How Can a Teacher Make Learning From Sparse Data Softer? Application to Business Relation Extraction

Farah Benamara, Hadjer Khaldi, Camille Pradel, Nathalie Aussenac-Gilles

How Can a Teacher Make Learning From Sparse Data Softer? Application to Business Relation Extraction

Hadjer Khaldi[1,2*] , **Farah Benamara**[2] , **Camille Pradel**[1] , **Nathalie Aussenac-Gilles**[2]
[1]Geotrend, France
[2]IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
{hadjer, camille}@geotrend.fr, {farah.benamara, nathalie.aussenac-gilles }@irit.fr,

Abstract

Business Relation Extraction between market entities is a challenging information extraction task that suffers from data imbalance due to the over-representation of negative relations (also known as *No-relation* or *Others*) compared to positive relations that corresponds to the taxonomy of relations of interest. This paper proposes a novel solution to tackle this problem, relying on *binary soft-labels supervision* generated by an approach based on knowledge distillation. When evaluated on a business relation extraction dataset, the results suggest that the proposed approach improves the overall performances, beating state-of-the art solutions for data imbalance. In particular, it improves the extraction of under-represented relations as well as the detection of false negatives.

1 Introduction

Nowadays, the web is considered as an important source of business and financial information that can be used to analyze business interactions between market entities. These interactions enable financial institutions to take well-informed decisions [Oberlechner and Hocking, 2004], as well as business professionals to sustain and innovate in a rapidly changing business world. However, structuring this information remains a challenging task, given the volume and velocity of the textual data generated online. Hence, the availability of systems that automatically extract business interactions between organizations (e.g., *startups*, *companies*, *non-profit organizations*, etc.) from textual content becomes crucial.

Business Relation Extraction (BRE) is an NLP task that aims at discovering relations involving different companies (e.g. company-customer, company-partner) at the sentence level [Zhao *et al.*, 2010]. For example, from the sentence in (1), extracted from BIZREL dataset [Khaldi *et al.*, 2021], a relation extraction system can infer that the company *Airbus* is a supplier for the company *Inmarsat*.

Example 1 *The Airbus group has signed a contract with Inmarsat for the delivery of three reconfigurable geostationary satellites in orbit.*

Dataset	# Sent.	#Rel.	% NR
TACRED	106,264	42	79.5
BioRel	533,560	125	50
BizRel	10,034	6	63

Table 1: NR in existing generic and domain specific datasets.

Recent works for BRE rely on supervised approaches, where neural models are trained on annotated datasets for business relations [Collovini *et al.*, 2020; De Los Reyes *et al.*, 2021; Reyes *et al.*, 2021; Khaldi *et al.*, 2021]. In general, supervised approaches consider relation extraction (RE) as a multi-class classification problem where each class corresponds to a predefined relation type [Zhang *et al.*, 2017; Wu, 2019]. In addition to the set of *positive relations* (henceforth PR) which corresponds to the taxonomy of relations of interest (like hypernymy, meronymy, and cause-effect relationships), most popular datasets manually annotated either for generic (e.g., SemEval-2010 Task 8 [Hendrickx *et al.*, 2010], TACRED [Zhang *et al.*, 2017]) or domain specific relations (e.g., ChemProt [Krallinger *et al.*, 2017], BizRel [Khaldi *et al.*, 2021]) include a *negative relation* (henceforth NR) to account either for the absence of a relation between two target entities (see NO-RELATION in TACRED), or any other types of relations not present in the annotation scheme (see OTHERS in SemEval-2010 and BizRel). NRs share two main characteristics: (C1) they have irregular and unstable linguistic realizations and (C2) are often over-represented making PR hard to predict due to the highly imbalanced nature of the problem (see the ratio of NR in Table 1).

Several solutions have been proposed to address NR: discard them during training [Doddington *et al.*, 2004], ignore them at the evaluation stage focusing only on the performances of PR as done in most RE shared tasks [Zhang *et al.*, 2017; Hendrickx *et al.*, 2010], or include them during training by treating all relations equally [Wu, 2019; Zhou and Chen, 2021]. These strategies however fail to deal with the sparseness of PR and the characteristics of NR in a real world scenario. To overcome the data imbalance problem, four main solutions have been proposed in the literature:

(1) *Data augmentation* where different strategies based on lexical variations are used to generate new instances for minority classes [Su *et al.*, 2021; Papanikolaou and Pierleoni, 2020].

[*]Contact Author

(2) *Cost-sensitive learning* by assigning higher wrong classification costs to classes with small proportion [Lin *et al.*, 2018; Zhang *et al.*, 2017] .

(3) *Multitask learning* where auxiliary tasks help the main task to improve performances of under-represented classes [Khaldi *et al.*, 2021; Wang and Hu, 2020].

(4) *Knowledge distillation* (henceforth KD) that aims to transfer knowledge from a complex teacher model to a small student model, where the outputs of the teacher network, called soft labels, are used to train a student network [Hinton *et al.*, 2015; Zhang *et al.*, 2020; Song *et al.*, 2021]. The basic idea behind KD is that the teacher's soft probabilities have more knowledge about classes than the one hot-encoded labels used usually to train the student.

The first three solutions rely on hard labels supervision, where the ground truth labels are represented using one-hot encoded vectors that are not able to represent the semantic information among relations. Indeed, NR can have unstable patterns, and can share similar linguistic realizations with PR. For example, the SemEval 2010 Task 8 dataset [Hendrickx *et al.*, 2010] includes the OTHERS relations for near misses of PR as NR instances, while in the BIZREL dataset [Khaldi *et al.*, 2021] sentences expressing PR and NR at the same time between different pair of entities are one of the main sources of false negatives. While hard label supervision successes to counter the class imbalance problem (i.e. (C2)), it does not however fully capture the dissimilarities between PR and NR, making the optimization of model's output probabilities hard. Recent studies show that soft labels generated via KD by a teacher model are more adequate to efficiently handle the inherent characteristics of NR (C1) [Song *et al.*, 2021]. In this paper, we aim to continue these efforts by proposing a new knowledge distillation approach based on *binary soft labels supervision (BSLS)*. The soft outputs generated by the teacher model trained for binary classification (PR vs. NR), are used to supervise the student model to perform multi-class RE. Our contributions are three folds:

- A new knowledge distillation approach to account for NR characteristics in imbalanced RE problem based on *binary soft labels supervision*. As far as we know, KD has never been used for business RE.
- A comparison of our approach against several state of the art hard labels (data augmentation, cost-sensitive learning, multitask) and soft labels approaches.
- An evaluation of the performances of our model on a business relation extraction dataset. Our results show that our approach improves the extraction of under-represented relations as well as the detection of false negatives, addressing therefore both (C1) and (C2).

This paper is organized as follows: We first present the related work, then describe our KD architecture. We finally detail the carried experiments and give our results.

2 Related Work

2.1 Knowledge Distillation for RE

The main idea behind KD is to design a simple student model that mimics the behavior of a complex, more informed, or a large teacher model in order to achieve comparable results in performing a specific task. It has first been proposed for model compression task [Hinton *et al.*, 2015].

KD has been recently proposed for RE. Zhang [2020] incorporates knowledge about type constraints between entities and relations into the teacher model then use knowledge distillation to generate well informed soft labels used to supervise a student model that is able to inherit this knowledge from its teacher. Song et al. [2021] integrate ground truth sentence-level identification information into the teacher network during training then transfer it to the student by sharing the classification layer to counter data imbalance problem. KD has also been used to alleviate the interference of noise from relation annotations in distant supervision via label softening [Li *et al.*, 2022].

Our work is close to [Song *et al.*, 2021] but instead of adding more features to the teacher model, we rather train the teacher and student models on two different complementary tasks: binary relation identification (PR vs. NR) and multi-class relation extraction. We assume that training a teacher model on binary relation identification helps to learn discriminative features that differentiate PR from NR, on a less imbalanced dataset, since all PR are merged into one class. The student model can therefore inherit from the teacher's produced binary soft labels the salient learnt features about PR and NR, to mitigate NR irregular patterns problem. We also experiment with different data-imbalance sensitive loss functions in the student model in order to alleviate the (PR vs. NR) imbalance problem.

2.2 Business Relation Extraction

Most existing works for BRE have used semi-supervised approaches relying either on lexico-syntactic patterns generated from dependency trees [Braun *et al.*, 2018], or lexical patterns based on a list of keywords which are specific to each predefined relation type [Lau and Zhang, 2011]. Recent works rely on supervised approaches, where neural models are trained on annotated datasets for business relations. For example, Collovini et al. [2020] extract relations between Fintech companies from news text using Bidirectional Gated Recurrent Units. Recently, De Los Reyes et al. [2021], Reyes et al. [2021], and Khaldi et al. [2021] relied on BERT pretrained language model [Devlin *et al.*, 2019] fine-tuned on annotated datasets to classify relations between financial and economic entities. Most works focus either on business relations classification [Braun *et al.*, 2018; Lau and Zhang, 2011] where NR is not considered, or on business relation identification where all relations are merged into one PR type [Reyes *et al.*, 2021; Collovini *et al.*, 2020]. Only few works handles both business relation identification (PR vs. NR) and business relation classification by including a NR in the set of relations to extract [Khaldi *et al.*, 2021; De Los Reyes *et al.*, 2021]. Our work continues these efforts by proposing a supervised model for BRE based on BERT, to perform both business relation identification and classification, while handling for the first time, as far as we know, business PR sparsity through knowledge distillation.

3 A Binary Soft-labels Supervision for Multi-class RE (BSLS)

Our *binary soft label supervision* approach for multi-class relation extraction is based on knowledge distillation where binary soft labels generated by a teacher model noted T are used to supervise the training of a student model noted S (cf. Figure 1). Following [Zhou and Chen, 2021], both S and T have the same architecture based on an improvement of R-BERT [Wu, 2019], a transformer model specifically designed to handle RE tasks. This architecture has two main components: **a)** *a sentence encoder* noted $Encoder_i$ with $i \in \{S, T\}$ based on the pre-trained BERT model [Devlin *et al.*, 2019] while using entity markers as sentence representation vectors, **b)** *a relation classifier* noted $Classifier_i$ composed of two linear layers followed by dropout layer then a softmax activation function.

An input sentence is first fed into $Encoder_i$, to get its contextual representations that are injected into $Classifier_i$ to predict the relation type. Let $P_i = (P_{i0}, ..., P_{in})$ the prediction probabilities generated by $Classifier_i$, with n being the number of relations to predict. Let P_{SoftT} the *soft labels*, i.e., the prediction probabilities generated by a pre-trained teacher binary classifier $Classifier_T$ whose weights are frozen and shared with S. Finally, let Y_b and Y_m be respectively the binary and multi-class *hard labels* that encode the ground-truth labels as one hot vectors. These soft and hard labels are used by two different losses in order to optimize the models parameters through back-propagation: $\mathcal{L}_{cT}$ (resp. $\mathcal{L}_{cS}$) , the classification loss that minimizes the errors between P_T and Y_b (resp. P_S and Y_m). and $\mathcal{L}_D$, the distillation loss calculated between a binarised form of P_S and P_{SoftT}.

The distillation algorithm consists in the following steps:

(1) First, train T on binary relation identification (PR Vs. NR), while optimizing the teacher classification loss $\mathcal{L}_{cT}$.

(2) Then $Classifier_T$'s weights are frozen and shared with S.

(3) S is trained on multi-class RE and supervised by both Y_m and P_{SoftT}, while optimizing both the student classification loss $\mathcal{L}_{cS}$ and the distillation loss $\mathcal{L}_D$. To this end, P_S are first binarised into P_{Sb} following the equation (1) where P_{S_0} refers to the prediction probability of NR as given by $Classifier_S$.

$$P_{Sb} = (P_{S0}, max(P_{S1}, ..., P_{Sn})) \tag{1}$$

(4) The weighted sum of $\mathcal{L}_{cS}$ and $\mathcal{L}_D$ is the final loss $\mathcal{L}_f$ optimized to train the student model, α =0.6 , $\beta = 0.4$, being loss weights.

$$\mathcal{L}_f = \alpha.\mathcal{L}_{cS} + \beta.\mathcal{L}_D \tag{2}$$

4 Data and Experiments

4.1 Baselines

We compare our model against four baseline models used to tackle data imbalance in RE: augmentation of the training data (DA), multitask architecture (MLT), optimizing using an adapted loss (ALS), and knowledge distillation (KD) via soft labels. We describe below each of these configurations.

1- Shortest dependency path data augmentation (DA$_{SDP}$) [Su *et al.*, 2021]: The main idea of data augmentation is to generate new instances that express the same relation. As the shortest dependency path is assumed to capture the required information to express a relation between two target entities [Bunescu and Mooney, 2005], the augmentation consists in extracting tokens located in this path, fixing them, then the rest of tokens are randomly transformed by: synonyms replacement, random swapping, and random deletion. In our experiment, this method augments the positive instances by 300%.

2- Multitask architecture (MLT$_{bin}$) [Khaldi *et al.*, 2021]: This is a multitask RE model that performs both relation identification (PR vs. NR) and relation extraction (multi-class classification). The relation identification task is an auxiliary task designed to help the main task of multi-class relation classification learn more features about PR vs. NR distinction. We use here a simplified version of MLT without considering any additional semantic features.

3- Adapted loss (ALS) : We rely on four adapted losses as follows:

– Weighted Cross Entropy loss (ALS$_{WCE}$) : A variant of cross-entropy loss that assigns to each class a pre-computed weight that corresponds to the penalty of miss-classifying its instances.

– Focal loss (ALS$_{FC}$) [Lin *et al.*, 2017]: This loss has shown to be very effective for object detection from highly imbalanced datasets since it down-weights easy examples and thus focus the training on hard negatives by adding a modulating factor to the cross-entropy loss.

–Adaptive scaling (ALS$_{AD}$) [Lin *et al.*, 2018]: It is a dynamic cost-sensitive learning algorithm that optimizes the F-score rather than the accuracy and adaptively scales costs of instances of different classes with a marginal utility that quantify the importance of positive/negative instances during training.

– Dice loss (ALS$_{DC}$) [Li *et al.*, 2020]: The Dice function is a widely used metric for evaluating image segmentation accuracy. It is the harmonic mean of precision and recall. It attaches equal importance to false positives and false negatives. We use here the weighted version of Dice loss to control the trade-off between precision and recall and down-weight easy examples. As far as we know, dice loss has never been used for RE.

4- Soft label supervision using knowledge distillation (KD$_{SLS}$): Soft labels generated by a teacher model trained on multi-class RE task are use to supervise a student model performing the same task. We use the focal loss to train the teacher model in order to handle class-imbalance when generating soft labels. This is the standard KD following [Hinton *et al.*, 2015], where the teacher and the student models perform

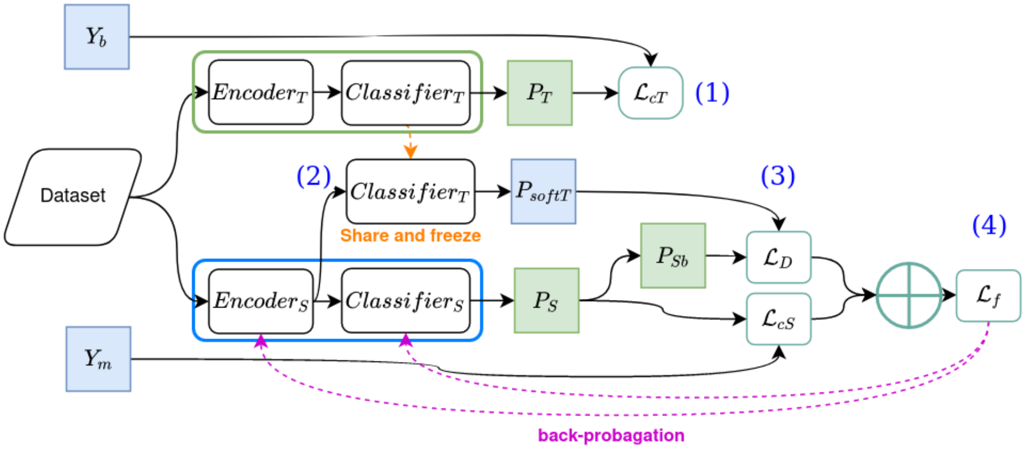


Figure 1: Binary soft-labels supervision architecture for Business Relation Extraction. (1) Teacher training, (2) Teacher classifier freezing and sharing, (3) Student training through knowledge distillation, (4) Final loss to train the student.

the same task, while only the teacher classifier is distilled as in [Song *et al.*, 2021]. Note that our teacher model is simpler as it does not include any additional features.

4.2 Data

We run experiments on the BIZREL dataset, [Khaldi *et al.*, 2021] a business relation extraction corpus freely available for research purposes.[1] The dataset has 10k relation instances between named entities of type *Organization*. It is composed of 5 positive relations (INVESTMENT, COMPETITION, COOPERATION, LEGAL-PROCEEDING, and SALE-PURCHASE) and one negative relation named OTHERS.

Data distribution per relation type and dataset type (train, test) are presented in Table 2. We can observe that the NR is over-represented compared to the other PR, representing 66.2% of the training data and 66.2% of the test. When looking at NR instances, we can notice that the patterns used to express this relation are irregular (see Examples 2 and 3), since a negative relation can be assigned to any other non-business relation such as: *list of sponsors*, *list of innovative companies*, or *employee's transfer from company A to company B*.

Example 2 *Shira Goodman, the former CEO of Framingham office supply retailer [**Staples**]$_{E1}$, has been elected to the board of directors of Los Angeles real estate giant [**CBRE Group**]$_{E2}$.*

Example 3 *Ten French entities were among the world's 100 most innovative organizations in 2016: three research centers (CNRS, CEA, IFP Energies Nouvelles) and seven companies (Alstom, [**Arkema**]$_{E1}$, [**Safran**]$_{E2}$, Saint-Gobain, Thales, Total, and Valeo).*

In addition, these patterns can be very close to the ones used to express PR. In Example 4, a NR is annotated between E_1 and E_3 while a PR of type COOPERATION exists between E_1 and E_2. We can notice that for both entity pairs, the pattern form E_1 *partners with* E_2 exists.

Data	Inv.	Com.	Coo.	Leg.	Sal.	Oth.	**#Tot.**
Train	281	1,675	627	50	248	5,647	8,528
Test	50	296	111	8	44	997	1,506

Table 2: BIZREL dataset distribution per relation type and per dataset type (train, test).

	Inv.	Com.	Coo.	Leg.	Sal.	Oth.
Avg. w_per_s	32	44	35	29	32	40
Avg. e_per_s	4	8	5	3	4	7
Avg. v_per_s	3	2	3	3	2	2

Table 3: BIZREL dataset complexity per relation type.

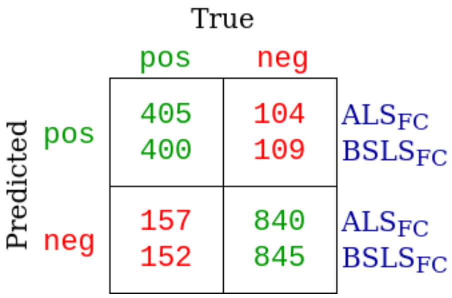
Example 4 *While [**Airbus**]$_{E1}$ partners with [**Audi**]$_{E2}$, Boeing is cozying to [**Adient**]$_{E3}$, Mercedes-Benz, and even General Motors.*

To measure the complexity of business relations in BIZREL dataset and their syntactic richness, we compute the average count of words, verbs, and entities per relation type (*Avg. w_per_s*, *Avg. v_per_s*, and *Avg. e_per_s* respectively). Table 3 shows the results. We observe that sentences contain on average from 3 to 8 named entities of type *organization*, therefore, potentially a maximum of 6 to 28 relations could occur in a single sentence between different entity pairs. In addition, sentences are complex containing in average from 2 to 3 verbs and the context surrounding a given relation instance varies from 29 to 44 tokens on average. Overall, these measures confirm the diversity and complexity of business relations expressed in BIZREL. This is more salient for OTHERS and COMPETITION where the average number of entities per sentence is 7 and 8 respectively, while the context (i.e., number of words per sentence) is respectively of 40 and 44.

[1] Link to BizRel dataset

Model	P	R	F1
ALS$_{CE}$	62.5	72.5	66.7
ALS$_{WCE}$	63.1	**75.1**	68.1
ALS$_{FC}$ [Lin *et al.*, 2017]	65.9	71.7	68.5
ALS$_{DC}$ [Li *et al.*, 2020]	66.9	65.4	65.7
ALS$_{AD}$ [Lin *et al.*, 2018]	62.6	70.9	66.0
MLT$_{bin}$ [Khaldi *et al.*, 2021]	62.8	73.2	67.2
DA$_{SDP}$ [Su *et al.*, 2021]	**69.7**	67.8	68.2
KD$_{SLS}$ [Song *et al.*, 2021]	63.9	70.9	67.0
BSLS$_{CE}$	65.4	71.7	68.2
BSLS$_{WCE}$	63.0	73.2	67.1
BSLS$_{FC}$	66.1	75.0	**69.9**
BSLS$_{DC}$	66.7	69.8	68.1
BSLS$_{AD}$	66.6	69.8	67.6

Table 4: Experimental results on the BizRel dataset. Best results are in bold.



		True		
		pos	neg	
Predicted	pos	405 400	104 109	ALS$_{FC}$ BSLS$_{FC}$
	neg	157 152	840 845	ALS$_{FC}$ BSLS$_{FC}$

Figure 2: Confusion matrix to compare between business and non-business instance classification in our best model (BSLS$_{FC}$) and the best baseline (ALS$_{FC}$)

	Inv.	Com.	Coo.	Leg.	Sal.	Oth.
ALS$_{FC}$	61.0	**78.8**	65.0	**77.8**	41.9	**86.6**
BSLS$_{FC}$	**68.9**	77.2	**66.7**	73.7	**46.2**	**86.6**

Table 5: Best baseline (ALS$_{FC}$) and our best model (BSLS$_{FC}$) F1-score per relation type. Best results of each relation are in bold.

5 Results and Discussion

Results of the baselines and BSLS experiments are reported in Table 4, in terms of macro precision, recall, and F-score. [2]

Overall, we can observe that the proposed model based on *binary soft labels supervision* (BSLS) optimized using a focal loss (FC) is the best, achieving an F-score of 69.9%, outperforming therefore all the baselines (+1.4% over the best one). *Shortest dependency path* (AD$_{SDP}$) data augmentation obtains the best precision (69.7%) while the *weighted cross entropy loss* (ALS$_{WCE}$) the best recall (75.1%).

When comparing between knowledge distillation models, we can observe that our *binary soft labels* (BSLS) are more efficient than KD$_{SLS}$, the *multi-class soft labels* state-of-the art (+2.9% F-score).

When experimenting BSLS with different loss functions, we notice that, for most of the experiments, BSLS optimized using $loss_i$ outperforms the baseline model optimized using the same $loss_i$. For example, BSLS$_{CE}$ scores higher than ALS$_{CE}$ (+1.5 % F-score), BSLS$_{DC}$ is better than ALS$_{DC}$ (+2.4 % F-score), BSLS$_{AD}$ outperforms ALS$_{AD}$ (+1.6 % F-score), and finally BSLS$_{FC}$ outperforms ALS$_{FC}$ (+1.4 % F-score).

We further compare the performances of the best baseline (ALS$_{FC}$) with our best performing model (BSLS$_{FC}$). Figure 2 gives a confusion matrix that shows the number of false/true positives/negatives between PR and NR. We can see that BSLS$_{FC}$ was able to reduce the number of false negative instances (from 157 to 152), and increase the true negative (from 840 to 845). We can also observe the impact of these changes on the recall where our model achieve one of the best score. It was however not able to reduce misclassifications due to false positive, leading therefore to a decrease in the precision when compared to the best precision.

A closer look into the results per class for the best baseline and best performing model (cf. Table 5) shows that our model is able to improve the performances of most under-represented positive relations, namely: INVESTMENT, COOPERATION and SALE-PURCHASE that represent 3.3%, 7.3% and 2.9% of test set. NR results remain stable and this was expected as our approach was specifically designed to handle under-represented PR. A final interesting finding is that PR with less frequencies are the one that benefits the most from *binary soft labels*. For example, an improvement of +7.9 % (resp. +4.3 %) in terms of F1 is observed for under-represented relation INVESTMENT (resp. SALE-PURCHASE) over the best baseline.

In order to gain insights into the main strengths of the current approach when compared to the best baseline, we analyse well classified instances by BSLS$_{FC}$, that ALS$_{FC}$ fails to classify correctly. We notice that our approach is able to identify the NR OTHERS in some cases where many relations are expressed between different target entities, unlike ALS$_{FC}$ (See example 5).

Example 5

While there were few mega acquisitions/ mergers primarily Chinese players acquiring European and US robotics/ automation companies (Kuka AG by [***Midea Group*** *]$_{E_1}$, Dematic by [* ***Kion Group*** *]$_{E_2}$ and KraussMaffei Automation by ChemChina) and few others by US industry giants (Affeymetrix by ThermoFisher and Intelligrated by Honeywel), most acquisitions were in the sub $ 500 M range .*

BSLS$_{FC}$'s correct label : OTHERS, **ALS$_{FC}$'s wrong label:** INVESTMENT

In addition, our model is also able to distinguish between semantically close PR such as INVESTMENT, SALE-PURCHASE, and COOPERATION, that uses the same lexical cues to be expressed such as *signing agreement*, *entering into a contract*. In example 6, the expression *entering into a contract* refers to *service-selling* contract rather than a COOPERATION relation.

[2] All models are based on `bert-base-cased`. Using FinBERT [Araci, 2019] did not improve the overall performances, where BSLS$_{FC}$ achieves the best F1 (68.9%), followed by ALS$_{FC}$ (68.7%).

Example 6

[***General Electric Corporation*** *]$_{E_1}$ has entered into a five - year, \$ 128,500 million contract with [* ***Electronic Data Systems*** *]$_{E_2}$ (EDS) to handle the corporation's desktop computer procurement, service, and maintenance activities.*

BSLS$_{FC}$'s correct label : SALE-PURCHASE, **ALS$_{FC}$'s wrong label:** COOPERATION

6 Conclusion

In this paper, we propose a novel solution to tackle PR vs. NR imbalance and NR irregular patterns problems, relying on *binary soft-labels supervision* generated by knowledge distillation. When evaluated on a business relation dataset, our approach improves the overall performances by enhancing the detection of under-represented relations and reducing false negative misclassification rates. As future work, we plan to evaluate our method to other generic and domain specific RE datasets in order to assess its adaptability to other domains.

References

[Araci, 2019] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. 2019.

[Braun *et al.*, 2018] D. Braun, A. Faber, A. Hernandez-Mendez, and F. Matthes. Automatic relation extraction for building smart city ecosystems using dependency parsing. In *Proceedings of the 2nd Workshop on Natural Language for Artificial Intelligence*, pages 29–39, 2018.

[Bunescu and Mooney, 2005] R. Bunescu and R. Mooney. A shortest path dependency kernel for relation extraction. In *Proceedings of HLT and EMNLP*, pages 724–731, 2005.

[Collovini *et al.*, 2020] S. Collovini, P. N. Gonçalves, G. Cavalheiro, J. Santos, and R. Vieira. Relation extraction for competitive intelligence. In *International Conference on Computational Processing of the Portuguese Language*, pages 249–258. Springer, 2020.

[De Los Reyes *et al.*, 2021] D. De Los Reyes, A. Barcelos, R. Vieira, and I. Manssour. Related named entities classification in the economic-financial context. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 8–15, 2021.

[Devlin *et al.*, 2019] J. Devlin, MW. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186, 2019.

[Doddington *et al.*, 2004] G. R Doddington, A. Mitchell, M. A Przybocki, L. A Ramshaw, S. M Strassel, and R. M Weischedel. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, pages 837–840, 2004.

[Hendrickx *et al.*, 2010] I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, and S. Szpakowicz. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, 2010.

[Hinton *et al.*, 2015] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. In *In Proc. of NeurIPS*, 2015.

[Khaldi *et al.*, 2021] H. Khaldi, F. Benamara, A. Abdaoui, N. Aussenac-Gilles, and E. Kang. Multilevel entity-informed business relation extraction. In *International Conference on Applications of Natural Language to Information Systems*, pages 105–118. Springer, 2021.

[Krallinger *et al.*, 2017] M. Krallinger, O. Rabal, S. A Akhondi, M. P. Pérez, J. Santamaría, G. P. Rodríguez, G. Tsatsaronis, and A. Intxaurrondo. Overview of the biocreative vi chemical-protein interaction track. In *Proceedings of the sixth BioCreative challenge evaluation workshop*, volume 1, pages 141–146, 2017.

[Lau and Zhang, 2011] R. Lau and W. Zhang. Semi-supervised statistical inference for business entities extraction and business relations discovery. In *SIGIR 2011 workshop*, pages 41–46, 2011.

[Li *et al.*, 2020] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li. Dice loss for data-imbalanced nlp tasks. In *Proceedings of the 58th ACL*, pages 465–476, 2020.

[Li *et al.*, 2022] R. Li, C. Yang, T. Li, and S. Su. Midtd: A simple and effective distillation framework for distantly supervised relation extraction. *ACM Transactions on Information Systems (TOIS)*, (4):1–32, 2022.

[Lin *et al.*, 2017] T.-Y. Lin, P. Goyal, R.s Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[Lin *et al.*, 2018] H. Lin, Y. Lu, X. Han, and L. Sun. Adaptive scaling for sparse detection in information extraction. *arXiv preprint arXiv:1805.00250*, 2018.

[Oberlechner and Hocking, 2004] T. Oberlechner and S. Hocking. Information sources, news, and rumors in financial markets: Insights into the foreign exchange market. *Journal of economic psychology*, pages 407–424, 2004.

[Papanikolaou and Pierleoni, 2020] Y. Papanikolaou and A. Pierleoni. Dare: Data augmented relation extraction with gpt-2. *arXiv preprint arXiv:2004.13845*, 2020.

[Reyes *et al.*, 2021] D.D.L. Reyes, D. Trajano, I. Manssour, R. Vieira, and R. Bordini. Entity relation extraction from news articles in portuguese for competitive intelligence based on bert. In *Brazilian Conference on Intelligent Systems*, pages 449–464. Springer, 2021.

[Song *et al.*, 2021] D. Song, J. Xu, J. Pang, and H. Huang. Classifier-adaptation knowledge distillation framework for relation extraction and event detection with imbalanced data. *Information Sciences*, 573:222–238, 2021.

[Su *et al.*, 2021] Peng Su, Yifan Peng, and K. Vijay-Shanker. Improving BERT model using contrastive learning for biomedical relation extraction. In *Proceedings of the 20th Workshop BioNLP*, pages 1–10. ACL, 2021.

[Wang and Hu, 2020] W. Wang and W. Hu. Improving relation extraction by multi-task learning. In *Proceedings of HPCCT'20 BDAI'20*, pages 152–157, 2020.

[Wu, 2019] Y. Wu, S.and He. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the CIKM'19*, pages 2361–2364, 2019.

[Zhang *et al.*, 2017] Y. Zhang, V. Zhong, D. Chen, G. Angeli, and C. D. Manning. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on EMNLP*, pages 35–45. ACL, 2017.

[Zhang *et al.*, 2020] Z. Zhang, X. Shu, B. Yu, T. Liu, J. Zhao, Q. Li, and L. Guo. Distilling knowledge from well-informed soft labels for neural relation extraction. In *Proceedings of the AAAI Conference*, pages 9620–9627, 2020.

[Zhao *et al.*, 2010] J. Zhao, P. Jin, and Y. Liu. Business relations in the web: Semantics and a case study. *Journal of Software*, (8):826–833, 2010.

[Zhou and Chen, 2021] W. Zhou and M. Chen. An improved baseline for sentence-level relation extraction. 2021.