Confronting Active Learning for Relation Extraction to a Real-life Scenario on French Newspaper Data

Cyrielle Mallart(1, 3) **Michel Le Nouy**(1) **Guillaume Gravier**(2) **Pascale Sébillot**(3)
cyrielle.mallart@irisa.fr, michel.lenouy@ouest-france.fr, guig@irisa.fr, pascale.sebillot@irisa.fr
(1) SIPA Ouest-France
(2) Univ Rennes, CNRS, Inria - IRISA, France
(3) Univ Rennes, CNRS, Inria, INSA Rennes - IRISA, France

Abstract

With recent deep learning advances in natural language processing, tasks such as relation extraction have been solved on benchmark data with near-perfect accuracy. However, in a realistic scenario, such as in a French newspaper company mostly dedicated to local information, relations are of varied, highly specific types, with virtually no data annotated for 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 model for detecting the relations that do exist, and a state-of-the-art model for relation classification. We compare several classification models of different depths, from simplistic word embedding averaging, to graph neural networks and Bert-based models, as well as several active learning query strategies, including a proposal for a balanced uncertainty-based strategy, in order to find the most cost-efficient yet accurate approach in our newspaper company's use case. Our findings highlight the unsuitability of deep models in this data-scarce scenario, as well as the need to further develop data-driven active learning strategies.

Relation extraction is a mature field of natural language processing (NLP) that aims at finding the relations between entities in a text. Due to the power of recent language models, and with the use of annotated benchmark datasets, the latest research on relation extraction has focused on the classification of the relations expressed in a sentence. However, issues arise when trying to adapt these powerful models to real-life data, such as extracting relations from content from a local newspaper. In this real-life scenario, data is of a nature very different from standard benchmark corpora, exhibiting certain features that are specific to the regional ecosystem, thus challenging off-the-shelf models. The main feature of this data is that, while it is abundant and new content is added daily, almost none of it is annotated, as human expert annotation is expensive and journalists cannot devote much time to annotating their articles. Other issues of the annotated data, beyond its scarcity, include the fact that the entities present are very specific to the local context of the newspaper, that is, entities that are mentioned only once, or that do not appear in other sources, as they belong to the social and economic fabric of the area covered by the newspaper. Additionally, many entities co-occur in the same sentence but are not related, and some types of relations appear more often in the content than others. These factors, added to ownership of journalistic sources and trust issues, makes using external data impossible. Similarly, eliminating cases of meaningless co-occurring entities pairs, or fallacious relations, is not straightforward, due to the complex language style and numerous entities mentions within articles. Indeed, our dataset contains on average 5 identified entities per sentence, with proper nouns making up 24% of all words in the content. The complexity of language varies between samples as well, as sentences lengths range from 3 to 112, with an average of 25 words per sentence, and the depth of the dependency tree of a sentence varies between 3 and 30,

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

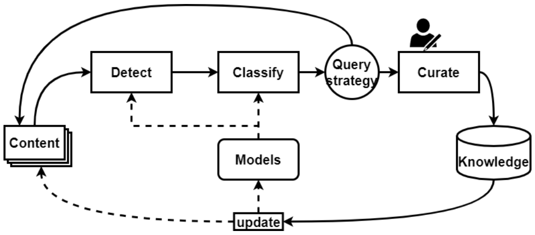

Figure 1: System architecture diagram

the average being 6. These numbers highlight the challenge of applying state-of-the-art relation extraction solutions onto data whose complexity varies largely across samples.

Previous works have been able to make use of unlabelled information, such as unsupervised relation extraction (Hasegawa et al., 2004; Takase et al., 2015), that clusters 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) instead uses such clusters of candidate relations as high-precision seeds to feed a second, semi-supervised model, but this requires redundant linguistic patterns or entity pairs to avoid semantic drift, which are unavailable here. On the supervised end of the learning methods, models range from 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 deep learning approaches, built on the transformer model (Devlin et al., 2019), such as (Yamada et al., 2020; Soares et al., 2019; Wu and He, 2019), and thus require large amount of labelled data or preexisting knowledge which we do not have. We develop an active learning approach to relation extraction in the newspapers' articles to deal with the scarcity and cost of labelled data, while still leveraging the deep state-of-the-art models in relation extraction and classification. 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 output by a model, with little additional cost to adding such a strategy to a probabilistic model. For deep learning, (Sener and Savarese, 2018) or (Siméoni et al., 2021) report no improvement by using uncertainty sampling in image classification tasks, while (Siddhant and Lipton, 2018) find that uncertainty-based sampling outperforms random across three NLP tasks. Our aim is to find how these results transfer to real-life data, and whether deep learning is truly adaptable in our context, or whether shallower systems are more adapted.

In the following, we describe the active learning system we devised to test the compatibility of deep learning and active learning for relation extraction in real life. In our experiment, we compare three active learning scenarios and three different classification models. We aim at finding a cost-effective active learning strategy 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.

1 Methodology

The proposed architecture of our active learning system 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 Section 1.2. Our pool of content consists of sentences extracted from articles of the newspaper. We consider each pair of entities appearing together in a sentence as a candidate relation, and a sample thus consists of a sentence with two highlighted entities between which there may be a relation. Every sample gathered in this active learning experiment has been seen and labelled by a human expert only once. Our annotators being experts of the content used in the articles, we considered that they were trustworthy enough to not have to rely on collective annotation, which in turn allowed for more samples to be annotated by experts whose time is precious so as to create a small database of annotated content for experimentation. Our samples fall in one of 13 relation types, revolving around the life of local companies, such as *dirige* (is_director_of), *a_son_siège_à* (has_headquarters_located_in), or *collègue_de* (is_colleague_of) as well as a *aucune* (none) category. The resulting dataset is unbalanced with respect to the type of relations, as fallacious relations make up 60% of the samples, and classes *dirige* and *a_son_siège_à* concentrate 30% of the remaining samples.

1.1 Proposed active learning query strategies

Three propositions for the choice of query strategy are as follows, most of them revolving around uncertainty-based sampling:

- Random: take k samples randomly from the entire pool of samples 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: gather all unlabelled samples predicted as belonging to a given relation class, for each of the l possible types of relations. In each of these l relation class bins, select the sample predicted to be in said class with the lowest probability. This results in l samples with low probability to belong to their predicted class, to which we add $k - l$ of the samples with the highest probability overall. This allows to control at the same time the border cases where the model does not distinguish relations very well, and to catch cases where the model is very confident of a wrong relation.

The least likely strategy, or least confident sampling, is the simplest version of uncertainty-based selection strategies: should the simplest and fastest form of uncertainty sampling work, it is likely that methods of sampling more tailored to the data can also succeed. The mixture strategy is an attempt to compensate for the unbalancedness of the data. We expect this to be particularly important for the classes that contain very few examples, as this strategy should select the examples most likely to be wrongly classified, which should be the rarest ones.

1.2 Models

The proposed learner is a pipeline of two models. We separate the task of detecting a relation from the classification of said relation. Our underlying idea here is that two light models are faster to train, and more likely to be correctly trained with a limited amount of data than a single large end-to-end model, with the classification model focusing on the small semantic differences between existing relations rather than focusing on the syntax to decide whether a relation exists or not. The detection model takes the word embedding[1], the part-of-speech tag and dependency tag of each word along the shortest dependency path that connects the two entities in a given sentence. These features are concatenated, and fed to an LSTM layer. The types of the two entities are fed to a simple fully-connected layer. The two outputs, encoding respectively the syntax of the sentence and the entities types, are merged through a dot-product, before a last sigmoid-activated fully connected layer.

For the classification model, we compare three models: a simple model based on average word embedding, C-GCN (Zhang et al., 2018)[2], that is recent, does not require a transformer layer, and is adaptable to French, and a model based on a French flavour of BERT, FlauBERT (Le et al., 2020).

The very simple base classification model (named here BASE) takes as input the average word embedding of the sentence in a sample, averaging every word vector obtained for the words of the sentence. These word embeddings are obtained from a pre-trained skip-gram model (Mikolov et al., 2013) obtained from Fauconnier who made the embeddings publicly available[1]. This input is fed to a fully connected layer, with a softmax output.

C-GCN considers the dependency parse of a sentence as a graph, and applies a graph convolution layer to obtain contextualized embeddings for each word. The final representation is obtained by concatenating the pooled representation for each entity and for the entire sentence, which is followed by a linear layer with a softmax activation.

The approach based on a pre-trained FlauBERT architecture (Le et al., 2020) (named here FlauBERT+FC), is inspired from (Alt et al., 2019) and (Shi and Lin, 2019). 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), we place our

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

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

Relation	Number of test samples	Number of training samples	Random	Least likely	Mixture
aucune (*none*)	60	774	0.47	0.43	**0.48**
dirige (*is_director_of*)	66	79	0.35	0.26	**0.43**
a_son_siège_à (*has_headquarters_located_in*)	46	75	0.31	**0.35**	**0.42**
collègue_de (*is_colleague_of*)	4	70	0.00	0.00	0.00
vit_à (*lives_in*)	13	58	0.06	**0.11**	**0.07**
sous_lieu_de (*is_geographical_subdivision_of*)	9	49	0.00	**0.24**	**0.14**
membre_de (*is_member_of*)	25	35	0.20	0.16	0.15
a_créé (*is_creator_of*)	23	32	0.35	0.20	**0.42**
filiale_de (*is_subsidiary_company_of*)	12	11	0.35	0.29	0.33
précède (*is_predecessor_of*)	3	11	0.40	0.29	0.19

Table 1: F1 score for each relation class present in the test set for all three strategies and the BASE model, after last iteration (Bold font indicates improvement over random)

Model	Random			Least likely			Mixture		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
FlauBERT	0.13	0.26	**0.16**	0.10	0.21	0.13	0.14	0.27	**0.16**
C-GCN	0.24	0.22	0.16	0.19	0.21	0.14	0.14	0.29	**0.19**
BASE	0.36	0.33	0.32	0.34	0.29	0.29	0.39	0.37	**0.37**

Table 2: Weighted scores of multiclass classification for the different strategies, after last iteration (Bold font indicates the best F1 score)

entities at the end of the masked sentence, in order to put more emphasis on the types of entities than on their names, as our very specific entities might not be well represented with a general pre-trained architecture. We feed this input sequence to a transformer, and this final vector representation is run through a linear layer activated with ReLU function, and a linear layer followed by a softmax layer.

2 Experimental results

We trained all three models using each of the proposed active learning strategies. The data consists of 1,271 annotations split in 13 categories, as well as two pools of 588 and 261 samples, respectively the "seed" to initialize the models, and the test set. All models are trained as part of a pipeline, where the detection model is trained jointly with the classification model. Except for FlauBERT+FC where we initialized the model with the pre-trained FlauBERT weights, all models were trained from scratch, including the detection model. The three tested strategies are respectively random, least likely and mixture. At each iteration, the model is re-trained on the entire annotated pool plus 20 annotations selected through the query strategy, 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.

We observed that the scores for all models show little improvement with the number of iterations, which means that adding annotated data yields marginally better results than the starting seed. The main culprit can be found in the highly imbalanced data. Indeed, most of the active learning samples display no relation, which leads to the detection model becoming very conservative, and discarding many samples as fallacious. In addition, two classes (*dirige*, is_director_of, and *a_son_siège_à*, has_headquarters_located_in) are disproportionately large, as visible in Table 1, which leads to a phenomenon of "concentration" on those two big classes, to varying degrees depending on the model. Notably, the C-GCN and FlauBERT+FC models classify all data in one of those two classes, *dirige*, and still reach a satisfying loss, therefore never learning any of the features on the smaller classes. The major issue with these large models is that there is not enough data of any class to sufficiently train them, let alone on examples belonging to the smallest classes.

The final scores, available in Table 2, are obtained for each model and each query strategy after the last iteration, that is, after the last sample has been processed or the stopping criterion has been met. A first observation is that the mixture query strategy consistently improves the F1 score over the least likely strategy, and either improves or is similar to the random strategy. The least likely strategy consistently performs more poorly than others, as it forces the model to only learn on the hardest samples. The mixture strategy mitigates this effect by incorporating to the uncertain samples some of the best predicted samples, therefore encouraging the model in some of its best predictions and feeding it with easy examples of each relation.

The mixture strategy improves both precision and recall on the detection task alone compared to the random strategy, with, for the BASE model, precision increasing from of 0.29 to 0.38, and recall increasing from 0.15 to 0.16. This is due to the mixture strategy selecting 40% of fallacious relations, compared to random that selects 60%. By decreasing the number of fallacious samples fed to the detection model, it trains over more examples of actual relations, and does not discard rare relations as fast. This mixture strategy shows only little improvement over random for the classification task, with a change in the results according to class size: slightly worse precision and recall on very small classes but better results on both large classes and some middle-sized classes, as shown in Table 1. The scores for the middle-sized classes do not improve much, contrary to what was expected. This gives an insight on the workings of the mixture strategy, which eventually acts as a guardrail for middle-sized class examples: selecting the least well predicted samples in each class tends to often select the samples whose relation is more present in the dataset, and catch when they have been wrongly predicted as belonging to one of the small classes, thus improving the scores on the largest and middle-sized classes. Examples from smaller classes are not well represented, as the mixture strategy selects from the predicted classes only, and not the real classes. As rare examples are more likely to be classified in the wrong category, our model does not necessarily train on an instance of a small class at each iteration. A possible explanation for the worsening of the performance on some of the middle-sized classes is that the vocabulary and turns of phrases are similar for these relations, such as *membre_de* (is_member_of) and *dirige* (is_director_of), to the point where a simplistic BASE model reaches its limits, and the larger models cannot learn to distinguish on so little training examples.

3 Conclusion

This work studied the relevance of deep learning in the context of active learning, in an attempt to use this framework to decrease the cost of annotation in a newspaper company context. We thus compared models of various depths, and also compared uncertainty-based strategies. All things equal, larger models such as C-GCN or transformers do not outperform a much simpler one for this small dataset. Even for a high-resource language such as French, with deep models that are extensively pre-trained, the more complex approaches require too much data to fine-tune on this task, and simpler models are therefore better suited, at least in the first steps of such models, when the amount of labelled data is so limited. This falls in line with previous results, concluding that active learning and deep learning do not mix well in the task of relation detection and classification. A second finding is that, given the larger amount of data needed to train larger models, it is also impossible in our setting to draw definitive conclusions about the influence of the number of iterations or the active learning strategy. Still, our choice of a hand-crafted strategy marginally improves final results compared to a random or least likely strategy, hinting to the fact that purely uncertainty-based strategies are not a good fit for our data, but that future works involving real-life scenarios will benefit from an active learning strategy more tailored to the use-case. A possible direction to explore is diversity sampling, that encourages the selection of samples that reflect the complexity and diversity of real-life scenarios. Within this paradigm, the strategies most promising for our use-case are cluster-based sampling, that selects samples from groups created though unsupervised clustering, and representative sampling, where the training domain differs from the target domain, thus geared towards cross-domain adaptation.

References

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Improving Relation Extraction by Pre-trained Language Representations. In *Proceedings of the 2019 Automated Knowledge Base Construction*. https://openreview.net/forum?id=BJgrxbqp67

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)*. 756–765. https://doi.org/10.18653/v1/P16-1072

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)*. 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* (Barcelona, Spain). Article 415. https://doi.org/10.3115/1218955.1219008

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*. 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*. 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)*. 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 Transactions on Audio, Speech and Language Processing* 24, 9 (9 2016), 1585–1594. https://doi.org/10.1109/TASLP.2016.2573050

Tomás 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. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 4058–4068. https://doi.org/10.18653/v1/D19-1417

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

Ozan Sener and Silvio Savarese. 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *Proceedings of the 2018 International Conference on Learning Representations*. https://openreview.net/forum?id=H1aIuk-RW

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* (New York, New York). 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. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2904–2909. https://doi.org/10.18653/v1/D18-1318

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

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*. 2895–2905. https://doi.org/10.18653/v1/P19-1279

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)*. 1298–1307. https://doi.org/10.18653/v1/P16-1123

Shanchan Wu and Yifan He. 2019. Enriching Pre-trained Language Model with Entity Information for Relation Classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 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*. 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*. 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 (08 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*. 2205–2215. https://doi.org/10.18653/v1/D18-1244