X-WikiRE: A Large, Multilingual Resource for Relation Extraction as Machine Comprehension

Mostafa Abdou, Cezar Sas, Rahul Aralikatte, Isabelle Augenstein and **Anders Søgaard**
{`abdou, sas, rahul, augenstein, soegaard`} @ `di.ku.dk`
University of Copenhagen

Abstract

Although the vast majority of knowledge bases (**KBs**) are heavily biased towards English, Wikipedias do cover very different topics in different languages. Exploiting this, we introduce a new multilingual dataset (**X-WikiRE**), framing relation extraction as a multilingual machine reading problem. We show that by leveraging this resource it is possible to robustly transfer models cross-lingually and that multilingual support significantly improves (zero-shot) relation extraction, enabling the population of low-resourced **KBs** from their well-populated counterparts.

1 Introduction

It is a widely lamented fact that linguistic and encyclopedic resources are heavily biased towards English. Even multilingual knowledge bases (KBs) such as Wikidata (Vrandečić and Krötzsch, 2014) are predominantly English-based (Kaffee and Simperl, 2018). This means that coverage is higher for English, and that facts of interest to English-speaking communities are more likely included in a KB. This work introduces a novel multilingual dataset (**X-WikiRE**) and explores techniques for automatically filling such language gaps by learning, from **X-WikiRE**, to add facts in other languages. Finally, we show that multilingual sharing is beneficial for knowledge base completion across all languages, including English.

The task of identifying potential KB entries in running text – i.e., relations that hold between two or more entities, is called *relation extraction* (RE). In the traditional, supervised setting (Bach and Badaskar, 2007), RE models are trained to identify a pre-specified set of relation types, which are observed during training. Models are meant to generalize to new entities, but *not* new *relations*. An alternative flavor is *open* RE (Fader et al.,
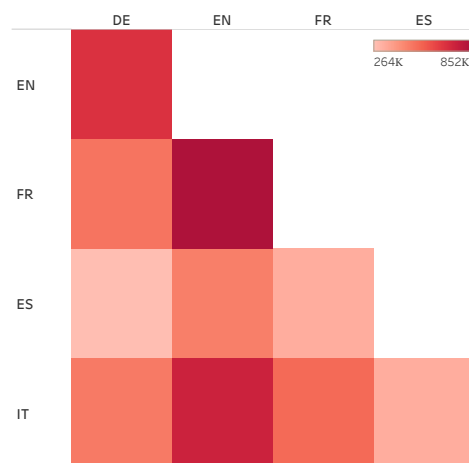

Figure 1: The overlap of triples between languages.

2011; Yates et al., 2007), which detects subject-verb-object triples and clusters semantically related verbs into coarse-grained semantic relations.

In this paper, we consider the middle ground, in which models are trained on a subset of pre-specified relations and applied to both seen and unseen entities, and unseen relations. The latter scenario is known as *zero-shot* RE (Rocktäschel et al., 2015).

Levy et al. (2017) present a reformulation of RE, where the task is framed as reading comprehension. In this formulation, each relation type (e.g. $author$, $occupation$) is mapped to at least one natural language question template (e.g. "*Who is the author of x?*"), where x is filled with an entity (e.g. "*Inferno*"). The model is then tasked with finding an answer ("*Dante Alighieri*") to this question with respect to a given context. They show that this formulation of the problem both outperforms off-the-shelf RE systems in the typical RE setting and, in addition, enables generalization to unspecified and unseen types of relations. **X-WikiRE** enables exploration of this reformulation of RE in a multilingual setting.
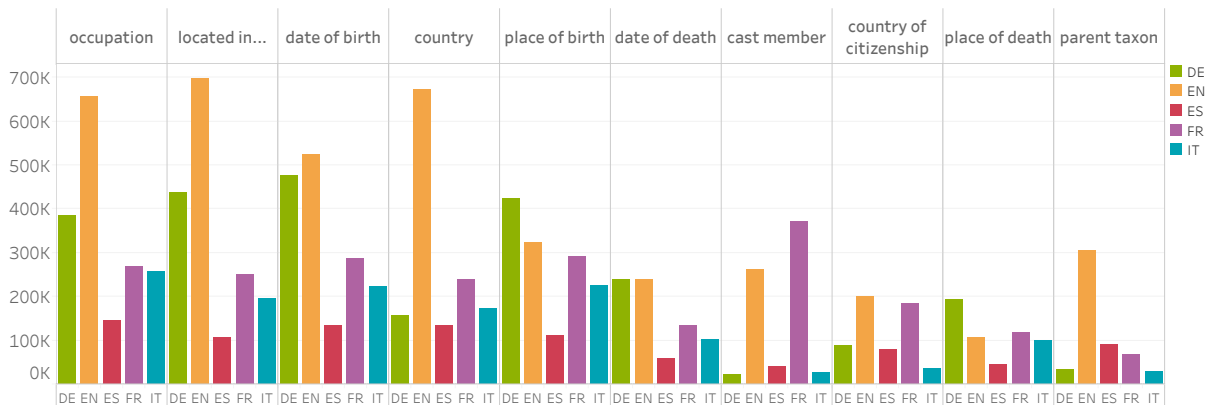
Figure 2: The number of triples for the top 10 properties in each language.

Contributions We introduce a new, large-scale multilingual dataset (**X-WikiRE**) of reading comprehension-based RE for English, German, French, Spanish, and Italian, facilitating research on multilingual methods for RE. Our dataset covers more languages (five) and is at least an order of magnitude larger than existing multilingual RE datasets, e.g., TAC 2016 (Ellis et al., 2015), which covers three languages and consists of ≈ 90k examples. We also a) perform cross-lingual RE showing that models pretrained on one language can be effectively transferred to others with minimal in-language finetuning; b) leverage multilingual representations to train a model capable of simultaneously performing (zero-shot) RE in all five languages, rivaling or outperforming its monolingually trained counterparts in many cases while requiring far fewer parameters per language; c) obtain considerable improvements by employing a more carefully designed nil-aware machine comprehension model.

2 Background

Relation extraction We begin with a brief description of our terminology. Given raw text, relation extraction is the task of identifying instances of relations $relation(entity_1, entity_2)$. We refer to these instances of relation and entity pairs as *triples*. Furthermore, throughout this work, we use the term *property* interchangeably with relation.

A large part of previous work on relation extraction has been concerned with extracting relations between unseen entities for a pre-defined set of relations seen during training (Zelenko et al., 2003; Zhou et al., 2005; Miwa and Bansal, 2016). For example, the instances (Barack Obama, Hawaii), (Niels Bohr, Copenhagen),

and (Jacques Brel, Schaerbeek) of the relation $born\_in(x, y)$ would be seen during the training phase, and then the model would be expected to correctly identify other instances of the relation such as (Jean-Paul Sartre, Paris) in running text. This is useful in closed-domain settings where it is possible to pre-select a set of relations of interest. In an open-domain setting, however, we are interested in the far more difficult problem of extracting unseen relation types. Open RE methods (Yates et al., 2007; Banko et al., 2007; Fader et al., 2011) do not require relation-specific data, but treat different phrasings of the same relation as different relations and rely on a combination of syntactic features (e.g. dependency parses) and normalisation rules, and so have limited generalization capacity.

Zero-shot relation extraction Levy et al. (2017) propose a novel approach towards achieving this generalization by transforming relations into natural language question templates. For instance, the relation $born\_in(x, y)$ can be expressed as "*Where was x born?*" or "*In which place was x born?*". Then, a reading comprehension model (Seo et al., 2016; Chen et al., 2017) can be trained on question, answer, and context examples where the x slot is filled with an entity and the y slot is either an answer if the answer is present in the context, or NIL. The model is then able to extract relation instances (given expressions of the relations as questions) from raw text. To test this *"harsh zero-shot"* setting of relation extraction, they build a dataset for RE as machine comprehension from WikiReading (Hewlett et al., 2016), relying on alignments between Wikipedia pages and Wikidata KB triples. They show that their read-

Lang	Question	Context & Answers
DE	In welchem land befindet man sich, wenn man **Amazonas** besucht?	Der Fluss Amazonas gab seinerseits dem Amazonasbecken sowie mehreren gleichnamigen Verwaltungseinheiten in **Brasilien**, **Venezuela**, **Kolumbien** . . .
EN	What country is **Amazon** located in?	The Amazon proper runs mostly through **Brazil** and **Peru**, and is part of the border between . . .
ES	¿En qué país se encuentra el **Amazonas**?	El río Amazonas es un río de América del Sur, que atraviesa **Perú**, **Colombia** y **Brasil**.
FR	Dans quel pays peux-tu trouver **Amazone**?	Le fleuve prend alors le nom d'Amazonas au **Pérou** et en **Colombie**, puis celui de rio Solimões en entrant au **Brésil** au . . .
IT	Di quale nazione fa parte il **Rio delle Amazzoni**?	Il Rio delle Amazzoni è un fiume dell'America Meridionale che attraversa **Perù**, **Colombia** e **Brasile** . . .

Table 1: Examples from our dataset of the same question-context pairs across all the languages with the correct answers highlighted in boldface.

ing comprehension model is able to use linguistic cues to identify relation paraphrases and lexico-syntactic patterns of textual deviation from questions to answers, enabling it to identify instances of new relations. Similar work (Obamuyide and Vlachos, 2018) recently also showed that RE can be framed as natural language inference.

3   X-WikiRE

X-WikiRE is a multilingual reading comprehension-based relation extraction dataset. Each example in the dataset consists of a *question*, a *context*, and an *answer*, where the question is a querified relation and the context may contain the answer or an indication that it is not present (NIL). Questions are obtained by transforming relations into question templates with slots where an entity is inserted. Within the RE framework described in Section 2, $entity_1$ is filled into a slot in the question template and $entity_2$ is the answer. Each triple[1] in the dataset can be identified uniquely across all languages. We construct **X-WikiRE** using the relevant parts of Wikidata and Wikipedia for each language. Wikidata is an open KB where the knowledge contained in each document is expressed as a set of statements, and each statement is a tuple (property_id, value_id) (e.g. statement (P50, Q1067) where P50 refers to *author* and Q1067 to "*Dante Alighieri*"). We perform data integration on Wikidata, as described by Hewlett et al. (2016): for each entity in Wikipedia

we take the corresponding Wikidata document, add the Wikipedia page text, and denormalize the statements. This consists of replacing the property and value ids of each statement in the document with the text label for values which are entities, and with the human readable form for numeric values (e.g. timestamps are converted to natural forms like "*25 May 1994*") obtaining a tuple $(property, entity)$.[2]

Slot-filling data   To extract the contexts for each triple in our dataset we use the distant supervision method described by Levy et al. (2017). For each Wikidata document belonging to a given $entity_1$ we take all the denormalized tuples $(property, entity_2)$ and extract the first sentence in the text containing both $entity_1$ and $entity_2$. Negatives (contexts without answers) are constructed by finding pairs of triples with common $entity_2$ type (to ensure they contain good distractors), swapping their context if $entity_2$ is not present in the context of the other triple.

Querification   Levy et al. (2017) created 1192 question templates for 120 Wikidata properties. A template contains a placeholder for an entity x (e.g. for property "*author*", some templates are "*Who wrote the novel x?*" and "*Who is the author of x?*"), which can be automatically filled in to create questions so that $question \approx template(property, x)$. For our multilingual dataset, we had these templates translated by human translators. The translators attempted to translate each of the original 1192 templates. If a template was difficult to translate, they were in-

---

[1]Not to be confused with an example as an example contains an instantiation of a relation in the form of a question. Thus, the different question templates for each relation share the same id.

[2]We make the simplification of referring to all values as entities.

Language	Pos	Neg	Pos*	Neg*
DE	2.5M	545K	11M	2.3M
EN	5.1M	1M	64M	12M
ES	1.2M	211K	5.5M	1.1M
FR	2.3M	867K	18M	6.8M
IT	1.9M	217K	10M	1.2M

Table 2: The number of positive and negative triples for each language with (*) and without templates.

structed to discard it. They were also instructed to create their own templates, paraphrasing the original ones when possible. This resulted in a varying number of templates for each of the properties across languages. In addition to the entity placeholder, some languages with richer morphology (Spanish, Italian, and German) required extra placeholders in the templates because of agreement phenomena (gender). We added a placeholder for definite articles, as well as one for gender-dependent filler words. The gender is automatically inferred from the Wikipedia page statistics and a few heuristics. Table 1 shows the same example across five languages.

Dataset statistics Table 2 shows the number of positive and negative triples and examples (i.e with and without consideration of the templates).

As expected (due to the size of its Wikidata), English has the highest number of triples for most properties. However, as Figure 2 shows, there are properties where it has fewer triples than other languages (e.g. French has more triples for film related properties such as $cast\_member$ and $nominated\_for$). Figure 1 shows the overlap in the number of triples between different languages. While it can be seen that English, once again, has the highest overall overlap with the other languages, there are interesting deviations from this pattern where for certain properties other languages share a larger intersection.

4 Method

In our framework, a machine comprehension model sees a question-context pair and is tasked with selecting an answer span within the context, or indicating that the context does not contain an answer (returning NIL). This 'nil-awareness' goes beyond the traditional reading comprehension setup where it is not required. It has, however, recently been incorporated into newer datasets (Trischler et al., 2017; Rajpurkar et al., 2018; Saha
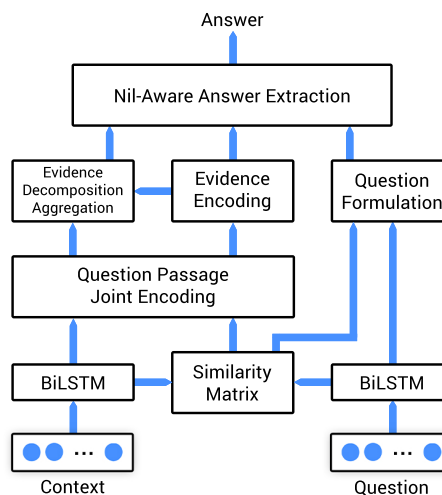

Figure 3: An overview of Namanda's architecture.

et al., 2018). We employ the architecture described in Kundu and Ng (2018) as our standard reading comprehension model for all the experiments. This nil-aware answer extraction framework (NAMANDA) is briefly described below. In a set of initial trials (see Table 3), we found that this model far outperformed the bias-augmented BiDAF model (Seo et al., 2016) used by Levy et al. (2017) on their dataset.

A Nil-aware machine comprehension model The reading comprehension model we employ, seen in Figure 3, encodes the question and context sequences and computes a similarity matrix between them. A column-wise softmax of the similarity matrix is multiplied with the question encoding to aggregate the most relevant parts of the question with respect to the context. Next, a joint-encoding of the question and context is created and a multi-factor self-attentive encoding is applied to accumulate evidence from the entire context. These representations are called the evidence vectors. Lastly, the evidence vectors are decomposed for every context word with orthogonal decomposition. The parallel components represent the relevant parts of the context and the orthogonal parts represent the irrelevant parts. These decompositions bias the decoder to either output a span or NIL.

Multilingual representations We compare two methods of obtaining multilingual representations. First, we employ fastText embeddings (Bojanowski et al., 2017) mapped to a multilingual space in a supervised fashion (Conneau et al., 2017). Second, we employ the newly released
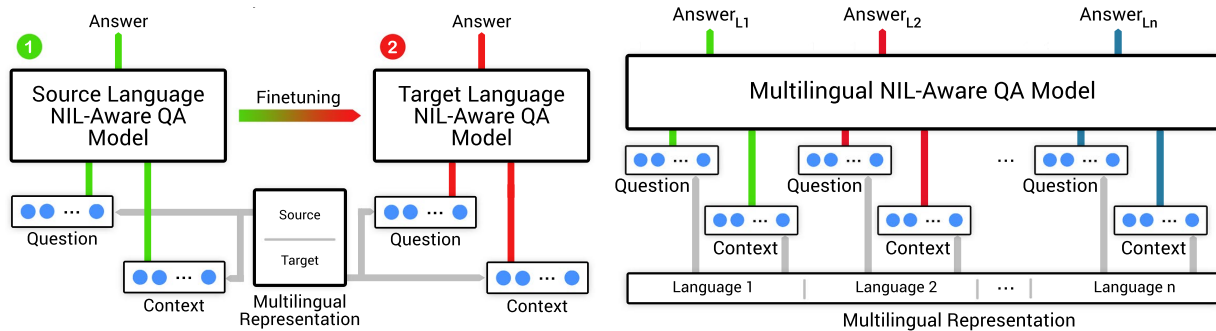
(a) Cross-lingual model transfer. In step (1), a source language model is trained until convergence. In step (2), it is finetuned on a limited amount of target language data.

(b) Joint multilingual training.

Figure 4: Our cross-lingual transfer and multilingual training setups.

multilingual BERT (Devlin et al., 2018) which is trained on the concatenation of the wikipedia corpora of 104 languages.[3] For BERT, we take the contexualized word representations from the final layer as input to our machine comprehension model's question and context Bi-LSTM encoders. We do not fine-tune the pre-trained model.

5 Experiments

Following Levy et al. (2017), we distinguish between the traditional RE setting where the aim is to generalize to unseen entities (**UnENT**) and the zero-shot setting (**UnREL**) where the aim is to do so for unseen relation types (see Section 2). Our goal is to answer these three questions: **A)** how well can RE models be transferred across languages? **B)** in the difficult **UnREL** setting, can the variance between languages in the number of instances of relations (see Figure 2) be exploited to enable more robust RE ? **C)** can one jointly-trained multilingual model which performs RE in multiple languages perform comparably to or outperform its individual monolingual counterparts? For all experiments, we take the *multiple templates* approach where a model sees different paraphrases of the same question during training. This approach was shown by Levy et al. (2017) to have significantly better paraphrasing abilities than when only one question template or simpler relation descriptions are employed.

Evaluation Our evaluation methodology follows Levy et al. (2017). We compute precision, recall and F1 by comparing spans predicted by the models with gold answers. Precision is equal to

the true positives divided by total number of non-nil answers predicted by a system. Recall is equal to the true positives divided by the total number of instances that are non-nil in the ground truth answers. Word order and punctuation are not considered.[4]

5.1 Monolingual Baselines

A baseline model is trained on the full monolingual training set (1 million instances) for each of the languages in both the **UnENT** and **UnREL** settings, which serve as a point of comparison for the cross-lingual transfer and multilingual models.

Comparison with Levy et al. (2017) In Table 3, the comparison between the nil-aware machine comprehension framework we employ (Mono) and the results reported by Levy et al. (2017) using the bias-augmented BiDAF model on their dataset (and splits) can be seen. The clear improvements obtained are in line with those reported by Kundu and Ng (2018) of NAMANDA over BiDAF on reading comprehension tasks.

Results Table 3 shows the results of the monolingual baselines. For the cross-lingual transfer experiments, these results can be viewed as a performance ceiling.

Observe that the results on our dataset are in general lower than those reported in Levy et al. (2017). This can be attributed to three factors: a) on average, the context length in our dataset is longer compared to theirs; b) the fastText word embeddings we employ to facilitate multilingual

---

[3]https://github.com/google-research/bert/blob/master/multilingual.md

---

[4]We do not exclude articles from the evaluation as separating them from entities is not as trivial for other languages as it is for English.
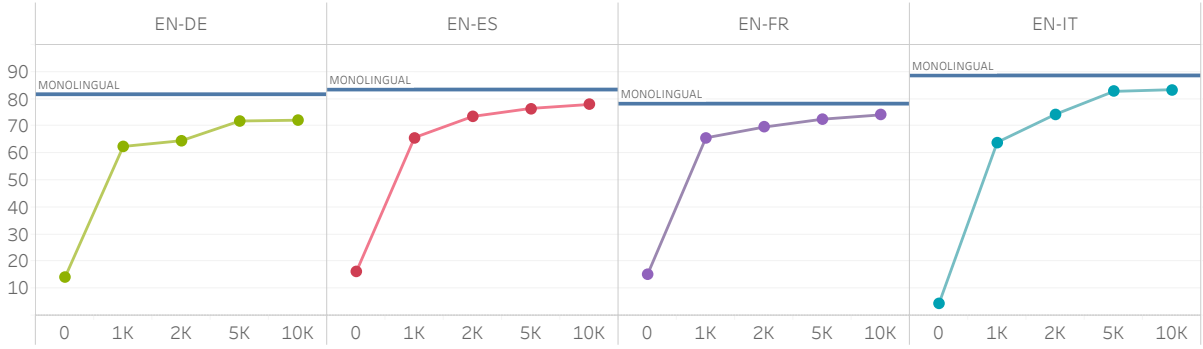
Figure 5: F1-scores for the cross-lingual transfer experiments in the **UnENT** setting. The MONOLINGUAL line shows the corresponding monolingual model's F1-score.

sharing have a lower coverage of the vocabularies of each language than the GloVe word embeddings employed in that work; c) in the **UnREL** setting, we employ a more challenging setup of 5-fold cross-validation (as opposed to 10-fold in their experiments), meaning that a lower number of relations is seen at training time and the test set contains a higher number of unseen relations.

5.2 Cross-Lingual Model Transfer

In this set of experiments, seen in Figure 4a, we test how well RE models can be transferred from a source language with a large number of training examples to target languages with no or minimal training data. In the **UnENT** experiments, we construct pairwise parallel test and development sets between English and each of the languages. An English RE model (built on top of the multilingual representations described in sub-section 4) is trained on a full English training set (1 million instances). We then evaluate how well this model can transfer to each of the four other languages in the following cases: with no finetuning or when 1000, 2000, 5000 or 10000 target language training examples are used for finetuning. Note that entities in the target languages' test and development sets are not seen in the English training data. We compare transfer performance with monolingual performance when a target language's full training set is employed.

A similar approach is followed for **UnREL** experiments. However, since the number of relations is relatively small, cross-validation with five folds is employed instead of fixed splits. Moreover, because this is a substantially more challenging setting we are interested in evaluating along another dimension (Question **B**): when relations are seen in the source language but not in the target language. Furthermore, unlike for **UnENT**, we directly use 10k examples for finetuning.

Results Figure 5 shows the results of the cross-lingual transfer experiments for **UnENT**, where transfer is accomplished through multilingually aligned fastText embeddings. In a parallel set of experiments, transfer was performed through the multilingual BERT encoder. The results of this showed a clear advantage for the former over the latter.[5] This is primarily due to the low vocabulary coverage of multilingual BERT which has a total vocabulary size of 100k tokens for 104 languages for coverage statistics). While it is clear that the models suffer from rather low recall when no finetuning is performed, the results show considerable improvements when finetuning with only 1000 target language examples. With 10K target language examples, it is possible to nearly match the performance of a model trained on the full target language monolingual training set.

Similarly, in the **UnREL** experiments, our results (Figure 6) show that it's possible to recover a large part of the fully-supervised monolingual models' performance. It can be seen, however, that with 10k target language examples, a lower proportion of the performance is recovered when compared to the **UnENT** setting. This indicates that it is more difficult to transfer the ability to identify *relation paraphrases* and *entity types* through global cues[6] which Levy et al. (2017) suggested are important for generalizing to new relations in this framework.

---

[5]We therefore continue the rest of our experiments in the paper using the multilingual fastText embeddings.

[6]When context phrasing deviates from the question in a way that is common between relations.
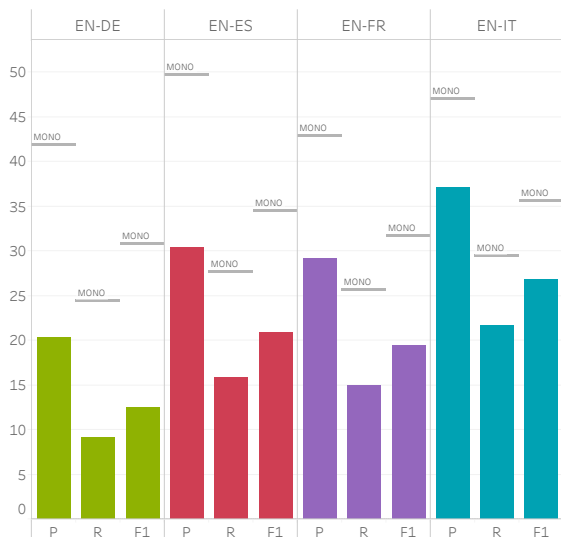
Figure 6: Precision, Recall and F1-scores for the cross-lingual transfer experiments in **UnREL** setting. The results are the mean of 5-fold cross-validation. The MONO line shows the corresponding monolingual model's F1-score.

5.3 One Model, Multiple Languages

We now examine the possibility of training one multilingual model which is able to perform relation extraction across multiple languages, as shown in Figure 4b. We are interested in the case when an entity may be seen in another language's training data, as this is a realistic cross-lingual KB completion scenario where different languages' KBs are better populated for different topics. To control for training set size we include 200k training instances per language, so that the total size of the training set is equal to that of the monolingual baseline. However, an additional benefit of multilingual training is that extra overall training data becomes available. To test the effect of that we also run an experiment where the full training set of each of the languages is employed (adding up to 5 million training examples).

In the **UnREL** experiments, 5-fold cross-validation is performed. We are once again interested in exploiting the fact that KBs are better populated for different properties across different languages. Our setup is therefore as follows: in each of the 5 folds, a test set relation for a particular language is not seen in that language's training set, but may be seen in any of the other languages. This amounts to maintaining the original zero-shot setting (where a relation is not seen) monolingually, but providing supervision by allowing the models to *peek across languages*.

Results In the **UnENT** setting the multilingual models trained on just 200k instances per language perform slightly below the monolingual baselines. This excludes for French where, surprisingly, the baseline performance is actually exceeded. When the full training sets of all languages are combined, the multilingual model outperforms the monolingual baselines for three (English, Spanish, and French) out of five languages and is slightly worse for two (German and Italian). This demonstrates that not only is it possible to utilize a single model to perform RE in multiple languages, but that the multilingual supervision signal will often lead to improvements in performance. These results are shown in the third and fourth columns of Table 3.

The multilingual **UnREL** model outperforms its monolingual counterparts by large margins for all languages reaching a near 100% F1-score improvement for most languages. This is largely in line with our premise that the natural topicality of KBs across languages can be exploited to provide cross-lingual supervision for relation extraction models.

5.4 Hyperparameters

In all experiments, models were trained for five epochs with a learning rate of 0.001 using Adam (Kingma and Ba, 2014). For finetuning in the cross-lingual transfer experiments, the learning rate was lowered to 0.001 to prevent forgetting and a maximum of 30 finetuning iterations over the small target language training set were performed with model selection using the target language development set F1-score. All monolingual models' word embeddings were initialised using fastText embeddings trained on each language's Wikipedia and common crawl corpora,[7] except for the comparison experiments described in sub-section 5.1 where GloVe (Pennington et al., 2014) was used for comparability with Levy et al. (2017).

6   Related Work

Multilingual NLU Advances in natural language understanding tasks have been as impressive as they have been fast-paced. Until recently, however, the multilingual aspect of such tasks has not received as much attention. This is primarily due to the costs associated with annotating data for multiple languages. Recent work such as Conneau et al. (2018); Agic and Schluter

---

[7]https://fasttext.cc/docs/en/crawl-vectors.html

Lang.		UnENT				UnREL		
		Levy et al. (2017)	Mono.	Multi. (S)	Multi. (L)	Levy et al. (2017)	Mono.	Multi.
EN*	P	87.66	**90.49**	n/a	n/a	43.61	**56.53**	n/a
	R	91.32	**94.87**	n/a	n/a	36.45	**44.74**	n/a
	F1	89.44	**92.63**	n/a	n/a	39.61	**49.85**	n/a
EN	P	n/a	74.09	74.33	**77.11**	n/a	46.75	**63.29**
	R	n/a	85.35	83.63	**86.42**	n/a	25.32	**44.40**
	F1	n/a	79.32	78.71	**81.50**	n/a	32.78	**51.99**
ES	P	n/a	81.79	80.60	**83.68**	n/a	49.77	**73.43**
	R	n/a	**85.02**	81.47	83.58	n/a	27.69	**62.82**
	F1	n/a	83.37	81.03	**83.63**	n/a	34.54	**67.64**
IT	P	n/a	**88.69**	86.23	88.43	n/a	47.09	**68.66**
	R	n/a	**88.10**	85.64	86.91	n/a	29.45	**55.24**
	F1	n/a	**88.39**	85.93	87.66	n/a	35.62	**61.13**
FR	P	n/a	82.36	80.82	**82.90**	n/a	42.93	**60.78**
	R	n/a	74.16	76.60	**78.10**	n/a	25.73	**47.09**
	F1	n/a	78.05	78.66	**80.43**	n/a	31.78	**53.06**
DE	P	n/a	**75.85**	69.88	73.67	n/a	41.94	**43.36**
	R	n/a	**88.21**	81.36	84.08	n/a	24.38	**25.32**
	F1	n/a	**81.57**	75.20	78.53	n/a	30.82	**31.97**

Table 3: Precision, Recall, and F1-score results for all languages' monolingual (Mono.) and multilingual (Multi.) models. (S) indicates the small multilingual model which was trained on 200k examples and (L) indicates the large on trained on 5 million examples. * is used to mark the results on Levy et al. (2017)'s English dataset.

(2018) offer important benchmarks for evaluating cross-lingual transfer of natural language inference models. Similarly, Cer et al. (2017) present the Semantic Textual Similarity dataset for four languages.

Multilingual relation extraction Previous investigations of multilingual RE have been few and far between. Faruqui and Kumar (2015) employed a pipeline of machine translation systems to translate to English, then Open RE systems to perform RE on the translated text, followed by cross-lingual projection back to source language. Verga et al. (2016) apply the universal schema framework (Riedel et al., 2013) on top of multilingual embeddings to extract relations from Spanish text without using Spanish training data. This approach, however, only enables generalization to unseen entities and does not have the flexibility to predict unseen relations. Furthermore, both of these works faced a fundamental difficulty with evaluation. The former resort to manual annotation of a small number of examples (1000) in each language and the latter use the 2012 TAC Spanish slot-filling evaluation dataset in which "*the coverage of facts in the available annotation is very small*". With the introduction of **X-WikiRE**, this work provides the first large-scale dataset and benchmark for the evaluation of multilingual RE

spanning five languages. While this paves the way for a wide range of research on multilingual relation extraction and knowledge base population, we hope to extend this to a larger variety of languages in future work, particularly as we have been able to show that the amount of training data required for cross-lingual model transfer is minimal, meaning that a small dataset (when only that is available) can go a long way.

7 Conclusion

We introduced **X-WikiRE**, a new, large-scale multilingual relation extraction dataset in which relation extraction is framed as a problem of reading comprehension to allow for generalization to unseen relations. Using this, we demonstrated that a) multilingual training can be employed to exploit the fact that KBs are better populated in different areas for different languages, providing a strong cross-lingual supervision signal which leads to considerably better zero-shot relation extraction; b) models can be transferred cross-lingually with a minimal amount of target language data for fine-tuning; c) better modelling of nil-awareness in reading comprehension models leads to improvements on the task. Our work is a step towards making KBs equally well-resourced across languages. To encourage future work in this direction, we release our code and dataset.

References

Zeljko Agic and Natalie Schluter. 2018. Baselines and Test Data for Cross-Lingual Inference. In *LREC*. European Language Resources Association (ELRA).

Nguyen Bach and Sameer Badaskar. 2007. A Review of Relation Extraction.

Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *IJCAI*, volume 7, pages 2670–2676.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity-Multilingual and Cross-lingual Focused Evaluation. *arXiv preprint arXiv:1708.00055*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie M Strassel. 2015. Overview of Linguistic Resources for the TAC KBP 2015 Evaluations: Methodologies and Results. In *TAC*.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying Relations for Open Information Extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545. Association for Computational Linguistics.

Manaal Faruqui and Shankar Kumar. 2015. Multilingual Open Relation Extraction Using Cross-lingual Projection. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1351–1356. Association for Computational Linguistics.

Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. 2016. WikiReading: A Novel Large-scale Language Understanding Task over Wikipedia. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1545. Association for Computational Linguistics.

Lucie-Aimée Kaffee and Elena Simperl. 2018. Analysis of Editors' Languages in Wikidata. In *OpenSym*, pages 21:1–21:5. ACM.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Souvik Kundu and Hwee Tou Ng. 2018. A Nil-Aware Answer Extraction Framework for Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4243–4252. Association for Computational Linguistics.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-Shot Relation Extraction via Reading Comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342. Association for Computational Linguistics.

Makoto Miwa and Mohit Bansal. 2016. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116. Association for Computational Linguistics.

Abiola Obamuyide and Andreas Vlachos. 2018. Zero-shot relation classification as textual entailment. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 72–78.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. 2013. Relation Extraction with Matrix Factorization and Universal Schemas.

In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 74–84. Association for Computational Linguistics.

Tim Rocktäschel, Sameer Singh, and Sebastian Riedel. 2015. Injecting logical background knowledge into embeddings for relation extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1119–1129.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. DuoRC: Towards Complex Language Understanding with Paraphrased Reading Comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1693. Association for Computational Linguistics.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A Machine Comprehension Dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200. Association for Computational Linguistics.

Patrick Verga, David Belanger, Emma Strubell, Benjamin Roth, and Andrew McCallum. 2016. Multilingual Relation Extraction using Compositional Universal Schema. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 886–896. Association for Computational Linguistics.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledge base.

Alexander Yates, Michele Banko, Matthew Broadhead, Michael J Cafarella, Oren Etzioni, and Stephen Soderland. 2007. TextRunner: Open Information Extraction on the Web. In *HLT-NAACL (Demonstrations)*, pages 25–26.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *Journal of Machine Learning Research*, 3:1083–1106.

GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. Exploring Various Knowledge in Relation Extraction. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 427–434. Association for Computational Linguistics.