who made the embeddings publicly available[1]. The path excludes the entities themselves, which are accessible on the first branch via their respective types.

3.2.2 GCN Classification Model

C-GCN + PA-LSTM (Zhang et al., 2018) was chosen as one of the classification model, as it does not necessitate any external knowledge base embedding or heavy transformer machinery, while still retaining performance close to the most recent models.

This relation extraction solution's idea is similar to many representation-based models, where a first part of the model to create contextualised embeddings of the words in the sentence, and a second part to output the predictions of the relation class from the embeddings. The aspect that sets aside the C-GCN model is the use of graphical neural networks over dependency parse trees to find contextualized vector representations of the tokens, which is essentially the computation of a few ma-

[1] http://fauconnier.github.io/#data

trix multiplications (2 times the number of GCN layers chosen), and therefore very easily distributed and fast to train.[2].

3.2.3 Base Classification Model

We replace the C-GCN in a second installment of our system, so as to verify whether the C-GCN model is adapted to the task of classifying existing relations, or too complex for the small amount of samples acquired via active learning. The replacement model takes as input the average word embedding, by averaging every word vector obtained for the words of the sentence. These word embeddings are the same as in our detection model. This input is fed to a fully connected layer, with a softmax output. This simplistic classification model will not be able to perform relation classification efficiently, but it acts as a baseline to evaluate the results of the other state-of-the-art models.

3.2.4 Bert-based Classification Model

Inspired from Alt et al. (2019) and Shi and Lin (2019), we also implement an approach based on a pre-trained BERT architecture for French, FlauBert (Le et al., 2020).

First, we construct the input sequence as *[[CLS] sentence [SEP] entity1 [SEP] entity2 [SEP]]*. To avoid over-fitting, the tokens of the input sentence corresponding to the entities are replaced by a special token representing the type of the entity ([PER], [LOC], [ORG] or [MISC]). Contrary to Alt et al. (2019), where the entities are placed before the masked sentence to bias the attention mechanism towards the representation of entities, we place ours at the end. Our very specific entities, such as original or new company names, might not be well represented with a pre-trained architecture, and we therefore put more emphasis on the known type of entities than on the name of the entities themselves. The input sequence is fed to the pre-trained Flaubert model. This model encodes the input representations over successive transformer blocks. Each of these transformer blocks is made of a masked multi-head attention layer followed by a position-aware feed-forward layer.

We thus obtain the final state representation h_L of the input sequence. The last state h_L^k, which represents a summary of the input sequence, is used to compute the probability distribution over all relation classes, by running it through a linear layer activated with ReLU function, and a last linear layer followed by a softmax layer.

[2]We used the code directly available from the authors, at https://github.com/qipeng/gcn-over-pruned-trees

4 Experiments

Two experiments, one comparing three relation classification models, and one comparing the three different uncertainty-based active learning strategies, aim at finding the best setting for relation extraction, taking into account cost limitations.

4.1 Data

To initialize and train our models, annotators from the company have created three datasets. Two small pools of data consist of 588 and 261 samples, respectively the seed and the testing set, with relations in various proportions, so as to reflect the reality of the contents of the newspaper.

For active learning purposes, we gathered another 1,271 annotations for the 13 categories, with a distribution shown in Tab. 1. As expected, most of the relation candidates do not actually depict a relation. The relations "créé en", "né à" and "contracté par" are not likely to be well predicted by any model, as they each are represented by at most 4 annotations. Besides, we added an *autre* (other) category. This allows firstly to avoid labelling samples that do depict a relation, albeit an unknown one, as fallacious, therefore reducing the noise for the detection model, and secondly, annotators can flag relations that we may not have identified earlier, giving us a path for future improvement of the system. Samples were annotated by only one person at a time, in an effort to acquire as many labelled samples as possible from a limited amount for annotators. This might lead to some bias, but considering that our annotators are experts, that will be on par with the expected use of the system: there will not be enough resources to have several journalists cross-validate potential samples, so once one sample is annotated by an expert in the field, it is considered properly labelled.

4.2 Comparison of Classification Models

This first experiment compares the three classification models: C-GCN, the base model (BASE) and the BERT-based approach (FlauBERT+FC). All seed data has been used to initialize the models. Classification models classify only on the 13 identified types of relation, without having to predict the class "aucune" (None), as the detection model already takes care of this class. The acquisition

Relation	Frequency in the data
aucune *(None)*	60,90%
dirige *(is the head of)*	6,22%
a_son_siège_à *(has its headquarters located in)*	5,90%
collègue_de *(is a colleague of/works with)*	5,51%
autre (other)	5,19%
vit_à *(lives in)*	4,56%
sous_lieu_de *(is a geographical subdivision of)*	3,86%
membre_de *(member of)*	2,75%
a_créé *(is the creator of)*	2,52%
précède *(is the predecessor of)*	0,87%
filiale_de *(is a subsidiary of)*	0,87%
créé.e_en *(created in)*	0,31%
né.e_à *(born in)*	0,31%
contracté.e_par *(has a contract with)*	0,24%

Table 1: Distribution of the labels of samples in the active learning pool

strategy is set to "random", with 50 annotations selected for each iteration, until either a criterion of a difference of micro-F1 score inferior to 0.001 is reached, or 60 iterations have been completed, which amounts to almost the entire training pool. Figure 3 shows the evolution of F1 score as a function of the iterations.

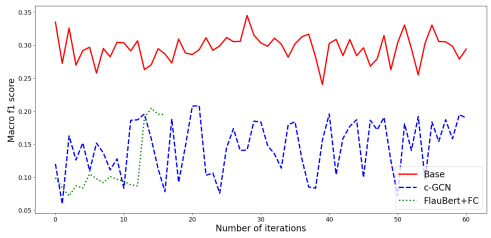


Figure 3: Macro F1 score for models BASE, C-GCN and FlauBERT+FC, along the active learning steps with a random query strategy.

Globally, the scores for all models only show little improvement with the number of iterations, which means that adding annotated data yields marginally better results than the simple starting seed. Upon analysis, the main culprit can be found in the use of a "random" sampling strategy on highly imbalanced data, which chooses imbalanced training samples to the model.

Firstly, most of the active learning samples display no relation. This leads to the detection model becoming very conservative, and discarding many samples as fallacious. The best achieved precision across training is 0.42, meaning that 42 % of samples labelled as "aucune" really are fallacious, and therefore 58 % of all samples labelled as fallacious actually depict a relation. As our model is a pipeline, it accumulates the errors: samples belonging to small classes, mistakenly classified as fallacious, do not make it to the classification model.

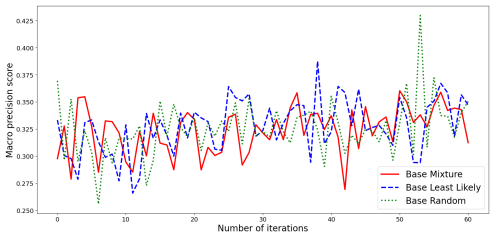
Secondly, two classes ("dirige" and "a son siège à") are disproportionately large, which leads to a phenomena of "concentration" on those two big classes, to varying degrees depending on the model. On the one hand, the C-GCN and FlauBERT+FC models directly classifies all data in one of those two classes, "dirige", and still reaches a satisfying loss, therefore never learning any of the features on the smaller classes. On the other hand, as learning progresses, the lighter model BASE progressively improves on the samples corresponding to the smallest classes, at the expense of the middle classes, with only a slight deterioration for the bigger classes. Results fluctuate largely due to the nature of the test dataset, which is small and therefore, one misclassification may lead to a large change in scores.

The take-home result is therefore that the state-of-the-art methods do not apply in a straightforward manner on this newspaper's data. The model that seems to adapt best to this imbalance of data is the simplest one, based on a vector representation of the sentence, as the GCN-based model completely misclassifies half of the dataset, and the Flaubert-based model does not have enough data to train its millions of weights properly.

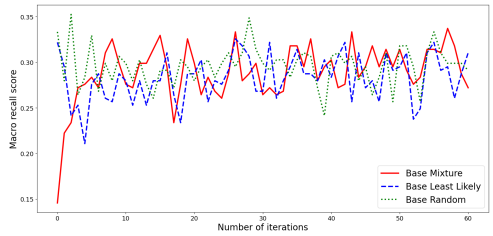
4.3 Active Learning Query Strategy

The first experiment used the "random" active learning strategy, to be fair for all models. We have shown that, with the over-representation of some classes in our data, this leads to bad performances for all models. To verify if the active learning acquisition strategy could help smooth this phenomena, we try three different active learning strategies with our BASE model and the C-GCN model. The three tested strategies are respectively random, least likely and mixture, the results for BASE being plotted in Figure 4. We do not plot the results for C-GCN as they do not solve the issue of all samples being predicted as "dirige", although we note a slight improvement on the recall of the "aucune" class for the mixture strategy.

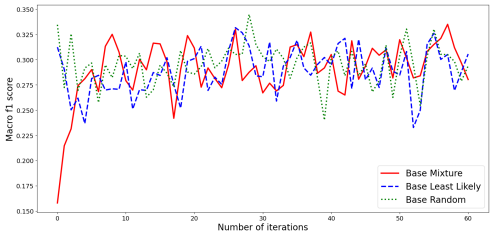
On the BASE model, the least likely strategy does not improve the general macro F1 score over



(a) Precision



(b) Recall



(c) F1

Figure 4: Macro scores for active learning strategies Random, Least likely and Mixture, along the active learning steps for a BASE classification model.

the random strategy. However, the scores within classes change: while the big classes are equally-well predicted under the two strategies, under the least likely strategy, the smallest classes (such as "filiale de" or "précède") get less well predicted, contrary to the more populated classes (such as "membre de" or "vit à").

The mixture strategy, on the other hand, improves both precision and recall on the detection task alone, which can be attributed to a different distribution of true and fallacious samples fed to the detection model for training: 60% of samples being fallacious under the random strategy drops to 40% under the mixture strategy. By allowing the model to train over more examples of actual relations, it does not discard rare relations as fast.

The mixture strategy, however, shows no improvement over random in terms of F1 score for the classification. Nonetheless, results are more consistent across classes, with slightly worse precision and recall on large classes but better results on the classes with few training samples. The result on the middle-sized classes surprisingly does not improve, when it was expected that a flatter distribution would lead to improvement on all classes but the largest ones. An explanation may be that, upon seeing a larger diversity of samples, the simplistic baseline model reaches its limits and cannot differentiate between samples containing similar vocabulary.

5 Conclusions

In this work, we reported on an experiment to develop an automatic relation extraction system adapted to our newspaper data in a context of scarcity of labelling, exploring various strategies towards the best possible setting. We got confronted to the divergence of objectives between deep learning and active learning, showing that shallower approaches are a safer bet in a context where labelled data is acquired via active learning. Although such simple models do reach their limits early, and deep learning consistently breaks records in the literature, deep architectures require more data to train than can be supplied through active learning with a real human as an oracle. We will therefore steer our future work in the direction of shallow classifiers with a small amount of weights to train. Given the small amount of data available, all linguistic information might need to be used, such as part-of-speech tags, semantic roles or dependency tags, not solely relying on contextualized word vector representation that either takes time to train or has to be obtained via an external source. Additionally, adopting a hand-made query strategy relying on sampling from each predicted relation to reduce the cost of annotation was found to be a small improvement compared to using only a small training set. Still, the increase in performance needs to be checked in the light of a classification model better suited for active learning. Besides, in order to be deployed, our system still needs to be able to take into account new types of relations. While we already store these "autre" relations during the annotation phase, we are yet to incorporate them into the classification model, in order to create a system that evolves with the content of the newspaper.

References

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Improving Relation Extraction by Pre-trained

Language Representations. In *Automated Knowledge Base Construction (AKBC)*. https://openreview.net/forum?id=BJgrxbqp67

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the Blanks: Distributional Similarity for Relation Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2895–2905. https://doi.org/10.18653/v1/P19-1279

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open Information Extraction from the Web. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*. Morgan Kaufmann Publishers Inc., 2670–2676. http://dl.acm.org/citation.cfm?id=1625275.1625705

Rui Cai, Xiaodong Zhang, and Houfeng Wang. 2016. Bidirectional Recurrent Convolutional Neural Network for Relation Classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1-Long Papers)*. Association for Computational Linguistics, 756–765. https://doi.org/10.18653/v1/P16-1072

Luciano Del Corro and Rainer Gemulla. 2013. ClausIE: Clause-based Open Information Extraction. In *Proceedings of the 22nd International Conference on World Wide Web*. Association for Computing Machinery, 355–366. https://doi.org/10.1145/2488388.2488420

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1-Long and Short Papers)*. Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Takaaki Hasegawa, Satoshi Sekine, and Ralph Grishman. 2004. Discovering Relations Among Named Entities from Large Corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Article 415. https://doi.org/10.3115/1218955.1219008

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 33–38. https://www.aclweb.org/anthology/S10-1006

Nanda Kambhatla. 2004. Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Extracting Relations. In *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics, Article 22. https://doi.org/10.3115/1219044.1219066

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised Language Model Pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 2479–2490. https://www.aclweb.org/anthology/2020.lrec-1.302

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. 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)*. Association for Computational Linguistics, 2124–2133. https://doi.org/10.18653/v1/P16-1200

Yang Liu, Sujian Li, Furu Wei, and Heng Ji. 2016. Relation Classification via Modeling Augmented Dependency Paths. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 24, 9 (Sept. 2016), 1585–1594. https://doi.org/10.1109/TASLP.2016.2573050

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, 55–60. http://www.aclweb.org/anthology/P/P14/P14-5010

Filipe Mesquita, Jordan Schmidek, and Denilson Barbosa. 2013. Effectiveness and Efficiency of Open Relation Extraction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 447–457. https://www.aclweb.org/anthology/D13-1043

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the 1st International Conference on Learning Representations, Workshop Track Proceedings*. http://arxiv.org/abs/1301.3781

Ameya Prabhu, Charles Dognin, and Maneesh Singh. 2019. Sampling Bias in Deep Active Classification: An Empirical Study. *CoRR* abs/1909.09389 (2019). arXiv:1909.09389 http://arxiv.org/abs/1909.09389

Benjamin Rosenfeld and Ronen Feldman. 2007. Clustering for Unsupervised Relation Identification. In *Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management*. Association for Computer Machinery, 411–418. https://doi.org/10.1145/1321440.1321499

Christopher Schröder and Andreas Niekler. 2020. A Survey of Active Learning for Text Classification using Deep Neural Networks. *CoRR* abs/2008.07267 (2020). arXiv:2008.07267 https://arxiv.org/abs/2008.07267

Ozan Sener and Silvio Savarese. 2017. Active Learning for Convolutional Neural Networks: A Core-Set Approach. *arXiv e-prints*, Article arXiv:1708.00489 (Aug. 2017), arXiv:1708.00489 pages. arXiv:1708.00489 [stat.ML]

Peng Shi and Jimmy Lin. 2019. Simple BERT Models for Relation Extraction and Semantic Role Labeling. *ArXiv* abs/1904.05255 (2019).

Yusuke Shinyama and Satoshi Sekine. 2006. Preemptive Information Extraction Using Unrestricted Relation Discovery. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics, 304–311. https://doi.org/10.3115/1220835.1220874

Aditya Siddhant and Zachary C. Lipton. 2018. Deep Bayesian Active Learning for Natural Language Processing: Results of a Large-Scale Empirical Study. *CoRR* abs/1808.05697 (2018). arXiv:1808.05697 http://arxiv.org/abs/1808.05697

Oriane Siméoni, Mateusz Budnik, Yannis Avrithis, and Guillaume Gravier. 2021. Rethinking deep active learning: Using unlabeled data at model training. In *2020 25th International Conference on Pattern Recognition (ICPR)*. 1220–1227. https://doi.org/10.1109/ICPR48806.2021.9412716

Sho Takase, Naoaki Okazaki, and Kentaro Inui. 2015. Fast and Large-scale Unsupervised Relation Extraction. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*. 96–105. http://aclweb.org/anthology/Y15-1012

Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. 2016. Relation Classification via Multi-Level Attention CNNs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1-Long Papers)*. Association for Computational Linguistics, 1298–1307. https://doi.org/10.18653/v1/P16-1123

Shanchan Wu and Yifan He. 2019. Enriching Pretrained Language Model with Entity Information for Relation Classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. ACM, New York, NY, USA, 2361–2364. https://doi.org/10.1145/3357384.3358119

Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. 2015. Classifying Relations via Long Short Term Memory Networks along Shortest Dependency Paths. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1785–1794. https://doi.org/10.18653/v1/D15-1206

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6442–6454. https://doi.org/10.18653/v1/2020.emnlp-main.523

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel Methods for Relation Extraction. *Journal of Machine Learning Research* 3 (2003), 1083–1106. https://doi.org/10.3115/1118693.1118703

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph Convolution over Pruned Dependency Trees Improves Relation Extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2205–2215. https://doi.org/10.18653/v1/D18-1244

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware Attention and Supervised Data Improve Slot Filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 35–45. https://nlp.stanford.edu/pubs/zhang2017tacred.pdf