DiS-ReX: A Multilingual Dataset for Distantly Supervised Relation Extraction

Abhyuday Bhartiya∗ Kartikeya Badola∗ Mausam
Indian Institute of Technology Indian Institute of Technology Indian Institute of Technology
New Delhi, India New Delhi, India New Delhi, India
bhartiyabhuyuday@gmail.com kartikeya.badola@gmail.com mausam@cse.iitd.ac.in

Abstract

Our goal is to study the novel task of distant supervision for multilingual relation extraction (Multi DS-RE). Research in Multi DS-RE has remained limited due to the absence of a reliable benchmarking dataset. The only available dataset for this task, RELX-Distant (Köksal and Özgür, 2020), displays several unrealistic characteristics, leading to a systematic overestimation of model performance. To alleviate these concerns, we release a new benchmark dataset for the task, named DiS-ReX. We also modify the widely-used bag attention models using an mBERT encoder and provide the first baseline results on the proposed task. We show that DiS-ReX serves as a more challenging dataset than RELX-Distant, leaving ample room for future research in this domain.

1 Introduction

Relation Extraction (RE) identifies the relation $r$ between a pair of entities $(e_1, e_2)$ given some text mentioning both of them. To avoid large manual annotation, RE is often trained via distant supervision (DS-RE) (Mintz et al., 2009). DS-RE uses facts $r(e_1, e_2)$ in an existing KB to associate a label $r$ with the bag containing all sentences that mention $e_1$ and $e_2$. Research in DS-RE has been mostly monolingual and limited to English. Our goal is to study multilingual RE via distant supervision (Multi DS-RE). We expect multilingual RE models to have several benefits over monolingual RE. First, training data from multiple languages may be pooled to create a large dataset, enabling cross-lingual knowledge transfer (Zoph et al., 2016; Feng et al., 2020). Second, it may encourage RE models to be consistent across languages (Lin et al., 2017), e.g., extraction of a fact already seen in one language should be easier in another.

To the best of our knowledge, RELX-Distant (Köksal and Özgür, 2020) is currently the only dataset available for Multi DS-RE, but even so, it has never been evaluated as a benchmark for the task. Our analysis reveals that it suffers from a poor selection of relation classes. Firstly, there are no examples of NA class (sentences with no relation between the two entities). Therefore, a model trained on RELX-Distant would find limited utility in any real world setting. Secondly, its choice of relation classes is highly disjoint, resulting in an absence of instances with multiple labels (unusual for a DS-RE dataset). Finally, it is highly imbalanced – even though it has 24 relation classes, over 50% bags belong to just one “country” relation.

Owing to these attributes, we observe that models trained on RELX-Distant end up classifying the instances of the minority class based on just the entity type information. Due to high skew, such mistakes have negligible impact on evaluation scores and the model achieves an AUC of 0.99 after only 5 training epochs. Such numbers are unheard of, especially when compared to benchmarking datasets in mono-lingual RE (mono-lingual variant of the same architecture obtains an AUC of 0.83 when trained and tested on the GDS dataset (Jat et al., 2018).

In response, we contribute a more realistic benchmark dataset for the task called DiS-ReX. Our dataset has over 1.8 million sentences in four languages: English, French, Spanish and German. It has 37 relation types including 1 No-Relation (NA) class and also has instances with multiple labels similar to the widely-used New York Times (NYT) dataset for English DS-RE (Riedel et al., 2010), thus comparing favorably to RELX-Distant.

We also adopt state-of-the-art DS-RE models in the multilingual setting by using the mBERT encoder (Devlin et al., 2019), to create a strong baseline for this task.

We achieve an AUC of 0.82 and a Micro-F1 of 0.76, suggesting that the dataset is not trivial to optimize on, and could act as a good benchmark.
for the task. We publicly release DiS-ReX and the baseline.¹

2 Related Work

Supervised RE datasets such as ACE05 (Walker et al., 2006) and KLUE (Park et al., 2021) are generally small, owing to the supervision needs per relation. Distant supervision (Mintz et al., 2009) is a popular alternative to large-scale human annotation, but necessitate more complex models to handle dataset noise. The standard English DS-RE dataset is New York Times (NYT) corpus (Riedel et al., 2010), which has served as the benchmark for research over the years. DS-RE models have evolved to use multi-instance learning (Hoffmann et al., 2011), multi-label learning (Surdeanu et al., 2012), corrections for false negatives (Ritter et al., 2013), and neural models such as piecewise CNNs (Zeng et al., 2015), intra-bag attention (Lin et al., 2016), and reinforcement learning (Qin et al., 2018).

Lin et al. (2017) and Wang et al. (2018) propose extensions of bag-attention models for bilingual (English-Mandarin) datasets. However, their adoption to multiple languages has been lacking, due to absence of a reliable multilingual dataset. Although RELX-Distant is the only Multilingual DS-RE dataset so far, it wasn’t originally used for Multi DS-RE task but to pre-train a model that gets fine-tuned for supervised RE task.

Contemporary to our work, other multilingual RE datasets and methods are being developed. These include a dataset for joint entity and relation extraction (Seganti et al., 2021), a model for multilingual KB completion (Singh et al., 2021), and an approach for automatic construction of cross-lingual training data for Open IE (Kolluru et al., 2022). Our proposed dataset, DiS-ReX, has already been used for further research on the Multilingual DS-RE task (Rathore et al., 2022).

3 Dataset Curation

All distant supervision datasets are curated by aligning known KB facts with sentences in a large corpus. We follow the same for DiS-ReX, while paying attention to cross-lingual normalization, and overall data and language statistics.

First, we harvest a large number of sentences from English, French, Spanish and German Wikipedias.² We use DBPedia language editions (Lehmann et al., 2015) for our KB – this gives us good coverage of entities that are local to different language speakers. DBPedia entities are associated with Wikidata IDs, which are normalized across languages. This enables us to fuse these DBPedia KBs and establish equivalence between entities like USA and Estados Unidos de América.

Next, we use a language-specific NER tagger, (we use the md variant of spaCy (Honnibal et al., 2020) NER taggers for each language), returning a rich set of sentences. In contrast, RELX-Distant finds entity mentions using Wikipedia hyperlinks. This severely limits its pool of sentences, since often only the first mention of an entity in a Wiki document has a hyperlink while others do not.

Linking each mention with its entity can be challenging, due to unavailability of high-quality entity linking software for every language. We take the pragmatic approach of using simple string matching, but only on the subset of entities that have an unambiguous surface form (or alias) in our fused KB. This maintains scalability to many languages, while ensuring high enough precision of linking.

For each entity-pair, we create a language-specific bag of all sentences that mention both. We also search for all relations between them in our fused KB. We associate the bag with all those relation labels, or “NA”, if no relation is found.

Our next steps select a balanced subset of this dataset, so that it can serve as a good benchmark for Multi DS-RE. We first select the subset of relations that have at least 50 bags in all languages.

---

¹https://github.com/dair-iitd/DiS-ReX

²Our pipeline applies to non-Wikipedia sentences too.
This yields the 36 positive relation types used in our data. For each relation type, we limit the number of bags in a language to a max of 10,000. This helps curb the skew due to highly frequent relations such as country and birthPlace. During this filtering, we ensure that bags of entity pairs common across more than one language are not removed, so that we have an abundant number of cross-lingual bags. Models can take advantage of such bags for establishing representation consistency across languages (Wang et al., 2018). Finally, we add bags of entity pairs that have no relation between them. Similar to NYT dataset, “NA” is the majority class in DiS-ReX (kept at roughly 70%).

Hence, we obtain a dataset with over 1.8 million sentences, and over 250,000 (non-NA) bags (see table 1 for more statistics). The 36 relations include frequent relations between persons, locations and organizations (e.g., capital, headquarter, works-at), and also some relations with fine-grained types such as bandMember, starring and recordLabel.

We estimate the percentage of bags satisfying “at-least one” assumption by manually labelling sentences across 50 randomly selected bags. We find that 82% of the bags satisfy “at-least one” assumption. For the test set of NYT Corpus, this percentage is close to 62% (Zhu et al., 2020).

Finally, we create train-dev-test splits by splitting the bags in the ratio 70 : 10 : 20. While splitting we ensure that entity-pairs in three sets are mutually exclusive, so the model does not extract by memorizing a fact.

4 Experiments and Data Analysis

4.1 Comparison: DiS-ReX vs. RELX-Distant

We now compare the two datasets: DiS-ReX and RELX-Distant. We find that the our dataset showcases several desirable properties expected from a challenging DS-RE dataset, including the presence of NA relations, inverse relations, multi-label bags, and better class balance.

70% of bags in DiS-ReX are NA bags, whereas RELX-Distant has none. We also note that a few relation pairs (from our 36 relations) represent inverses of each other, e.g., {influenced by, influenced}, {successor, predecessor}, and {associatedBand, bandMember}. Inverse relations test an extractor’s ability to output related relations from the same bag, but with different entity ordering. RELX-Distant has no inverse relations in its relation vocabulary.

| RELX-Distant | DiS-ReX |
|--------------|---------|
| Efficiency ($\eta$) | 0.522 0.856 |
| M-F1 (top 3) | 94.29 82.06 |
| M-F1 (bottom 3) | 49.47 63.28 |

Table 2: Key statistics representing class imbalance between RELX-Distant and DiS-ReX

| Lang. | RELX-Distant | DiS-ReX |
|-------|--------------|---------|
| AUC  | &micro;F1 | M-F1 | AUC | &micro;F1 | M-F1 |
| English | 0.99 | 0.95 | 0.78 | 0.78 | 0.71 | 0.69 |
| French | 0.99 | 0.96 | 0.79 | 0.81 | 0.75 | 0.68 |
| Spanish | 0.98 | 0.94 | 0.77 | 0.80 | 0.73 | 0.66 |
| German | 0.99 | 0.95 | 0.80 | 0.76 | 0.72 | 0.59 |
| All | 0.99 | 0.95 | 0.79 | 0.81 | 0.74 | 0.68 |

Table 3: Language-wise performance of mBERT + Att. &micro;F1 and M-F1 refer to micro and macro F1 scores.

A key characteristic of DS-RE problems is that they need multi-label modeling (Surdeanu et al., 2012), since multiple relations commonly exist between an entity pair. RELX-Distant has no such bags, primarily because its choice of relation types is such that almost no entity-pair can have multiple relations. E.g., its Person-Person relations are mother, spouse, father, sibling, partner, where multi-label bags are highly unlikely. In contrast, DiS-ReX has 21,642 bags that have more than one relation label. As an example, the entity pair (Isaac Newton, England) is associated with four relations – birthPlace, country, deathPlace and nationality.

To compare the imbalance amongst non-NA relation classes in DiS-ReX and RELX-Distant, we calculate normalized entropy (Shannon, 1948), also known an Efficiency ($\eta$). Value closer to 1 indicates that the class-wise distribution is closer to the uniform distribution. Results in Table 2 indicate that DiS-ReX is a more balanced dataset (more details regarding calculation of $\eta$ in appendix).

4.2 Baseline Performance

We implement three DS-RE baselines for our DiS-ReX dataset. Our first baseline is PCNN+Att (Lin et al., 2016), which uses a piece-wise CNN as the sentence encoder and performs bag-level multi-label classification using Intra-Bag attention. In this model, each language is trained and tested upon separately. Inspired by Ni and Florian (2019), we extend this to design a second baseline,
mBERT+Att. It replaces PCNN encoders with a shared mBERT encoder (Devlin et al., 2019) and retains the intra-bag attention architecture for constructing the bag representation. Our last baseline is mBERT+MNRE, which adapts the MNRE model (Lin et al., 2017) to our setting. MNRE introduced cross-lingual attention for bilingual RE. We extend this attention module to more than two languages and also replace its language-specific CNN encoders with a shared mBERT encoder. More details on baselines and training are in appendix.

We first compare mBERT+Att model on both DiS-ReX and RELX-Distant in Table 3. We find that RELX-Distant achieves an unreasonably high AUC and micro-F1. Since Micro-F1 may be overwhelmed by a few highly frequent relations, we also report Macro-F1 scores. Even the Macro-F1 scores of RELX-Distant are over 10 pt higher, suggesting that DiS-ReX is a more challenging dataset for our task. We also report the Macro-avg of F1 scores of 3 most frequent and 3 least frequent classes of both the datasets in Table 2. The performance drops by 45pts in RELX-Distant, more than double the decrease observed in our dataset, corroborating that the RELX-Distant model is not learning infrequent relations effectively. For that model, we notice that the person-person relation types, which are minority classes, obtain the lowest F1 scores. It gets confused between mother and spouse or between father and sibling. In some cases, the confidence is as high as 95% on such errors. This suggests that the model is making predictions based solely on head-tail entity types in instances belonging to the person-person relation classes. But, such mistakes depress the Micro-F1 and AUC scores only negligibly, due to severe class imbalance. Thus, the high scores do not reflect high model quality.

We report results of three models on DiS-ReX in Table 4 – mBERT+MNRE achieves 0.82 AUC and 0.76 micro-F1, establishing the best baseline performance on our task.

### 4.3 Error Analysis

We find that due to incorporation of NA class, multi-label bags and fine-grained relation classes, DiS-ReX offers several new challenges. We observe that on multi-label bags, micro-F1 falls drastically from roughly 0.84 (bags with 1 label) to 0.35 (4 labels), primarily due to reducing recall (statistics in Table 5).

| #relations | Micro-F1 | Precision | Recall |
|------------|----------|-----------|--------|
| 1          | 0.842    | 0.865     | 0.820  |
| 2          | 0.673    | 0.934     | 0.525  |
| 3          | 0.518    | 0.959     | 0.354  |
| 4          | 0.348    | 0.937     | 0.214  |

Table 5: Comparing performance of mBERT+MNRE on entity pairs with different number of labels in the ground truth in the DiS-ReX dataset

We also perform manual error analysis of 100 random and 100 most confident mistakes made by the model trained on DiS-ReX. For errors where a non-NA relation is incorrectly predicted as another, we find one major error class – highly confident mistakes in predicting closely related relation types that have high overlaps, such as \{author, director\}, and \{homeTown, birthPlace\}. Some model errors correspond to confusion in predicting inverse relations such as \{successor, predecessor\} and \{influenced, influencedBy\}. Such cases are absent in the RELX-Distant test set. We found less than 10% errors within the confident errors are due to entity disambiguation mistakes in ground truth, however, we found no such data error in the 100 random errors, suggesting that this failure mode is not the most frequent, and the test data is relatively clean.

We additionally divide the errors made on the entire test set by the best performing model into three variants.

- **Type-1 Error**: Model predicts a positive (Non-NA) relation label \(R_1\) and ground label is also a positive (non-NA) relation label \(R_2\) but \(R_2\) is not the same as \(R_1\).

- **Type-2 Error**: Model predicts NA relation label but ground label is a positive (non-NA) relation label.

- **Type-3 Error**: Model predicts positive (non-NA) relation label but ground label is NA relation label.

We present the distribution of these three errors in Table 6. Predicting non-NA as NA and NA as non-NA relation make up most (55-85%) of the errors. We believe that eliminating such kinds of errors would be an important focus area in DS-RE research, especially for datasets which are better representative of real world settings.
| Language | Type-1 Error (%) | Type-2 Error (%) | Type-3 Error (%) |
|----------|------------------|------------------|------------------|
| English  | 44.49            | 31.17            | 24.33            |
| French   | 29.69            | 36.14            | 34.15            |
| Spanish  | 35.08            | 36.37            | 28.54            |
| German   | 14.94            | 45.28            | 39.77            |

Table 6: Types of Errors made in different languages for mBERT+MNRE on DiS-ReX

4.4 Is mBERT+Att Language Agnostic?

It is believed that sharing mBERT encoder across languages is advantageous for cross-lingual transfer (Wu and Dredze, 2019). This is reflected in our experiments too where mBERT+Att strongly outperforms PCNN+Att.

mBERT+Att produces a single embedding for a multilingual bag, summarizing mBERT embeddings of individual sentences. We posit that for this model to achieve its true potential on DiS-ReX, mBERT encoder must learn to map all sentences to a language-agnostic representation space, or else the downstream bag attention model may get confused between intra-language and inter-language variability. We investigate this further by raising the question: is the mBERT encoder learning language agnostic embeddings?

For this we encode all sentences in multilingual bags (that contain all languages) using the encoder of trained mBERT+Att model and plot the sentence embeddings using tSNE. We show an illustrative figure for the bag (Swiss, Switzerland) in Figure 1. We find that mBERT clusters sentences of one language together, irrespective of their content (more figures in Appendix). This suggests that mBERT embeddings strongly retain language information, and are not language-agnostic.

This may prove to be a significant obstacle towards progress on our task, since the noise-filtering intra-bag attention may end up capturing variance across languages more than variance in semantics. This may also explain why mBERT+MNRE performs better, since it generates embeddings of sub-bags of each language separately, instead of a single embedding for a multilingual bag.

5 Conclusion

We propose DiS-ReX, a novel dataset for Multi DS-RE in 4 languages. We show that it is a more realistic and challenging benchmark compared to the existing dataset. DiS-ReX has a fairly well-represented distribution of relation types, includes instances with no-relation between entity-pairs and the relation-types selected show several real-world characteristics like inverse relations, different relations with high overlap, etc. We also publish first baseline numbers on the task of Multi DS-RE by extending existing state-of-the-art models. A detailed analysis of model performance suggests several research challenges for future: (1) learning language-agnostic sentence embeddings, (2) robustness to related relations (inverse; overlapping but semantically different), and (3) handling multi-label entity-pairs. Recently, Rathore et al. (2022) develop a multilingual DS-RE model named PARE, which reports improved performance on the DiS-ReX dataset.

Acknowledgements

This work is primarily supported by a grant from Huawei. It is also supported by grants from Google, Jai Gupta Chair professorship and a Visvesvaraya faculty award by the Govt. of India to Mausam. We thank IIT Delhi HPC facility for compute resources. We thank Vipul Rathore for his useful feedback on evaluating the quality of the dataset, and Vipul Rathore and Keshav Kolluru for their helpful comments on an earlier draft of the paper.

References

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), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfie Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic sentence embedding. arXiv preprint arXiv:2007.01852.

Xu Han, Tianyu Gao, Yuan Yao, Deming Ye, Zhiyuan Liu, and Maosong Sun. 2019. OpenNRE: An open and extensible toolkit for neural relation extraction. In Proceedings of EMNLP-IJCNLP: System Demonstrations, pages 169–174.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, pages 541–550.

Matthew Honnibal, Ines Montani, Sofie Van Lardeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Sharmistha Jat, Siddhesh Khandelwal, and Partha Talukdar. 2018. Improving distantly supervised relation extraction using word and entity based attention. arXiv preprint arXiv:1804.06987.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Abdullatif Köksal and Arzucan Özgür. 2020. The RELX dataset and matching the multilingual blanks for cross-lingual relation classification. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 340–350, Online. Association for Computational Linguistics.

Keshav Kolluru, Mohammed Muqeeth, Shubham Mittal, Soumen Chakrabarti, and Mausam. 2022. Alignment-Augmented Consistent Translation for Multilingual Open Information Extraction. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland. Association for Computational Linguistics.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia—a large-scale, multilingual knowledge base extracted from wikipedia. Semantic web, 6(2):167–195.

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), pages 2124–2133.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In International Conference on Learning Representations.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pages 1003–1011.

Jian Ni and Radu Florian. 2019. Neural cross-lingual relation extraction based on bilingual word embedding mapping. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 399–409, Hong Kong, China. Association for Computational Linguistics.

Sungjoon Park, JiHyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, JooHong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seonbgo Jang, Seungwon Do, Sunkyong Kim, Kyungtae Lim, Jongwon Lee, Kyeunim Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jungwoo Ha, and Kyunghyun Cho. 2021. KLUE: Korean language understanding evaluation. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

Pengda Qin, Weiran Xu, and William Yang Wang. 2018. Robust distant supervision relation extraction via deep reinforcement learning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, pages 2137–2147. Association for Computational Linguistics.

Vipul Rathore, Kartikeya Badola, Parag Singla, and Mausam. 2022. PARE: A simple and strong baseline for monolingual and multilingual distantly supervised relation extraction. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 148–163. Springer.
Alan Ritter, Luke Zettlemoyer, Mausam, and Oren Etzioni. 2013. Modeling missing data in distant supervision for information extraction. *Trans. Assoc. Comput. Linguistics*, 1:367–378.

Alessandro Seganti, Klaudia Firlag, Helena Skowronska, Michał Satława, and Piotr Andruszkiewicz. 2021. Multilingual entity and relation extraction dataset and model. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1946–1955, Online. Association for Computational Linguistics.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Harkanwar Singh, Soumen Chakrabarti, Prachi Jain, Sharod Roy Choudhury, and Mausam. 2021. Multilingual knowledge graph completion with joint relation and entity alignment. In *3rd Conference on Automated Knowledge Base Construction, AKBC 2021, Virtual, October 4-8, 2021*.

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*, pages 2895–2905.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 455–465.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus.

Xiaozhi Wang, Xu Han, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2018. Adversarial multi-lingual neural relation extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1156–1166.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. 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.

Tong Zhu, Haitao Wang, Junjie Yu, Xiabing Zhou, Wenliang Chen, Wei Zhang, and Min Zhang. 2020. Towards accurate and consistent evaluation: A dataset for distantly-supervised relation extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6436–6447, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.
Appendix

B Calculation of Efficiency

For a dataset of size \( n \) over \( k \) classes, where \( i^{th} \) class has \( n_i \) instances:

\[
Efficiency = - \sum_{i=1}^{k} \frac{n_i \log n_i}{\log k}
\]

Efficiency lies between 0 and 1. A higher value suggests that the class-distribution is closer to uniform.

C Baseline architecture

C.1 BERT Encoder

To obtain a distributed representation of a sentence \( x \), we use mBERT. In order to encode positional information into the model we use Entity Markers scheme introduced by (Soares et al., 2019). We add special tokens \([E1], [E1]\) to mark start and end of the head entity and \([E2], [E2]\) to mark start and end of the tail entity. This modified sentence is fed into a pretrained BERT model and the output head and tail tokens are concatenated to get the final sentence representation \( \tilde{x}^j_i \) for each sentence \( x^j_i \) in our bag.

C.2 Intra Bag Attention

To obtain representation of bag \( B \), we apply selective sentence-level attention (Lin et al., 2016). We obtain real-valued vector \( \tilde{B} \) for the bag as a weighted sum of sentence representations \( \tilde{x}^j_i \):

\[
\tilde{B} = \sum_{i,j} \alpha^j_i \cdot \tilde{x}^j_i
\]

where \( \alpha^j_i \) measures attention score of \( \tilde{x}^j_i \) with a specific relation \( r \):

\[
\alpha^j_i = \frac{\exp(\tilde{x}^j_i \cdot r)}{\sum_{k,l} \exp(\tilde{x}^k_l \cdot r)}
\]

This reduces the effect of noisy labels on the final bag representation.

Finally, we obtain conditional probability \( p(r|B, \theta) = \text{softmax}(o) \). Here we obtain \( o \) which represents scores for all relation types.

\[
o = RB + d
\]

\( R \) is the matrix of relation representations. Our objective function is the cross-entropy loss and is defined as follows:

\[
L(\theta) = \sum_{i=1}^b p(r_i|B_i, \theta)
\]

where \( b \) denotes the number of bags in our training data

C.3 MNRE and Cross-Lingual Attention

In order to extend the Intra Bag Attention to multilingual setting, (Lin et al., 2017) introduce separate relation embeddings for each language and propose creating several representations of a bag by taking attention of sentences in language \( j \) with relation embedding of language \( k \). Formally, the cross-lingual representation \( S_{jk} \) is defined as a weighted sum of those sentence vectors \( \tilde{x}^j_i \) in the \( j^{th} \) language where \( \alpha^j_{ik} \) is the attention score of each sentence with respect to the \( k^{th} \) language.

\[
S_{jk} = \sum_{i} \alpha^j_{ik} \cdot \tilde{x}^j_i
\]

\[
\alpha^j_{ik} = \frac{\exp(\tilde{x}^j_i \cdot r_k)}{\sum_{j,l} \exp(\tilde{x}^j_l \cdot r_k)}
\]

\[
o = (R_k + M)S_{jk} + d
\]

\( R_k \) is the matrix of relation representations \( (r_k) \) in language \( k \) and \( M \) is a global relation matrix initialized randomly. Similar to (Lin et al., 2016), probability \( p(r|S_{jk}, \theta) = \text{softmax}(o) \). To obtain score of relation \( r \) for bag \( B \):

\[
f(B, r) = \sum_{jk} \log p(r|S_{jk}, \theta)
\]

Loss function is negative log likelihood over all bags in the dataset.
### D Training details

For training we use AdamW optimizer (Kingma and Ba, 2017; Loshchilov and Hutter, 2019), with \( lr=0.001, \text{betas}=(0.9, 0.999), \text{eps}=1e^{-08} \). Weight decay is 0.01 for all parameters except bias and layer norm parameters. Hyperparameters were selected using manual tuning on the dataset. We train the mBERT models for 5 epochs and the PCNN+Att model for 60 epochs. We follow the framework of OpenNRE (Han et al., 2019) and select bag size = 2 for all models. For testing, we choose the weights with best validation AUC. Correct prediction of NA class is not counted in the calculation of Micro F1 and AUC. We use a single Tesla V100 32 GB GPU for all of our experiments.

mBERT+MNRE baseline takes 8 hours for 1 epoch. mBERT+Att takes 3 hours for 1 epoch. PCNN+Att takes 3 hours for 60 epochs.

Training, validation and testing splits for both DiS-ReX and RELX-Distant are in the ratio of 7:1:2. We made sure that the bags in testing set do not overlap with the bags in the training set.

### E Detailed Statistics of mBERT Baselines

In Table 7, we present results on all languages for our three baselines on DiS-ReX. In tables 8, 9 , we present the distribution of errors made by the mBERT+Att and mBERT+MNRE models

In Table 10 and 11, we present the results on bags having 1,2,3 and 4 labels in ground truth using mBERT+Att and mBERT+MNRE respectively.

In Table 12, we present the results on all classes of the best baseline model (mBERT+MNRE) when run on our DiS-ReX dataset.

| Language | Type-1 Error (%) | Type-2 Error (%) | Type-3 Error (%) |
|----------|------------------|------------------|------------------|
| English  | 43.44            | 26.66            | 29.90            |
| French   | 29.73            | 30.45            | 39.82            |
| Spanish  | 33.82            | 30.61            | 35.57            |
| German   | 15.03            | 39.60            | 45.37            |

Table 8: Types of Errors made in different languages for mBERT+Att

| Language | Type-1 Error (%) | Type-2 Error (%) | Type-3 Error (%) |
|----------|------------------|------------------|------------------|
| English  | 44.49            | 31.17            | 24.33            |
| French   | 29.69            | 36.14            | 34.15            |
| Spanish  | 35.08            | 36.37            | 28.54            |
| German   | 14.94            | 45.28            | 39.77            |

Table 9: Types of Errors made in different languages for mBERT+MNRE
| Number of relation labels | Micro-F1 | Precision | Recall |
|---------------------------|----------|-----------|--------|
| 1                         | 0.842    | 0.865     | 0.820  |
| 2                         | 0.673    | 0.934     | 0.525  |
| 3                         | 0.518    | 0.959     | 0.354  |
| 4                         | 0.348    | 0.937     | 0.214  |

Table 11: Comparing performance of mBERT+MNRE on entity pairs with different number of labels in the ground truth
| Relation Label   | F1  | Precision | Recall |
|------------------|-----|-----------|--------|
| predecessor      | 67.58 | 76.31 | 60.65 |
| nationality      | 67.29 | 64.68 | 70.12 |
| artist           | 76.78 | 74.79 | 78.87 |
| region           | 81.43 | 81.14 | 81.73 |
| department       | 95.08 | 95.28 | 94.88 |
| successor        | 72.16 | 75.32 | 69.26 |
| location         | 69.82 | 65.36 | 74.93 |
| bandMember       | 73.45 | 73.45 | 73.45 |
| isPartOf         | 66.50 | 59.52 | 75.33 |
| hometown         | 73.03 | 70.14 | 76.17 |
| previousWork     | 68.83 | 64.89 | 73.27 |
| riverMouth       | 72.63 | 78.97 | 67.24 |
| team             | 81.66 | 85.85 | 77.86 |
| recordLabel      | 86.85 | 87.24 | 86.46 |
| associatedBand   | 71.26 | 61.69 | 84.36 |
| author           | 78.87 | 83.30 | 74.88 |
| influenced       | 61.35 | 65.81 | 57.46 |
| birthPlace       | 75.00 | 75.52 | 74.48 |
| formerBandMember | 57.94 | 59.62 | 56.36 |
| leaderName       | 71.16 | 70.97 | 71.35 |
| deathPlace       | 66.24 | 64.15 | 68.46 |
| city             | 78.96 | 81.93 | 76.19 |
| province         | 78.82 | 78.73 | 78.92 |
| influencedBy     | 59.29 | 65.26 | 54.32 |
| locationCountry  | 62.58 | 64.76 | 60.55 |
| related          | 75.94 | 74.35 | 77.59 |
| director         | 83.59 | 79.36 | 88.29 |
| capital          | 53.68 | 48.69 | 59.82 |
| largestCity      | 65.89 | 71.57 | 61.04 |
| NA               | 95.08 | 95.56 | 94.61 |
| country          | 86.57 | 85.77 | 87.39 |
| starring         | 86.32 | 86.52 | 86.12 |
| subsequentWork   | 71.65 | 70.23 | 73.12 |
| producer         | 53.30 | 51.20 | 55.58 |
| headquarter      | 68.54 | 66.08 | 71.18 |
| state            | 82.54 | 78.32 | 87.26 |
| locatedInArea    | 72.23 | 70.44 | 74.10 |
| All relations    | 70.67 | -     | -     |

Table 12: Class-wise performance scores for MNRE (our best performing model)

G Qualitative Analysis

In this section, we give some examples of randomly selected non NA instances in our dataset:

**English:**

- **Sentence:** another dialect spoken in tioman island is a distinct malay variant and most closely related to riau archipelago malay subdialect spoken in natuna and anambas islands in the south china sea together forming a dialect continuum between the bornean malay with the mainland malay

**Entities:** (tioman island, the south china sea)
• **Sentence:** in 2017 jenny durkan was elected as the first openly lesbian mayor of seattle
  **Entities:** (jenny durkan, seattle)
  **Relations:** http://dbpedia.org/ontology/birthPlace

German:

• **Sentence:** danach kamen abgeleitete klassen hinzu ein strengeres typsystem und während stroustrup "c with classes" ("c mit klassen") entwickelte woraus später c++ wurde schrieb er auch cfront einen compiler der aus c with classes zunächst c-code als erzeugte
  **Entities:** (c,c++)
  **Relations:** http://dbpedia.org/ontology/influenced

• **Sentence:** früher auch ur ist ein 96.1 km langer nebenfluss der sauer entlang der grenze von deutschland zu den westlichen nachbarstaaten belgien und luxemburg
  **Entities:** (sauer, deutschland)
  **Relations:** http://dbpedia.org/ontology/locatedInArea

French:

• **Sentence:** à la mort de boleslas v le pudique duc princeps de pologne la guerre civile en mazovie empêche conrad de revendiquer le trône de cracovie
  **Entities:** (boleslas v le pudique, cracovie)
  **Relations:** http://dbpedia.org/ontology/deathPlace

• **Sentence:** les entreprises masson masson est le dirigeant effectif des trois entreprises du groupe cette situation se reflète désormais dans l actionnariat et les raisons sociales des sociétés qui deviennent joseph masson sons and company (montréal) masson langevin sons and company (québec) masson sons and company (glasgow) cette dernière société basée en écosse a surtout vocation de gérer les achats
  **Entities:** (joseph masson, québec)
  **Relations:** http://dbpedia.org/ontology/birthPlace

Spanish:

• **Sentence:** en 2003 apareció en anything else película de woody allen junto a christina ricci y jason biggs además actuó en la película para televisión l
  **Entities:** (anything else, jason biggs)
  **Relations:** http://dbpedia.org/ontology/starring

• **Sentence:** es una comuna y población de francia en la región de borgoña departamento de yonne en el distrito de sens y cantón de sens-ouest
  **Entities:** (sens, yonne)
  **Relations:** http://dbpedia.org/ontology/department

H Additional Dataset Statistics

In Table 13, we present the number of bags common across 2,3 and all 4 languages. In table 14 and 15, we present the number of bags and sentences in each class on all 4 languages in our dataset. In figure 3 we present a histogram depicting number of bags present for each relation class.
### Table 13: Number of bags common across 2, 3 and all languages

| Number of languages | Number of Bags |
|---------------------|----------------|
| 2                   | 59709          |
| 3                   | 9494           |
| 4                   | 1488           |

Figure 3: Number of bags vs relation class in DiS-ReX (all languages combined)
| Relation Label       | English | French | German | Spanish | All languages |
|----------------------|---------|--------|--------|---------|---------------|
| NA                   | 149874  | 142467 | 149034 | 148806  | 590181        |
| isPartOf             | 2548    | 645    | 465    | 490     | 4148          |
| state                | 1882    | 1762   | 3537   | 429     | 7610          |
| largestCity         | 265     | 342    | 199    | 393     | 1199          |
| birthPlace          | 7861    | 9532   | 3341   | 9484    | 30218         |
| deathPlace          | 4377    | 5629   | 277    | 4709    | 14992         |
| nationality          | 2205    | 4413   | 143    | 2265    | 9026          |
| country              | 10024   | 9618   | 3065   | 9808    | 32515         |
| capital              | 544     | 651    | 397    | 891     | 2483          |
| city                 | 1415    | 4257   | 7930   | 1844    | 15446         |
| author               | 1483    | 1224   | 94     | 460     | 3261          |
| previousWork         | 348     | 696    | 305    | 1127    | 2476          |
| location             | 5655    | 1300   | 1180   | 1685    | 9820          |
| riverMouth           | 464     | 880    | 3303   | 154     | 4801          |
| locatedInArea        | 1324    | 785    | 5715   | 608     | 8432          |
| hometown             | 1689    | 435    | 163    | 4474    | 6761          |
| successor            | 1574    | 2959   | 74     | 1618    | 6225          |
| influenced           | 820     | 453    | 61     | 188     | 1522          |
| headquarter          | 1122    | 922    | 460    | 1895    | 4399          |
| province             | 225     | 1121   | 1272   | 2405    | 5023          |
| associatedBand       | 3669    | 384    | 107    | 2555    | 6715          |
| subsequentWork       | 390     | 760    | 344    | 1248    | 2742          |
| locationCountry      | 925     | 799    | 2237   | 361     | 4322          |
| bandMember           | 1327    | 1909   | 300    | 3092    | 6628          |
| director             | 1258    | 3003   | 1592   | 2089    | 7942          |
| team                 | 1329    | 564    | 461    | 634     | 2988          |
| artist               | 1188    | 3891   | 1241   | 2670    | 8990          |
| related              | 1439    | 375    | 117    | 6262    | 8193          |
| producer             | 1381    | 2848   | 1401   | 3044    | 8674          |
| predecessor          | 475     | 2814   | 81     | 273     | 3643          |
| leaderName           | 353     | 236    | 270    | 223     | 1082          |
| formerBandMember     | 960     | 1153   | 174    | 1345    | 3632          |
| recordLabel          | 791     | 881    | 199    | 2107    | 3978          |
| region               | 1529    | 3673   | 1907   | 2249    | 9358          |
| influencedBy         | 954     | 533    | 86     | 291     | 1864          |
| starring             | 3040    | 7018   | 3087   | 4179    | 17324         |
| department           | 99      | 5486   | 323    | 3157    | 9065          |
| All relations        | 216806  | 226418 | 194942 | 229512  | 876743        |

Table 14: Comprehensive bag-wise statistics of the dataset
| Relation Label          | English | French | German | Spanish | All languages |
|------------------------|---------|--------|--------|---------|---------------|
| isPartOf               | 16085   | 2794   | 2566   | 1880    | 23325         |
| state                  | 11979   | 13135  | 13705  | 1405    | 40224         |
| largestCity            | 18811   | 4163   | 8949   | 3136    | 35059         |
| birthPlace             | 15738   | 16624  | 4376   | 14359   | 51097         |
| deathPlace             | 11498   | 12208  | 539    | 8888    | 33133         |
| nationality            | 5848    | 9560   | 219    | 4330    | 19957         |
| country                | 88787   | 43911  | 13148  | 64660   | 210506        |
| capital                | 19887   | 4713   | 17227  | 5318    | 47145         |
| city                   | 4490    | 11156  | 23631  | 3740    | 43017         |
| author                 | 3387    | 4121   | 335    | 1417    | 9260          |
| previousWork           | 6507    | 1276   | 450    | 2318    | 10551         |
| location               | 15538   | 4757   | 4656   | 6014    | 30965         |
| riverMouth             | 1172    | 2442   | 12467  | 420     | 16501         |
| locatedInArea          | 4320    | 4152   | 18890  | 1904    | 29266         |
| hometown               | 7648    | 796    | 1067   | 8971    | 18482         |
| successor              | 4700    | 6963   | 128    | 3118    | 14909         |
| influenced             | 2416    | 1147   | 635    | 394     | 4592          |
| headquarter            | 5419    | 2399   | 2030   | 5736    | 15584         |
| province               | 1082    | 2472   | 2710   | 11672   | 17936         |
| associatedBand         | 7390    | 713    | 136    | 8437    | 16676         |
| subsequentWork         | 6541    | 1318   | 517    | 2526    | 10902         |
| locationCountry        | 3204    | 2836   | 8226   | 1229    | 15495         |
| bandMember             | 3592    | 5910   | 475    | 8763    | 18740         |
| director               | 2005    | 7811   | 2970   | 3961    | 16747         |
| team                   | 1830    | 814    | 694    | 1396    | 4734          |
| artist                 | 2893    | 9591   | 3156   | 6472    | 22112         |
| related                | 4526    | 928    | 171    | 17432   | 23057         |
| producer               | 2459    | 6398   | 2647   | 6384    | 17888         |
| predecessor            | 2592    | 7003   | 162    | 600     | 10357         |
| leaderName             | 1549    | 1074   | 452    | 448     | 3523          |
| formerBandMember       | 2975    | 3452   | 279    | 4091    | 10797         |
| recordLabel            | 1320    | 1214   | 219    | 4149    | 6902          |
| region                 | 5836    | 11860  | 5901   | 4485    | 28082         |
| influencedBy           | 2524    | 1482   | 913    | 536     | 5455          |
| starring               | 4484    | 14578  | 4616   | 6676    | 30354         |
| department             | 196     | 15807  | 693    | 4997    | 21693         |
| All relations          | 532499  | 409087 | 438315 | 456418  | 1858012       |

Table 15: Comprehensive sentence-wise statistics of the dataset