Separating Retention from Extraction in the Evaluation of End-to-end Relation Extraction

Bruno Taillé[1], **Vincent Guigue**[1], **Geoffrey Scoutheeten**[2] **and Patrick Gallinari**[1,3]

[1]Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, LIP6

[2]BNP Paribas [3]Criteo AI Lab

{bruno.taille, vincent.guigue, patrick.gallinari}@lip6.fr

geoffrey.scoutheeten@bnpparibas.com

Abstract

State-of-the-art NLP models can adopt shallow heuristics that limit their generalization capability (McCoy et al., 2019). Such heuristics include lexical overlap with the training set in Named-Entity Recognition (Taillé et al., 2020a) and *Event* or *Type* heuristics in Relation Extraction (Rosenman et al., 2020). In the more realistic end-to-end RE setting, we can expect yet another heuristic: the mere retention of training relation triples. In this paper we propose several experiments confirming that retention of known facts is a key factor of performance on standard benchmarks. Furthermore, one experiment suggests that a pipeline model able to use intermediate type representations is less prone to over-rely on retention.

1 Introduction

Information Extraction (IE) aims at converting the information expressed in a text into a predefined structured format of knowledge. This global goal has been divided into subtasks easier to perform automatically and evaluate. Hence, Named Entity Recognition (NER) and Relation Extraction (RE) are two key IE tasks among others such as Coreference Resolution (CR), Entity Linking or Event Extraction. Traditionally performed as a pipeline (Bach and Badaskar, 2007), these two tasks can be tackled jointly in order to model their interdependency, alleviate error propagation and obtain a more realistic evaluation setting (Roth and Yih, 2002; Li and Ji, 2014).

Following the general trend in Natural Language Processing (NLP), the recent quantitative improvements reported on Entity and Relation Extraction benchmarks are at least partly explained by the use of larger and larger pretrained Language Models (LMs) such as BERT (Devlin et al., 2019) to obtain contextual word representations. Concurrently, there is a realization that new evaluation protocols are necessary to better understand the strengths and shortcomings of the obtained neural network models, beyond a single holistic metric on an hold-out test set (Ribeiro et al., 2020).

In particular, generalisation to unseen data is a key factor in the evaluation of deep neural networks. It is all the more important in IE tasks that revolve around the extraction of mentions: small spans of words that are likely to occur in both the evaluation and training datasets. This lexical overlap has been shown to be correlated to neural networks performance in NER (Augenstein et al., 2017; Taillé et al., 2020a). For pipeline RE, Rosenman et al. (2020) and Peng et al. (2020) expose shallow heuristics in neural models: relying too much on the type of the candidate arguments or on the presence of specific triggers in their contexts.

In end-to-end Relation Extraction, we can expect that these NER and RE heuristics are combined. In this work, we argue that current evaluation benchmarks measure both the desired ability to extract information contained in a text but also the capacity of the model to simply retain labeled (head, predicate, tail) triples during training. And when the model is evaluated on a sentence expressing a relation seen during training, it is hard to disentangle which of these two behaviours is predominant. However, we can hypothesize that the model can simply retrieve previously seen information acting like a mere compressed form of knowledge base probed with a relevant query. Thus, testing on too much examples with seen triples can lead to overestimate the generalizability of a model.

Even without labeled data, LMs are able to learn some relations between words that can be probed with cloze sentences where an argument is masked (Petroni et al., 2019). This raises the additional question of lexical overlap with the orders of magnitude larger unlabeled LM pretraining corpora that will remain out of scope of this paper.

Code for reproducing our evaluation settings is available at github.com/btaille/retex

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10438–10449
November 7–11, 2021.

2 Datasets and Models

We study three recent end-to-end RE models on **CoNLL04** (Roth and Yih, 2004), **ACE05** (Walker et al., 2006) and **SciERC** (Luan et al., 2018). They rely on various pretrained LMs and for a fairer comparison, we use BERT (Devlin et al., 2019) on ACE05 and CoNLL04 and SciBERT (Beltagy et al., 2019) on SciERC[1] .

PURE (Zhong and Chen, 2021) follows the pipeline approach. The NER model is a classical span-based model (Sohrab and Miwa, 2018). Special tokens corresponding to each predicted entity span are added and used as representation for Relation Classification. For a fairer comparison with other models, we study the approximation model that only requires one pass in each encoder and limits to sentence-level prediction. However, it still requires finetuning and storing two pretrained LMs instead of a single one for the following models.

SpERT (Eberts and Ulges, 2020) uses a similar span-based NER module. RE is performed based on the filtered representations of candidate arguments as well as a max-pooled representation of their middle context. While Entity Filtering is close to the pipeline approach, the NER and RE modules share a common entity representation and are trained jointly. We also study the ablation of the max-pooled context representation that we denote **Ent-SpERT**.

Two are better than one (TABTO) (Wang and Lu, 2020) intertwines a sequence encoder and a table encoder in a Table Filling approach (Miwa and Sasaki, 2014). Contrary to previous models the pretrained LM is frozen and both the final hidden states and attention weights are used by the encoders. The prediction is finally performed by a Multi-Dimensional RNN (MD-RNN). Because it is not based on span-level predictions, this model cannot detect nested entities, e.g. on SciERC.

3 Partitioning by Lexical Overlap

Following (Augenstein et al., 2017; Taillé et al., 2020a), we partition the entity mentions in the test set based on lexical overlap with the training set. We distinguish *Seen* and *Unseen* mentions and also extend this partition to relations. We denote a relation as an *Exact Match* if the same (head, predicate, tail) triple appears in the train set; as a *Partial Match* if one of its arguments appears in the same position in a training relation of same type; and as *New* otherwise.

We implement a naive **Retention Heuristic** that tags an entity mention or a relation exactly present in the training set with its majority label. We report micro-averaged Precision, Recall and F1 scores for both NER and RE in Table 1.

An entity mention is considered correct if both its boundaries and type have been correctly predicted. For RE, we report scores in the **Boundaries** and **Strict** settings (Bekoulis et al., 2018; Taillé et al., 2020b). In the Boundaries setting, a relation is correct if its type is correct and the boundaries of its arguments are correct, without considering the detection of their types. The Strict setting adds the requirement that the entity type of both argument is correct.

3.1 Dataset Specificities

We first observe very different statistics of Mention and Relation Lexical Overlap in the three datasets, which can be explained by the singularities of their entities and relations. In CoNLL04, mentions are mainly Named Entities denoted with proper names while in ACE05 the surface forms are very often common names or even pronouns, which explains the occurrence of training entity mentions such as "it", "which", "people" in test examples. This also leads to a weaker entity label consistency (Fu et al., 2020a): "it" is labeled with every possible entity type and appears mostly unlabeled whereas a mention such as "President Kennedy" is always labeled as a person in CoNLL04. Similarly, mentions in SciERC are common names which can be tagged with different labels and they can also be nested. Both the poor label consistency as well as the nested nature of entities hurt the performance of the retention heuristic.

For RE, while SciERC has almost no exact overlap between test and train relations, ACE05 and CoNLL04 have similar levels of exact match. The larger proportion of partial match in ACE05 is explained by the pronouns that are more likely to co-occur in several instances. The difference in performance of the heuristic is also explained by a poor relation label consistency.

3.2 Lexical Overlap Bias

As expected, this first evaluation setting enables to expose an important lexical overlap bias, already

[1]More implementation details in Appendix A

$\mu F1$	NER			RE Boundaries				RE Strict			
	Seen	Unseen	All	Exact	Partial	New	All	Exact	Partial	New	All
						ACE05					
proportion	*82%*	*18%*		*23%*	*63%*	*14%*		*23%*	*63%*	*14%*	
heuristic	59.2	-	55.1	37.9	-	-	23.0	34.3	-	-	20.8
Ent-SpERT	$89.0_{0.1}$	$74.1_{1.0}$	$86.5_{0.2}$	$77.0_{1.1}$	$52.2_{1.1}$	$38.9_{1.0}$	$57.0_{0.8}$	$75.1_{1.2}$	$48.4_{1.0}$	$36.3_{2.0}$	$53.9_{0.8}$
SpERT	$89.4_{0.2}$	$74.2_{0.8}$	$86.8_{0.2}$	$84.8_{0.8}$	$59.6_{0.7}$	$42.3_{1.1}$	$64.0_{0.6}$	$82.6_{0.8}$	$55.6_{0.7}$	$38.4_{1.1}$	$60.6_{0.5}$
TABTO	$89.7_{0.1}$	$77.4_{0.8}$	$87.5_{0.2}$	$85.9_{0.9}$	$62.6_{1.8}$	$44.6_{2.9}$	$\mathbf{66.4}_{1.3}$	$81.6_{1.5}$	$58.1_{1.6}$	$38.5_{3.1}$	$\mathbf{61.7}_{1.1}$
PURE	$90.5_{0.2}$	$80.0_{0.3}$	$\mathbf{88.7}_{0.1}$	$86.0_{1.3}$	$60.5_{1.0}$	$47.1_{1.6}$	$\mathbf{65.1}_{0.7}$	$84.1_{1.1}$	$57.9_{1.3}$	$44.0_{2.0}$	$\mathbf{62.6}_{0.9}$
						CoNLL04					
proportion	*50%*	*50%*		*23%*	*34%*	*43%*		*23%*	*34%*	*43%*	
heuristic	86.0	-	59.7	90.9	-	-	35.5	90.9	-	-	35.5
Ent-SpERT	$95.9_{0.3}$	$81.9_{0.2}$	$\mathbf{88.9}_{0.2}$	$92.3_{1.4}$	$60.8_{1.4}$	$54.6_{1.3}$	$64.8_{0.9}$	$92.3_{1.4}$	$60.8_{1.4}$	$54.2_{1.2}$	$64.7_{0.8}$
SpERT	$95.4_{0.4}$	$81.2_{0.4}$	$88.3_{0.2}$	$91.4_{0.6}$	$67.0_{1.1}$	$59.0_{1.4}$	$69.3_{1.2}$	$91.4_{0.6}$	$66.9_{1.1}$	$58.5_{1.4}$	$69.0_{1.2}$
TABTO	$95.4_{0.4}$	$83.1_{0.7}$	$\mathbf{89.2}_{0.5}$	$92.6_{1.5}$	$72.6_{2.1}$	$64.8_{1.0}$	$\mathbf{74.0}_{1.4}$	$92.6_{1.5}$	$72.1_{1.8}$	$64.7_{1.1}$	$\mathbf{73.8}_{1.2}$
PURE	$95.0_{0.2}$	$81.8_{0.2}$	$88.4_{0.2}$	$90.1_{1.3}$	$66.6_{1.0}$	$58.6_{1.5}$	$68.3_{1.0}$	$89.9_{1.4}$	$66.6_{1.0}$	$58.5_{1.5}$	$68.2_{0.9}$
						SciERC					
proportion	*23%*	*77%*		*<1%*	*30%*	*69%*		*<1%*	*30%*	*69%*	
heuristic	31.3	-	20.1	-	-	-	0.7	-	-	-	0.7
Ent-SpERT	$77.6_{1.0}$	$64.0_{0.6}$	$\mathbf{67.3}_{0.6}$	-	$48.1_{0.7}$	$41.9_{0.6}$	$43.8_{0.5}$	-	$38.1_{1.9}$	$29.4_{1.1}$	$32.1_{1.2}$
SpERT	$78.5_{0.5}$	$64.2_{0.4}$	$\mathbf{67.6}_{0.3}$	-	$53.1_{1.2}$	$46.0_{1.0}$	$\mathbf{48.2}_{1.1}$	-	$43.0_{1.6}$	$33.2_{1.1}$	$\mathbf{36.2}_{1.0}$
PURE	$78.0_{0.5}$	$63.8_{0.6}$	$\mathbf{67.2}_{0.4}$	-	$54.0_{0.7}$	$44.8_{0.4}$	$\mathbf{47.6}_{0.3}$	-	$42.2_{0.7}$	$32.6_{0.7}$	$\mathbf{35.6}_{0.6}$

Table 1: Test NER and RE F1 Scores separated by lexical overlap with the training set. Exact Match RE scores are not reported on SciERC where the support is composed of only 5 exactly seen relation instances. Average and standard deviations on five runs.

discussed in NER, in end-to-end Relation Extraction. On every dataset and for every model micro F1 scores are the highest for Exact Match relations, then Partial Match and finally totally unseen relations. This is a first confirmation that retention plays an important role in the measured overall performance of end-to-end RE models.

3.3 Model Comparisons

While we cannot evaluate TABTO on SciERC because it is unfit for extraction of nested entities, we can notice different hierarchies of models on every dataset suggesting that there is no one-size-fits-all best model, at least in current evaluation settings.

The most obvious comparison is between SpERT and Ent-SpERT where the explicit representation of context is ablated. This results in a loss of performance on the RE part and especially on partially matching or new relations for which the entity representations pairs have not been seen. Ent-SpERT is particularly effective on Exact Matches on CoNLL04, suggesting its retention capability.

Other comparisons are more difficult, given the numerous variations between the very structure of each model as well as training procedures. However, the PURE pipeline setting seems to only be more effective on ACE05 where its NER performance is significantly better, probably because learning a separate NER and RE encoder enables to learn and capture more specific information for each distinctive task. Even then, TABTO yields better Boundaries performance only penalized on the Strict setting by entity types confusions. On the contrary, on CoNLL04, TABTO significantly outperforms its counterparts, especially on unseen relations. This indicates that it proposes a more effective incorporation of contextual information in this case where relation and argument types are mapped bijectively.

On SciERC, performance of all models is already compromised at the NER level before the RE step, which makes further distinction between model performance even more difficult.

4 Swapping Relation Heads and Tails

A second experiment to validate that retention is used as a heuristic in models' predictions is to modify their input sentences in a controlled manner

	Sentence	Ground Truth Relation
Original	**John Wilkes Booth** , who assassinated **President Lincoln** , was an actor .	(John Wilkes Booth, Kill, President Lincoln)
Swapped	**President Lincoln** , who assassinated **John Wilkes Booth** , was an actor .	(President Lincoln, Kill, John Wilkes Booth)

Table 2: Example of Swapped sentence. The Triple (John Wilkes Booth, Kill, President Lincoln) is present in the training set and the retention behaviours lead models to extract this triple when probed with the swapped sentence expressing the reverse relation.

F1	NER ↑		RE ↑		revRE ↓	
	O	S	O	S	O	S
Kill						
Ent-SpERT	91.6	91.7	85.1	35.4	-	58.5
SpERT	91.4	**92.6**	86.2	35.0	-	57.8
TABTO	**92.0**	**92.8**	**89.6**	27.6	-	59.5
PURE	90.5	90.7	84.1	**52.3**	-	**14.3**
Located in						
Ent-SpERT	**90.0**	87.0	78.3	30.3	-	24.8
SpERT	88.6	87.7	75.0	24.9	-	33.5
TABTO	**90.1**	**88.9**	**85.3**	36.1	-	34.9
PURE	89.0	83.7	81.2	**59.3**	-	**5.1**

Table 3: Performance on CoNLL04 test set containing exactly one relation of the corresponding type in its original form (O) and where the relation head and tail are swapped (S). NER F1 score is micro-averaged while strict RE score only takes these relations into account. The revRE score corresponds to unwanted extraction of the reverse relation, symptomatic of the retention effect in the swapped setting.

similarly to what is proposed in (Ribeiro et al., 2020). We propose a very focused experiment that consists in selecting asymmetric relations that occur between entities of same type and swap the head with the tail in the input. If the model predicts the original triple, then it over relies on the retention heuristic, whereas finding the swapped triple is an evidence of broader context incorporation. We show an example in Table 2.

Because of the requirements of this experiment, we have to limit to two relations in CoNLL04: “Kill” between people and “Located in” between locations. Indeed, CoNLL04 is the only dataset with a bijective mapping between the type of a relation and the types of its arguments and the consistent proper nouns mentions makes the swaps mostly grammatically correct. For each relation type, we only consider sentences with exactly one instance of corresponding relation and swap its arguments. We only consider this relation in the RE scores reported in Table 3. We use the strict RE score as well as **revRE** which measures the extraction of the reverse relation, not expressed in the sentence.

For each relation, the hierarchy of models corresponds to the overall CoNLL04. Swapping arguments has a limited effect on NER, mostly for the "Located in" relation. However, it leads to a drop in RE for every model and the revRE score indicates that SpERT and TABTO predict the reverse relation more often than the newly expressed one. This is another proof of the retention heuristic of end-to-end models, although it might also be attributed to the language model to the language model. In particular for the ”Located in“ relation, swapped heads and tails are not exactly equivalent since the former are mainly cities and the latter countries.

On the contrary, the PURE model is less prone to information retention, as shown by its revRE scores significantly smaller than the standard RE scores on swapped sentences. Hence, it outperforms SpERT and TABTO on swapped sentences despite being the least effective on the original dataset.The important discrepancy in results can be explained by the different types of representations used by these models. The pipeline approach allows the use of argument type representations in the Relation Classifier whereas most end-to-end models use lexical features in a shared entity representation used for both NER and RE.

These conclusions from quantitative results are validated qualitatively. We can observe that the four predominant patterns are intuitive behaviours on sentences with swapped relations: retention of the incorrect original triple, prediction of the correct swapped triple and prediction of none or both triples. We report some examples in Table 9 and Table 10 in the Appendix.

5 Related Work

Several works on generalization of NER models mention lexical overlap with the training as a key indicator of performance. Augenstein et al. (2017) separate mentions in the test set as seen and unseen during training and measure out-of-domain generalization in an extensive study of two CRF based models and SENNA combining a Convolutional Neural Network with a CRF (Collobert and Weston, 2011). Taillé et al. (2020a) compare the effect of introducing contextual embeddings in the classical BiLSTM-CRF architecture in a similar setting and show that they help close the performance gap on unseen mentions and domains. Arora et al. (2020); Fu et al. (2020b,a) study the influence of several properties such as lexical overlap, label consistency and entity length on state-of-the-art models performance. They model these properties as continuous scores associated to each mention and bucketized for evaluation. Lexical overlap has also been mentioned in Coreference Resolution (Moosavi and Strube, 2017) where coreferent mentions tend to co-occur in the test and train sets. In this line of works, the impact of lexical overlap is measured either by separating performance depending on the property of mentions (seen or unseen) or with out-of-domain evaluation with a test set from a different dataset with lower lexical overlap with the train set.

Another recently proposed method for fine-grained evaluation of NLP models beyond a single benchmark score is to modify the test sentences in a controlled manner. McCoy et al. (2019) expose lexical overlap as a shallow heuristic adopted by state-of-the-art Natural Language Inference models, especially by swapping subject and object of verbs in the hypothesis of some examples where the premise entails the hypothesis. While such a modification changes the label of these examples to non-entailment, all models tested show a spectacular drop of accuracy on these models. Ribeiro et al. (2020) propose a broader set of test set modifications to individually test robustness of NLP models to several patterns such as the introduction of negation, swapping words with synonyms, changing tense and much more.

In pipeline RE where ground truth candidate arguments are given, models often use intermediate representations based on entity types that reduce lexical overlap issues. However, Rosenman et al. (2020) show that they still tend to adopt shallow heuristics based on the type of the arguments and the presence of triggers indicative of the presence of a relation. They propose hard cases with several mentions of same types for which Relation Classifiers struggle connecting the correct pair. Concurrently, Peng et al. (2020) confirm that RE benchmarks present shallow cues such as the type of the candidate arguments that can be used alone to infer the relation.

We propose to extend previous work on NER and RE to the more realistic end-to-end RE setting with two of the previously described approaches: 1) separating performance by lexical overlap of mentions or argument pairs and 2) modifying some CoNLL04 test examples by swapping relations heads and tails.

6 Conclusion

In this paper, we study three state-of-the-art end-to-end Relation Extraction models in order to highlight their tendency to retain seen relations. We confirm that retention of seen mentions and relations play an important role in overall RE performance and can explain the relatively higher scores on CoNLL04 and ACE05 compared to SciERC. Furthermore, our experiment on swapping relation heads and tails tends to show that the intermediate manipulation of type representations instead of lexical features enabled in the pipeline PURE model makes it less prone to over-rely on retention.

While the limited extend of our swapping experiment is an obvious limitation of this work, it shows limitations of both current benchmarks and models. It is an encouragement to propose new benchmarks that might be easily modified by design to probe such lexical overlap heuristics. Contextual information could for example be contained in templates of that would be filled with different (head, tail) pairs either seen or unseen during training.

Furthermore, pretrained Language Models can already capture relational information between phrases (Petroni et al., 2019) and further experiments could help distinguish their role in the retention behaviour of RE models.

Acknowledgments

We thank the anonymous reviewers for their thoughtful comments. Part of this work was performed while Bruno Taillé was an employee of BNP Paribas and supported by the French Ministry of Higher Education, Research and Innovation under the CIFRE convention 2018/0327.

References

Simran Arora, Avner May, Jian Zhang, and Christopher Ré. 2020. Contextual embeddings: When are they worth it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2663, Online. Association for Computational Linguistics.

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83.

Nguyen Bach and Sameer Badaskar. 2007. A Review of Relation Extraction. *Literature review for Language and Statistics II 2*, page 15.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Adversarial training for multi-context joint entity and relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2830–2836, Brussels, Belgium. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Ronan Collobert and Jason Weston. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

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.

Markus Eberts and Adrian Ulges. 2020. Span-based Joint Entity and Relation Extraction with Transformer Pre-training. In *Proceedings of the 12th European Conference on Artificial Intelligence (ECAI)*.

Jinlan Fu, Pengfei Liu, and Graham Neubig. 2020a. Interpretable multi-dataset evaluation for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6058–6069, Online. Association for Computational Linguistics.

Jinlan Fu, Pengfei Liu, Qi Zhang, and Xuanjing Huang. 2020b. Rethinking Generalization of Neural Models: A Named Entity Recognition Case Study. In *AAAI 2020*.

Qi Li and Heng Ji. 2014. Incremental Joint Extraction of Entity Mentions and Relations. pages 402–412. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Makoto Miwa and Yutaka Sasaki. 2014. Modeling joint entity and relation extraction with table representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1858–1869, Doha, Qatar. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2017. Lexical features in coreference resolution: To be used with caution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Vancouver, Canada. Association for Computational Linguistics.

Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. Learning from Context or Names? An Empirical Study on Neural Relation Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3661–3672, Online. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Shachar Rosenman, Alon Jacovi, and Yoav Goldberg. 2020. Exposing Shallow Heuristics of Relation Extraction Models with Challenge Data. In *Proceedings of the 2020 Conference on Empirical Methods*

in Natural Language Processing (EMNLP), pages 3702–3710, Online. Association for Computational Linguistics.

Dan Roth and Wen-Tau Yih. 2002. Probabilistic Reasoning for Entity & Relation Recognition. In *COLING: The 19th International Conference on Computational Linguistics*.

Dan Roth and Wen-tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, pages 1–8, Boston, Massachusetts, USA. Association for Computational Linguistics.

Mohammad Golam Sohrab and Makoto Miwa. 2018. Deep exhaustive model for nested named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2843–2849, Brussels, Belgium. Association for Computational Linguistics.

Bruno Taillé, Vincent Guigue, and Patrick Gallinari. 2020a. Contextualized Embeddings in Named-Entity Recognition: An Empirical Study on Generalization. In *Advances in Information Retrieval*, pages 383–391. Springer International Publishing.

Bruno Taillé, Vincent Guigue, Geoffrey Scoutheeten, and Patrick Gallinari. 2020b. Let's Stop Incorrect Comparisons in End-to-end Relation Extraction! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3689–3701, Online. Association for Computational Linguistics.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 Multilingual Training Corpus. Linguistic Data Consortium.

Jue Wang and Wei Lu. 2020. Two are better than one: Joint entity and relation extraction with table-sequence encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1706–1721, Online. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. A Frustratingly Easy Approach for Entity and Relation Extraction. In *NAACL 2021*.

A Implementation Details

For every model, we use the original code associated with the papers with the default best performing hyperparameters unless stated otherwise. We run 5 runs on a single NVIDIA 2080Ti GPU for each of them on each dataset. For CoNLL04 and ACE05, we train each model with both the cased and uncased versions of BERT$_{BASE}$ and only keep the best performing setting.

PURE (Zhong and Chen, 2021) [1] We use the approximation model and limit use a *context window* of 0 to only use the current sentence for prediction and be able to compare with other models. For ACE05, we use the standard *bert-base-uncased* LM but use the *bert-base-cased* version on CoNLL04 which results in a significant +2.4 absolute improvement in RE Strict micro F1 score.

SpERT (Eberts and Ulges, 2020) [2] We use the original implementation as is with *bert-base-cased* for both ACE05 and CoNLL04 since the uncased version is not beneficial, even on ACE05 where there are fewer proper nouns. For the Ent-SpERT ablation, we simply remove the max-pooled context representation from the final concatenation in the RE module. This modifies the RE classifier's input dimension from the original 2354 to 1586.

Two are better than one (TABTO) (Wang and Lu, 2020) [3] We use the original implementation with *bert-base-uncased* for both ACE05 and CoNLL04 since the cased version is not beneficial on CoNLL04.

B Datasets Statistics

We present general datasets statistics in Table 4.

We also compute average values of some entity and relation attributes inspired by (Fu et al., 2020a) and reported in Table 5.

We report two of their entity attributes: **entity length** in number of tokens (**eLen**) and **entity label consistency (eCon)**. Given a test entity mention, its label consistency is the number of occurrences in the training set with the same type divided by its total number of occurrences. It is zero for unseen mentions. Because eCon reflects both the ambiguity of labels for seen entities and the proportion of unseen entities, we propose to introduce the **eCon*** score that only averages label consistency of seen mentions and **eLex**, the proportion of entities with lexical overlap with the train set.

We introduce similar scores for relations. **Relation label consistency (rCon)** extends label consistency for triples. **Argument types label constistency (aCon)** considers the labels of every pair of mentions of corresponding types in the training set. Because pairs of types are all seen during training we do not decompose aCon into aCon* and aLex. **Argument length (aLen)** is the sum of the lengths of the head and tail mentions. **Argument distance (aDist)** is the number of tokens between the head and the tail of a relation.

We present a more complete report of overall Precision, Recall and F1 scores that can be interpreted in light of these statistics in Table 6.

ACE05	Train	Dev	Test
Sentences	10,051	2,424	2,050
Mentions	26,473	6,338	5,476
Relations	4,788	1,131	1,151
CoNLL04	Train	Dev	Test
Sentences	922	231	288
Mentions	3,377	893	1,079
Relations	1,283	343	422
SciERC	Train	Dev	Test
Sentences	1,861	275	551
Mentions	5,598	811	1,685
Relations	3,219	455	974

Table 4: Datasets Statistics

[1] github.com/princeton-nlp/PURE
[2] github.com/lavis-nlp/spert
[3] github.com/LorrinWWW/two-are-better-than-one

	Entities				Relations					
	eCon	eCon*	eLex	eLen	rCon	rCon*	rLex	aCon	aLen	aDist
ACE05	65%	78%	82%	1.1	15%	62%	23%	7.1%	2.3	2.8
CoNLL04	49%	98%	50%	1.5	21%	91%	23%	29%	3.8	5.8
SciERC	17%	74%	23%	1.6	0.4%	74%	0.5%	13%	4.7	5.3

Table 5: Average of some entity and relation attributes in the test set.

$\mu F1$	NER			RE Boundaries			RE Strict		
	P	R	F1	P	R	F1	P	R	F1
				ACE05					
heuristic	44.7	71.9	55.1	23.6	22.3	23.0	21.4	20.2	20.8
Ent-SpERT	$86.7_{0.3}$	$86.3_{0.3}$	$86.5_{0.2}$	$56.7_{1.0}$	$57.4_{0.7}$	$57.0_{0.8}$	$53.5_{1.0}$	$54.2_{0.8}$	$53.9_{0.8}$
SpERT	$87.2_{0.2}$	$86.5_{0.3}$	$86.8_{0.2}$	$68.1_{1.1}$	$60.5_{0.5}$	$64.0_{0.6}$	$64.4_{1.1}$	$57.2_{0.4}$	$60.6_{0.5}$
TABTO	$86.7_{0.3}$	$88.3_{0.6}$	$87.5_{0.2}$	$71.0_{2.7}$	$62.5_{2.5}$	$66.4_{1.3}$	$66.1_{2.6}$	$58.1_{2.1}$	$61.8_{1.1}$
PURE	$88.8_{0.3}$	$88.6_{0.1}$	$88.7_{0.1}$	$67.4_{0.8}$	$63.0_{0.8}$	$65.1_{0.7}$	$64.8_{1.0}$	$60.5_{1.0}$	$62.6_{0.9}$
				CoNLL04					
heuristic	75.9	49.2	59.7	84.1	22.5	35.5	84.1	22.5	35.5
Ent-SpERT	$88.4_{0.6}$	$89.3_{0.7}$	$88.9_{0.2}$	$59.3_{0.5}$	$71.3_{1.5}$	$64.8_{0.9}$	$59.2_{0.5}$	$71.2_{1.5}$	$64.7_{0.8}$
SpERT	$87.9_{0.6}$	$88.7_{0.3}$	$88.3_{0.2}$	$69.7_{2.3}$	$69.0_{0.5}$	$69.3_{1.2}$	$69.4_{2.3}$	$68.7_{0.6}$	$69.0_{1.2}$
TABTO	$89.0_{0.7}$	$89.3_{0.3}$	$89.2_{0.5}$	$75.6_{3.2}$	$72.6_{1.9}$	$74.0_{1.4}$	$75.4_{3.1}$	$72.4_{1.8}$	$73.8_{1.2}$
PURE	$88.3_{0.4}$	$88.5_{0.5}$	$88.4_{0.2}$	$68.6_{2.0}$	$68.2_{1.6}$	$68.3_{1.0}$	$68.5_{2.0}$	$68.1_{1.5}$	$68.2_{0.9}$
				SciERC					
heuristic	18.8	21.5	20.1	3.5	0.4	0.7	3.5	0.4	0.7
Ent-SpERT	$68.0_{0.3}$	$66.6_{0.9}$	$67.3_{0.6}$	$44.8_{0.7}$	$42.9_{1.0}$	$43.8_{0.5}$	$32.9_{0.9}$	$31.5_{1.5}$	$32.1_{1.2}$
SpERT	$67.6_{0.5}$	$67.6_{0.2}$	$67.6_{0.3}$	$49.3_{1.4}$	$47.2_{1.3}$	$48.2_{1.1}$	$37.0_{1.3}$	$35.4_{1.0}$	$36.2_{1.0}$
PURE	$68.2_{0.6}$	$66.2_{0.9}$	$67.2_{0.4}$	$50.2_{0.9}$	$45.2_{1.0}$	$47.6_{0.3}$	$37.6_{1.2}$	$33.8_{0.7}$	$35.6_{0.6}$

Table 6: Overall micro-averaged Test NER and Strict RE Precision, Recall and F1 scores. Average and standard deviations on five runs. We can observe that the recall of the heuristic is correlated with the proportions of seen entities or triples (eLex or rLex). Its particularly high precision on CoNLL04 seems rather linked to the important label consistency of seen entities and relation (eCon* and rCon*).

Dataset	Entity Types	Relation Types
ACE05	Facility, Geo-political Entity, Location, Person, Vehicle, Weapon	Artifact, Gen-affiliation, Org-affiliation, Part-whole, Person-social, Physical
CoNLL04	Location, Organization, Other, Person	Kill, Live in, Located in, Organization based in, Work for
SciERC	Generic, Material, Method, Metric, Other Scientific Term, Task	Compare*, Conjunction*, Evaluate for, Feature of, Hyponym of, Part of, Used for

Table 7: Entity and Relation Types of end-to-end RE datasets. SciERC presents two types of symmetric relations denoted with a *.

			NER ↑			RE Strict ↑			Reverse RE Strict ↓		
			P	R	F1	P	R	F1	P	R	F
Kill	Original	Ent-SpERT	$91.7_{0.4}$	$91.5_{0.7}$	$91.6_{0.4}$	$82.9_{2.7}$	$87.6_{1.8}$	$85.1_{0.9}$	-	-	-
		SpERT	$91.7_{2.1}$	$91.0_{1.0}$	$91.4_{1.2}$	$88.1_{3.1}$	$84.4_{1.4}$	$86.2_{1.4}$	-	-	-
		TABTO	$91.8_{0.6}$	$92.2_{0.5}$	$\mathbf{92.0}_{0.4}$	$88.8_{1.6}$	$90.7_{3.3}$	$\mathbf{89.6}_{1.3}$	-	-	-
		PURE	$91.5_{0.9}$	$89.6_{0.6}$	$90.5_{0.6}$	$87.2_{2.1}$	$81.3_{1.1}$	$84.1_{1.2}$	-	-	-
	Swap	Ent-SpERT	$91.3_{0.9}$	$92.1_{0.7}$	$91.7_{0.7}$	$31.8_{5.3}$	$40.0_{8.3}$	$35.4_{6.5}$	$52.8_{5.6}$	$65.8_{7.2}$	$58.5_{5.7}$
		SpERT	$92.6_{1.8}$	$92.6_{0.8}$	$\mathbf{92.6}_{1.2}$	$33.0_{4.4}$	$37.3_{7.4}$	$35.0_{5.6}$	$54.8_{5.1}$	$61.3_{4.1}$	$57.8_{4.0}$
		TABTO	$92.8_{0.8}$	$92.7_{0.9}$	$\mathbf{92.8}_{0.7}$	$26.8_{3.6}$	$28.4_{4.1}$	$27.6_{3.8}$	$57.8_{3.1}$	$61.3_{3.0}$	$59.5_{2.8}$
		PURE	$92.0_{0.5}$	$89.5_{1.0}$	$90.7_{0.5}$	$65.2_{6.0}$	$44.0_{7.4}$	$\mathbf{52.3}_{6.5}$	$17.8_{2.3}$	$12.0_{2.3}$	$\mathbf{14.3}_{2.2}$
Located in	Original	Ent-SpERT	$90.1_{0.8}$	$89.8_{1.5}$	$\mathbf{90.0}_{0.7}$	$80.8_{3.7}$	$76.2_{3.2}$	$78.3_{2.4}$	-	-	-
		SpERT	$89.8_{1.2}$	$87.5_{1.5}$	$88.6_{1.1}$	$77.2_{2.8}$	$73.0_{3.0}$	$75.0_{2.0}$	-	-	-
		TABTO	$90.1_{1.3}$	$90.0_{1.8}$	$\mathbf{90.1}_{1.5}$	$93.0_{3.3}$	$78.9_{4.6}$	$\mathbf{85.3}_{3.9}$	-	-	-
		PURE	$88.6_{1.1}$	$89.4_{1.8}$	$89.0_{1.0}$	$89.3_{4.0}$	$74.6_{3.7}$	$81.2_{2.6}$	-	-	-
	Swap	Ent-SpERT	$86.7_{1.9}$	$87.4_{2.7}$	$87.0_{2.1}$	$38.0_{8.5}$	$25.4_{2.8}$	$30.3_{4.6}$	$30.2_{5.2}$	$21.1_{5.8}$	$24.8_{5.7}$
		SpERT	$87.3_{1.4}$	$88.0_{0.9}$	$87.7_{1.1}$	$34.8_{14.8}$	$19.5_{6.7}$	$24.9_{9.2}$	$45.6_{17.0}$	$26.5_{10.5}$	$33.5_{13.0}$
		TABTO	$89.0_{0.6}$	$88.8_{0.9}$	$\mathbf{88.9}_{0.8}$	$46.5_{6.6}$	$29.7_{5.7}$	$36.1_{5.8}$	$45.2_{5.2}$	$28.6_{3.7}$	$34.9_{3.6}$
		PURE	$82.7_{0.8}$	$84.6_{0.8}$	$83.7_{0.5}$	$74.9_{7.6}$	$49.7_{4.7}$	$\mathbf{59.3}_{3.0}$	$6.5_{1.8}$	$4.3_{1.3}$	$\mathbf{5.1}_{1.5}$

Table 8: Detailed results of the Swap Relation Experiment with Precision, Recall and F1 scores.

1	The Warren Commission determined that on Nov. 22 , 1963 , **A** fired a high-powered rifle at **B** 's motorcade from the sixth floor of what is now the Dallas County Administration Building , where he worked .	
A, B	Lee Harvey Oswald, Kennedy	Kennedy, Lee Harvey Oswald
Ent-SpERT	(A,B)	(B,A)
SpERT	(A,B)	(B,A)
TABTO	(A,B)	(B,A)
PURE	(A,B)	(B,A)
2	Today 's Highlight in History : Twenty years ago , on June 6 , 1968 , at 1 : 44 a.m. local time , **B** died at Good Samaritan Hospital in Los Angeles , 25 -LCB- hours after he was shot at the Ambassador Hotel by **A** .	
A, B	Sirhan Bishara Sirhan, Sen. Robert F. Kennedy	Sen. Robert F. Kennedy, Sirhan Bishara Sirhan
Ent-SpERT	(A,B)	(B,A)
SpERT	(A,B)	(B,A)
TABTO	(A,B)	(B,A)
PURE	(A,B)	-
3	In 1968 , authorities announced the capture in London of **A** , suspected of the assassination of civil rights leader **B** .	
A, B	James Earl Ray, Dr. Martin Luther King Jr	Dr. Martin Luther King Jr, James Earl Ray
Ent-SpERT	(A,B)	(A,B) (B,A)
SpERT	(A,B)	(A,B) (B,A)
TABTO	(A,B)	(A,B)
PURE	(A,B)	(A,B)
4	The Warren Commission determined that **A** fired at **B** from the sixth floor of what is now the Dallas County Administration Building .	
A, B	Oswald, Kennedy	Kennedy, Oswald
Ent-SpERT	(A,B)	-
SpERT	(A,B)	(A,B) (B,A)
TABTO	(A,B)	(B,A)
PURE	(A,B)	(A,B)

Table 9: Some qualitative examples of models' predictions on original (left column) and swapped (right) CoNLL04 sentences for the "Kill" relation. Despite a perfect Relation Extraction in the original sentences for all models, swapping head and tails results in several types of errors mainly regarding the direction of the relation. Predictions of incorrect original triples are in red. These examples are obtained from models trained with the same seed ($s = 0$).

1	Reagan recalled that on the 40th anniversary of the Normandy landings he read a letter from a young woman whose late father had fought at **A** , a **B** sector .	
A, B	Omaha Beach, Normandy	Normandy, Omaha Beach
Ent-SpERT	(A,B)	-
SpERT	(A,B)	-
TABTO	(A,B)	-
PURE	(A,B)	(A,B)
2	**A** , **B** (AP)	
A, B	MILAN, Italy	Italy, MILAN
Ent-SpERT	(A,B)	(A,B)
SpERT	(A,B)	(A,B)
TABTO	(A,B)	(B,A)
PURE	(A,B)	-
3	In **A** , downed tree limbs interrupted power in parts of **B** .	
A, B	Indianapolis, Indiana	Indiana, Indianapolis
Ent-SpERT	(A,B)	(B,A)
SpERT	(A,B)	(B,A)
TABTO	(A,B)	(B,A)
PURE	(A,B)	(B,A)
4	The plane , owned by Bradley First Air , of **A** , **B** , was carrying cargo to Montreal for Emery Air Freight Corp. , an air freight courier service with a hub at the Dayton airport .	
A, B	Ottawa, Canada	Canada, Ottawa
Ent-SpERT	(A,B) (Dayton airport, Canada)	(Dayton airport, Ottawa)
SpERT	(A,B) (Dayton airport, Canada)	-
TABTO	(A,B)	(A,B)
PURE	(A,B)	(A,B)

Table 10: Some qualitative examples of models' predictions on original (left column) and swapped (right) CoNLL04 sentences for the "Located in" relation. This relation is often simply expressed by an apposition of the head and tail separated by a comma. Predictions of incorrect original triples are in red. These examples are obtained from models trained with the same seed ($s = 0$).