An autoencoder-based representation for noise reduction in distant supervision of relation extraction

García-Mendoza, Juan Luis; Villaseñor-Pineda, Luis; Orihuela-Espina, Felipe; Bustio-Martínez, Lázaro

DOI:
10.3233/JIFS-219241

Document Version
Peer reviewed version

Citation for published version (Harvard):
García-Mendoza, JL, Villaseñor-Pineda, L, Orihuela-Espina, F & Bustio-Martínez, L 2022, 'An autoencoder-based representation for noise reduction in distant supervision of relation extraction', *Journal of Intelligent and Fuzzy Systems*, vol. 42, no. 5, pp. 4523-4529. https://doi.org/10.3233/JIFS-219241

Undefined 0 (0) 1
IOS Press

An Autoencoder-based Representation for Noise Reduction in Distant Supervision of Relation Extraction

Juan-Luis García-Mendoza [a,*]
Luis Villaseñor-Pineda [a]
Felipe Orihuela-Espina [a]
Lázaro Bustio-Martínez [b]

[a] *Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Puebla, México*
[b] *Universidad Iberoamericana, DEII, Prolongación Paseo de Reforma 880, 01219, CDMX, México.*

Abstract. Distant Supervision is an approach that allows automatic labeling of instances. This approach has been used in Relation Extraction. Still, the main challenge of this task is handling instances with noisy labels (e.g., when two entities in a sentence are automatically labeled with an invalid relation). The approaches reported in the literature addressed this problem by employing noise-tolerant classifiers. However, if a noise reduction stage is introduced before the classification step, this increases the macro precision values. This paper proposes an Adversarial Autoencoders-based approach for obtaining a new representation that allows noise reduction in Distant Supervision. The representation obtained using Adversarial Autoencoders minimize the intra-cluster distance concerning pre-trained embeddings and classic Autoencoders. Experiments demonstrated that in the noise-reduced datasets, the macro precision values obtained over the original dataset are similar using fewer instances considering the same classifier. For example, in one of the noise-reduced datasets, the macro precision was improved approximately 2.32% using 77% of the original instances. This suggests the validity of using Adversarial Autoencoders to obtain well-suited representations for noise reduction. Also, the proposed approach maintains the macro precision values concerning the original dataset and reduces the total instances needed for classification.

*Corresponding author. E-mail: juanluis@inaoep.mx

Keywords: noise reduction, adversarial autoencoders, distant supervision

1. Introduction

Relation Extraction (RE) is concerned with “detecting and classifying predefined relationships between entities identified in-text" [15]. In RE, a text sentence is analyzed to retrieve two named entities of interest and a specific association between them. Often, there is an interest associations collection with a rich set of entities participating in the associations. In the RE classification task, the classes are different association labels, and the classification consists of assigning the most likely label expressing the relation between the entities. Variants of RE may depart from the raw text or pre-extracted entities or a combination of them, but the one departing from the pre-extracted entities is the purest form of the RE problem.

RE has been addressed in different ways [17], including Distant Supervision (DS) [13]. According to [17], “Distant supervision combines the benefits of semisupervised and unsupervised relation extraction approaches and leverages a knowledge base as a source of training data". The main idea of DS is automatically labeling a dataset leveraging on existing knowledge; generally, in the form of knowledge bases [13]. This automatic labeling requires a heuristic or assumption to annotate the instances in the dataset construction. There are two main assumptions in the literature. The first one was proposed by Mintz et al. [13] (we will name it *Mintz assumption*) who assumed that “if two entities participate in a relation, *any* sentence that contains those two entities might express that relation". Riedel et al. [16] relaxed this assumption (we

*Corresponding author. E-mail: juanluis@inaoep.mx

will name it *Riedel assumption*), instead of assuming that "if two entities participate in a relation, *at least one* sentence that mentions these two entities might express that relation". Both assumptions are inadequate in many cases, such as when no sentence expresses the relation, leading to the introduction of false positives (noise in the labels) in the annotation of the dataset. Often, a pair of entities in a sentence does not imply a relationship or may express several relations concomitantly depending on the context, as depicted in Fig 1.

When the outcome of the annotation of a dataset by DS is later served to a classification problem, noise in the labels might have a detrimental effect. Whether the noise in the labels arises from the failure of the assumption made by the DS labeling or any other different process is circumstantial. The original noise is irrelevant for the classifier, but the classifier has to deal with it nonetheless. In general, a method is robust if it can operate (e.g., find the correct solution) in the presence of noise and/or outliers. Still, it is worth noting that robustness is never universal, and all robust methods have a critical limit of noise that they can tolerate before failing. Regardless, several works reported in the literature have a certain amount of tolerance to noise by combining the *Riedel assumption* with Deep Neural Networks [21,10,9,18,22,20]. However, perhaps the most obvious way to improve the performance of the classifiers is to use cleaning methods in a previous step. This is used to alleviate the noise presence in the class labels [4], with the additional benefit that in any case, this solution can be combined with the use of robust classifiers to achieve a good classification. Besides, defining a separate task for explicitly cleaning the dataset can yield cleaned datasets useful for purposes other than classifying.

The main contribution of this paper is to obtain a new data representation for noise reduction in DS using Adversarial Autoencoders. The proposed data representation will allow obtaining datasets with less noise which implies that the macro precision will be improved. As a direct consequence, once an instance is classified with a relation, it is more likely to be correct. To validate this hypothesis, we use noise-tolerant classifier BGWA (BiGRU-based word attention model) [9] to measure the macro precision on the new datasets obtained.

2. Related Work

Nowadays, there are many noise-tolerant methods in DS [21,10,9,18,22,20]. One of the earliest approaches based on DNN was the Piecewise Convolutional Neural Networks (PCNN) proposed by Zeng et al. [21]. This network builds bags of instances from the entities pairs that are considered correct if at least one of the sub-networks labels it positively (Multi-instance Learning). In other DDN-based approaches [10,22,9, 20] different attention mechanisms have been incorporated to deal with noise. Examples are sentence-level attention [10], word-level attention to dynamically highlight important parts of the sentence [22], attention over words to identify such key phrases is used (BGWA) [9], and intra-bag and inter-bag attention [20].

In addition, noisy labels are also frequently dealt with using data cleaning methods [4]. In principle, they may be evident to a potential subsequent classification exercise, but when classification follows an assessment by wrapping, it often guides the cleaning exercise. Depending on how conservative they are, data cleaning methods can eliminate too few or too many instances, thus reducing the performance of the potential subsequent classifiers [12]. Brodley and Friedl [2] advocate that it is preferable to eliminate several instances correctly labeled than to maintain instances with noisy labels. However, this is only possible or convenient when the acquisition of instances is cheap, and the instances are abundant, which might not always be the case. Notwithstanding, the cleaning or filtering of labels with current methods does not guarantee the total elimination of noise. Complete elimination of noise is certainly achievable, even if silly. It suffices with deannotating every instance. However, such an extreme approach has no practical application for obvious reasons. Therefore in practice, a compromise between confidence intervals and false-positive acceptance rate ought to be sought. According to [14] the Autoencoders (AE) can be used when there are noisy instances.

3. Sentence Embeddings

Sentence Embeddings (SE), like Word Embeddings (WE), are real values that contain the semantic meaning, in this case, of the complete sentence, distributed in a k-dimensional vector [3]. As with WE, there are pre-trained models [3].

– Pre-trained SE proposed by [3]: Two models are presented that are optimized for texts that have more than one word, such as sentences or paragraphs. Given a certain text, these models return their cor-

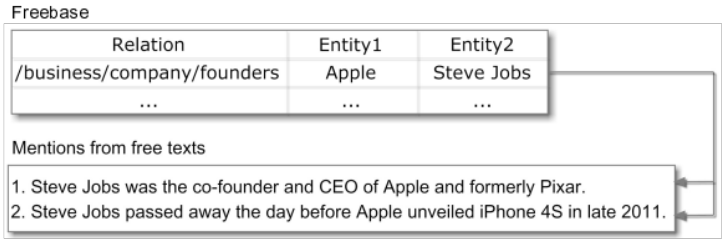


Fig. 1. In this example, it is included a pair of entities that not express the same relation. Considering the *founders* relation, the first one will be correctly labeled. While the second will not. Example reproduced from [21].

responding vector. One model is based on the use of transformers (Transformers) that presents greater accuracy at the cost of greater consumption of resources and a more complex model [19]. The other model is based on Deep Averaging Networks, which has lower accuracy but higher efficiency [8]. The model was trained using sources from Wikipedia, web news, question and answer pages, and discussion forums. In this work, we will refer to these SE as TRANSF[1] and DAN[2] respectively.

– Pre-trained SE proposed by [1]: An architecture is proposed to learn joint multilingual sentence representations for 93 languages. This architecture uses a language-agnostic BiLSTM encoder to build the SE, coupled with an auxiliary decoder and trained on parallel corpora. The authors named this SE as LASER[3].

4. Autoencoders

Autoencoders (see Fig. 2a) are models that projects the input into the output [5]. Perhaps the most classical AE is Principal Component Analysis (PCA). AE transforms the input data X to a latent space Z using a function f (the encoder), and returns $X^{'}$ from Z using a decoder function g. If f is invertible ($\exists\ f^{-1} : g = f^{-1}$), the recovery will be errorless. Otherwise, the goal is to reconstruct the input $X^{'}$ minimizing the error $L = |X - g(f(X))|^2$ in reconstruction, or some variant of L e.g. regularized, generalized projections, etc.

Adversarial AE (AAE) (see Fig. 2b) are a particular AE coupled with a Generative Adversarial Networks (GAN) [6]. AAE is trained to fulfill two objectives: (1) minimizing the error L in the reconstruction of the input, while (2) fitting the vectors of the latent space Z to a previously known distribution [11]. AAE can be used for semi-supervised classification, unsupervised clustering, dimensionality reduction, and data visualization [11]. Furthermore, adjusting the representation to a known distribution allows us to detect instances far from this distribution and consider them as noisy.

5. Methods

Let:

– A set of sentences $\mathcal{S} = \{s_i | i = 1 \dots I\}$ and I the cardinality of $\mathcal{S}$.
– A set of labels (relations) $\mathcal{R} = \{r_j | j = 1 \dots J\}$ and J the cardinality of $\mathcal{R}$, and one of these is the $\mathcal{NA}$ relation which is the negative class. The set $\mathcal{R}^{\dagger} = \mathcal{R}/\mathcal{NA}$.
– A set of observations $\mathcal{X} = \{x_k | x_k = (s_i, r_j) \in \mathcal{S} \times \mathcal{R}^{\dagger}\}$.
– A partition over $\mathcal{X}$, $\mathcal{P}(\mathcal{X}) = \{\mathcal{X}_{train}, \mathcal{X}_{test}\}$.

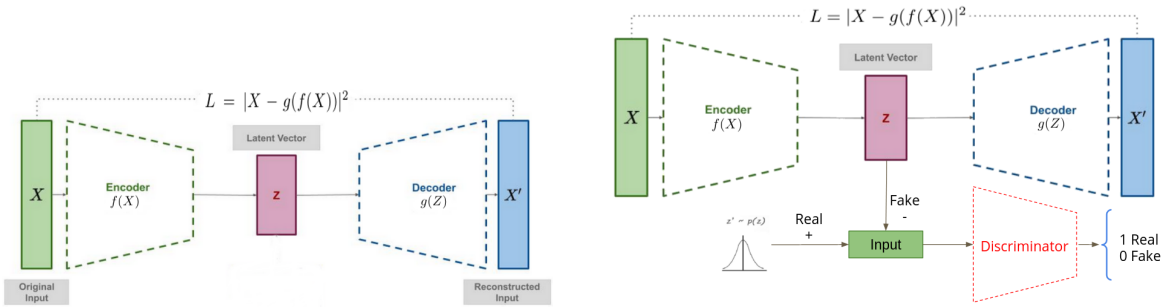
The problem of classification in RE is given a sentence $s_i \in \mathcal{S}$, assign a relation label $r_j \in \mathcal{R}^{\dagger}$ where the classifier $C : \mathcal{S} \rightarrow \mathcal{R}^{\dagger}$ with association rule $r_j = C(s_i)$ ought to be learned in advance from $\mathcal{X}_{train}$ with some given minimization of error in the observations set $\mathcal{X}_{test}$. But this classification is going to occur over a noisy set of observations. Therefore we define the cleaning problem. Let:

– A partition over the training set $\mathcal{X}_{train}$ defined by the relation r_j, $\mathcal{P}(\mathcal{X}_{train}) = \{\mathcal{X}^j_{train} | \forall j \in J - 1 : \mathcal{X}^j_{train} = \{x^j_{train} = (s_i, r_j) \in \mathcal{X}_{train}\}\}$, that is, it does not consider the $\mathcal{NA}$ relation

We define an encoder for each $\mathcal{X}^j_{train}$:

$$\begin{aligned} encoder^j : &\mathcal{S} \rightarrow \mathcal{V} \\ &(s_i[x^j_{train}], v_{i \times j}) \end{aligned} \tag{1}$$

[1] https://tfhub.dev/google/universal-sentence-encoder-large/5
[2] https://tfhub.dev/google/universal-sentence-encoder/4
[3] https://github.com/facebookresearch/LASER



(a) AE (b) AAE

Fig. 2. General architecture of the AE and AAE (reproduced and adapted from `https://deepnotes.io/deep-clustering` respectively).

where $\mathcal{V}^j = \{v_{i\times j} | v_{i\times j} \in \mathbb{R}^n\}$ is the vector representation of each sentence s_i according to the relation r_j. This representation is generated by the encoder $encoder^j$.

We define a cleaning function for each $\mathcal{V}^j$:

$$cleaning^j : \mathcal{S} \overset{encoder^j}{\rightarrow} \mathcal{V}^j \rightarrow \mathcal{S}$$

$$cl_i^j = cleaning^j(s_i) = \begin{cases} s_i \; if \; \neg noisy(v_{i\times j}) \\ \emptyset \quad otherwise \end{cases} \tag{2}$$

where $noisy$ is when the distance (Cosine or Euclidean are used) of v_i exceed the average distances between all $v_{i\times j}$; and the union of all cl_i^j, $\mathcal{X}'_{train} = \bigcup_{j\ldots J, i\ldots I} cl_i^j$ forms the new training set for training the classifier C. Figure 3 summarises the proposed method.

The functions $encoder^j$ used for obtaining the new representations $\mathcal{V}^j$ were generated by the following encoders:

– *f_laser*, *f_ae_laser*, and *f_aae_laser*: We only used LASER embeddings, and we use it as input for AE (Fig. 4a) and AAE (Fig. 4b), respectively.
– *f_dan*, *f_ae_dan*, and *f_aae_dan*: We only used DAN embeddings, and we use it as input for AE (Fig. 4a) and AAE (Fig. 4b), respectively.
– *f_transf*, *f_ae_transf*, and *f_aae_transf*: We only used TRANSF embeddings, and we use it as input for AE (Fig. 4a) and AAE (Fig. 4b), respectively.

The AE and AAE architectures are composed of two dense layers (1000 units and a ReLu-like activation function), both in the encoder f_θ and the decoder g_θ (see Fig. 4a and 4b). The input is a vector representing all the sentences in the text. This vector is obtained using some available sentence representation such as LASER [1], DAN [3], or TRANSF [3] pre-trained embeddings. Our proposal as $encoder^j$ is the AAE under the assumption that an observation (s_i, r_j), where the relation r_j is noisy, will not fit correctly to the distribution of the rest of the observations, and will remain far away. The discriminator input is one-third of each $\mathcal{X}^j_{train}$, while the autoencoder is the remainder. I.e., an attempt is made to adjust two-thirds of each $\mathcal{X}^j_{train}$ to the distribution of one-third. The way to select the instances to train the discriminator will be studied in deep as future work. In this research, instances were selected randomly.

We obtained the following $\mathcal{X}'_{train}$ training sets for each $encoder^j$ functions:

– *base_laser*, *aae_laser*, and *ae_laser*: We obtain these sets using *f_laser*, *f_ae_laser*, and *f_aae_laser* like $encoder^j$ function.
– *base_dan*, *aae_dan*, and *ae_dan*: We obtain the sets using *f_dan*, *f_ae_dan*, and *f_aae_dan* like $encoder^j$ function.
– *base_transf*, *aae_transf*, and *ae_transf*: We obtain these sets *f_transf*, *f_ae_transf*, and *f_aae_transf* like $encoder^j$ function.

6. Experiments and evaluation

In this section, we present the experiments conducted to demonstrate the validity of the proposed representation.

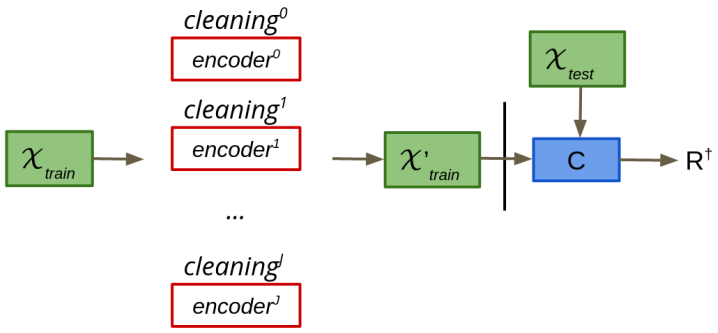
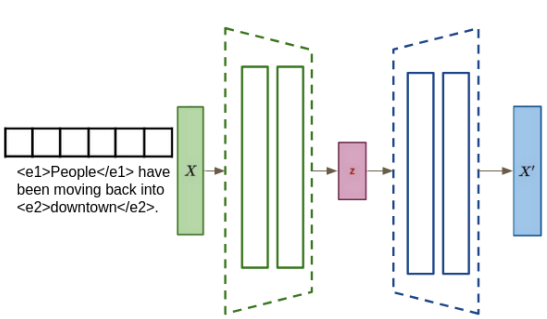
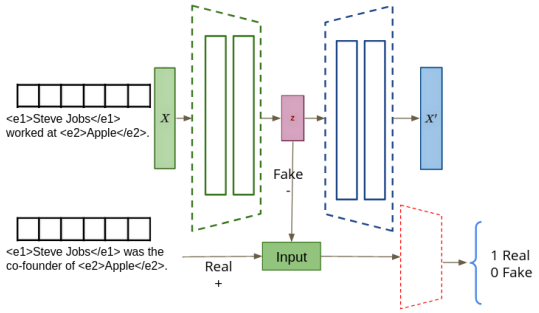


Fig. 3. Overview of the methodology proposed in this work.



(a) AE with dense layers and one vector of input. The autoencoder input is $\mathcal{X}^j_{train}$.



(b) AAE with dense layers and one input vector. The discriminator input is one third of each $\mathcal{X}^j_{train}$ while the autoencoder is the remainder.

Fig. 4. Architectures used in this work.

6.1. *Assessment of the model representation capability*

We performed an evaluation of the intra-cluster distances over each $\mathcal{V}^j$ obtained with $encoder^j$ using the Cosine distance[4] (see Fig. 5). This evaluation is carried out to determine how close the instances of the same relation are. The results in Fig. 5 show that the representation $\mathcal{V}^j$ obtained using AAE-based functions minimize the mean intra-cluster distance concerning the others. Within AAE-based functions, *f_aae_laser* obtains the best mean intra-cluster distance value allowing more compact groups.

Fig. 6 compares the representations obtained with *f_laser* (Fig. 6a), *f_ae_laser* (Fig. 6b) and *f_aae_laser* (Fig. 6c) on a subset of instances. The visual representation is achieved using PCA maintaining 3 principal components. This subset has a total of $4,000$ randomly chosen instances, where $2,000$ were taken from */people/person/nationality* relation and the other $2,000$ were taken from the remaining relations. It can be observed how the representations obtained with *f_aae_laser* (Fig. 6c) tend to form 2 clusters, while the representations that used *f_laser* (Figure 6a) and *f_ae_laser* (Fig. 6b) are concentrated in the same region of the embedded manifold.

6.2. *Convenience of using the noise cleaning representation proposed for classification*

To determine the convenience of using the noise reduction approach proposed in this research, the BGWA method was used as the evaluation classifier considering the macro precision measure as evaluation metric (Table 1). We take as a baseline the results of the BGWA on the original NYT2010 ($\mathcal{X}_{train}$). The total number of instances varies for each $\mathcal{X}'_{train}$ obtained set, while the $\mathcal{X}_{test}$ contains $4,801$ instances (labeled

[4]Other distances were evaluated, but the best results were obtained using the cosine distance

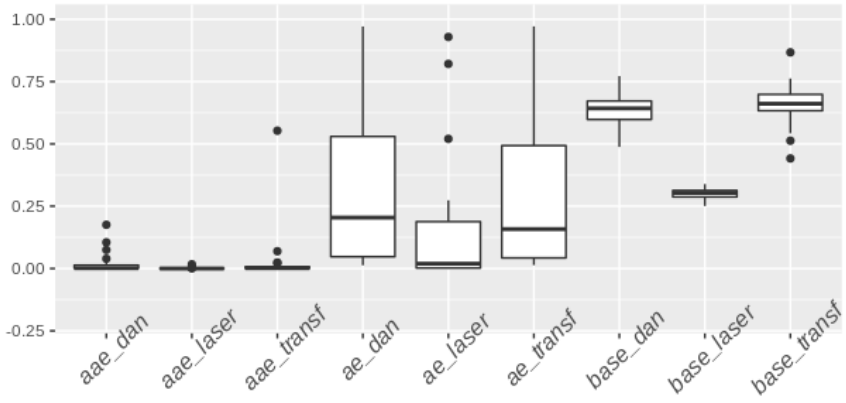
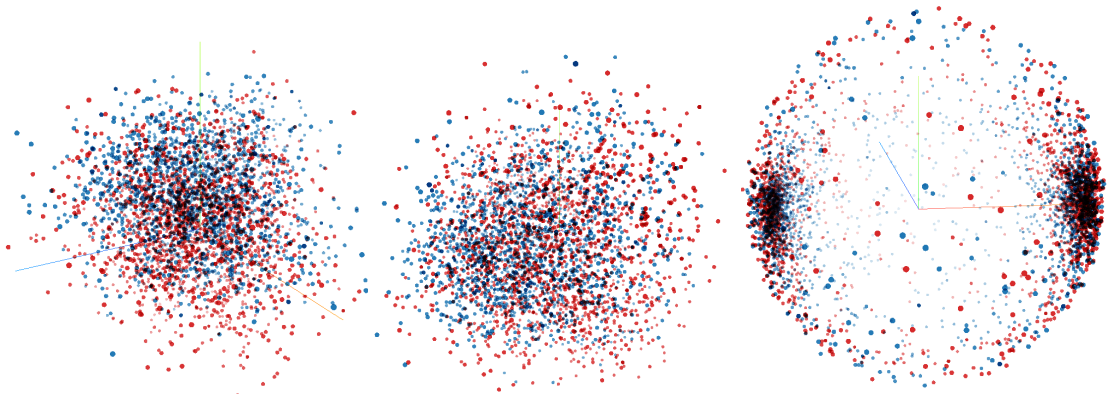


Fig. 5. Intra-cluster distances over each $\mathcal{V}^j$ obtained with $encoder^j$.



(a) *f_laser* (b) *f_ae_laser* (c) *f_aae_laser*

Fig. 6. PCA projections of a instances subset with the generated representations using the functions *f_laser*, *f_ae_laser* and *f_aae_laser*. [**blue**]: instances from */people/person/nationality* relation. [**red**]: other instances. Figure generated with [`https://projector.tensorflow.org/`]

through crowdsourcing using MTurk[5]). The original NYT2010 dataset contains 53 classes, where 30 relations were used due to they exist in each sets not including $\mathcal{NA}$ ($\mathcal{R}^{\dagger}$).

The Anova One Way test was applied on macro precision for determining if there exist significant differences. In this case significant differences were found (Anova: $F(9, 40) = 5.68, p < 4.75e^{-05}$) with an effect size of $\eta^2 = 0.561$. In the pairwise post-hoc comparisons with the Holm Correction [7] and *t-test* we only find significant differences between *base_dan* and *aae_laser*. The results obtained on the $\mathcal{X}'_{train}$ did not show significant differences concerning $\mathcal{X}_{train}$ but it can be taken as an initial set within an iterative process. Despite this, similar results were achieved using only the 77% instances (*aae_laser*) in the $\mathcal{X}_{train}$ set. This indicates that the noise reduction methods did not eliminate instances that significantly affect the BGWA classifier's results for macro precision. In the confusion matrices, it was observed that there are three relations in which most of the errors of the method are concentrated on each set of data. These relations are */location/country/administrative_divisions*, */location/country/capital* and */location/location/contains*, which have common characteristics, like the same type of the pair of entities (locations). Furthermore, these three relations represent 61.4% of the $\mathcal{X}_{test}$ set.

[5]Mechanical Turk, MTurk, is a human annotation service provided by Amazon.

Table 1
Macro precision values after 5 executions of BGWA on each dataset (Cosine distance).

Dataset	Instances	Macro precision
baseline	154929	0.560±0.012
base_dan	85465	0.523±0.011
base_transf	85426	0.541±0.015
base_laser	90308	0.546±0.012
ae_dan	79139	0.540±0.014
ae_transf	78140	0.530±0.018
ae_laser	86680	0.542±0.017
aae_dan	95775	0.537±0.016
aae_transf	91925	0.544±0.010
aae_laser	120024	**0.573±0.008**

Table 2
Macro precision values after 5 executions of BGWA on each dataset (Euclidean distance).

Dataset	Instances	Macro precision
baseline	154929	0.560±0.012
base_dan	84095	0.492±0.021
base_transf	83648	0.478±0.010
base_laser	83351	0.486±0.007
ae_dan	83634	0.466±0.024
ae_transf	81864	0.442±0.020
ae_laser	87216	0.486±0.007
aae_dan	101086	0.492±0.018
aae_transf	103890	0.484±0.010
aae_laser	106218	0.506±0.017

The function $cleaning^j$ considers fewer instances as noise over the representations obtained with *f_aae_laser*. Besides, BGWA obtains higher macro precision values over *aae_laser* with respect to $\mathcal{X}_{train}$ and the others obtained $\mathcal{X}'_{train}$ sets.

From Table 2 it can be noticed that the Cosine distance performs better than the Euclidean distance. This is caused because the macro precision obtained using the Euclidean distance in any dataset outperforms the macro precision of the baseline. The same happens with other similarity distances evaluated.

7. Conclusions

The representations obtained with AAE-based functions allow grouping instances according to their relations in more compact groups. This allows reducing potentially noisy instances in the original dataset.

The proposed noise reduction approach obtains similar macro precision values, considering fewer instances, concerning the original dataset using the BGWA classifier as the evaluator. We only found significant differences between *base_dan* and *aae_laser*. The macro precision obtained over the original dataset was improved using 77% of the instances in this last dataset. This fulfilled the objective proposed in this work. The obtained results verify the importance of using the proposed data representation for noise-cleaning before classifying relations in the DS task. Furthermore, it suggests the usefulness of AAE for obtaining representations for noise reduction without significantly lowering macro precision values.

We are currently working on the AAE architecture to obtain a dataset that improves the resulting measures of the original dataset.

8. Acknowledgments

The present work was supported by CONACyT/México (scholarship 937210 and grant CB-2015-01-257383). Additionally, the authors thank CONACYT for the computer resources provided through the INAOE Supercomputing Laboratory's Deep Learning Platform for Language Technologies. Finally, we would like to thank Dr. Miguel Á. Álvarez-Carmona from CICESE-UT3 for his comments and suggestions.

References

[1] Mikel Artetxe and Holger Schwenk. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019.

[2] Carla E. Brodley and Mark A. Friedl. Identifying Mislabeled Training Data. *Journal of Artificial Intelligence Research*, 11:131–167, 1999.

[3] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal Sentence Encoder. *arXiv:1803.11175v2 [cs.CL]*, page 7, 2018.

[4] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014.

[5] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[7] Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[8] Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. Deep Unordered Composition Rivals Syntactic Methods for Text Classification. In *Proceedings of the 53rd Annual Meeting ofthe Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1681–1691, 2015.

[9] Sharmistha Jat, Siddhesh Khandelwal, and Partha Talukdar. Improving Distantly Supervised Relation Extraction using Word and Entity Based Attention. *arXiv:1804.06987v1 [cs.CL]*, April 2018.

[10] Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting ofthe Association for Computational Linguistics*, pages 2124–2133, Berlin, Germany, 2016. Association for Computational Linguistics.

[11] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial Autoencoders. *arXiv:1511.05644v2 [cs.LG]*, 2015.

[12] N Matic, I Guyon, L Bottou, J. Denker, and V. Vapnik. Computer aided cleaning of large databases for character recognition. In *11th IAPR International Conference on Pattern Recognition. Vol. II. Conference B: Pattern Recognition Methodology and Systems*, pages 330–333. IEEE, 1992.

[13] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the 47th Annual Meeting of the ACL*, pages 1003–1011, Suntec, Singapore, 2009.

[14] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van den Hengel. Deep Learning for Anomaly Detection: A Review. *ACM Computing Surveys*, 1(1):3569–3570, 2020.

[15] Jakub Piskorski and Roman Yangarber. Information extraction: Past, Present and Future. In *Multi-source, Multilingual Information Extraction and Summarization 11*, pages 23–49. Springer-Verlag Berlin Heidelberg, 2013.

[16] Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163, Berlin, 2010. Springer.

[17] Alisa Smirnova and Philippe Cudré-Mauroux. Relation Extraction Using Distant Supervision: A Survey. *ACM Computing Surveys*, 51(5):1–35, November 2018.

[18] Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha Talukdar. Reside: Improving Distantly-Supervised Neural Relation Extraction using Side Information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1257–1266, Brussels, Belgium, 2018. Association for Computational Linguistics.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, and Lukasz Kaiser. Attention Is All You Need. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, pages 5998–6008, 2017.

[20] Zhi-Xiu Ye and Zhen-Hua Ling. Distant Supervision Relation Extraction with Intra-Bag and Inter-Bag Attentions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2810–2819, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[21] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762, Lisbon, Portugal, 2015. Association for Computational Linguistics.

[22] Peng Zhou, Jiaming Xu, Zhenyu Qi, Hongyun Bao, Zhineng Chen, and Bo Xu. Distant supervision for relation extraction with hierarchical selective attention. *Neural Networks*, 108:240–247, 2018.