Multilingual Entity and Relation Extraction Dataset and Model

Alessandro Seganti[*2], Klaudia Firląg[1], Helena Skowrońska[*3],
Michał Satława[1], and Piotr Andruszkiewicz[1,4]

[1]Samsung R&D Institute Poland
[2]Equinix
[3]NextSell, ODC Group
[4]Warsaw University of Technology
alessandro.seganti@gmail.com
{k.firlag, m.satlawa, p.andruszki2}@samsung.com
arvala@wp.pl, p.andruszkiewicz@ii.pw.edu.pl

Abstract

We present a novel dataset and model for a multilingual setting to approach the task of Joint Entity and Relation Extraction. The SMiLER dataset consists of 1.1 M annotated sentences, representing 36 relations, and 14 languages. To the best of our knowledge, this is currently both the largest and the most comprehensive dataset of this type. We introduce HERBERTa, a pipeline that combines two independent BERT models: one for sequence classification, and the other for entity tagging. The model achieves micro F_1 81.49 for English on this dataset, which is close to the current SOTA on CoNLL, SpERT.

1 Introduction

The majority of the – constantly growing – amount of openly accessible knowledge is locked in unstructured text, and hence inefficiently utilized by any systems. The NLP tasks related to this problem include Information Extraction, Relation Extraction, Named Entity Retrieval, as well as Joint Entity and Relation Extraction.

Our contribution is twofold. First, we present SMiLER (Samsung MultiLingual Entity and Relation Extraction dataset): an open-domain corpus of annotated sentences, created for the Joint Entity and Relation Extraction task. With 1.1 M sentences, 36 relation types, and 14 languages, SMiLER seems to be both the largest and the most diversified corpus for the task in existence, to the best of our knowledge. The corpus was semi-automatically created from Wikipedia and DBpedia, and partly checked by linguists.

Our second contribution is HERBERTa – Hybrid Entity and Relation extraction BERT for a multilingual setting that consists of two independently pretrained BERT models (Devlin et al., 2018). The first one classifies the input sequence as belonging to one of our 36 relations (including *no_relation*). Its output – the relation – is then fed to the second BERT, together with the same input sequence. The second model performs entity tagging and outputs two spans for the two entities selected from the input sequence. Our model is close to SpERT (Eberts and Ulges, 2019), the current state-of-the-art for Joint Entity and Relation Extraction, achieving micro F_1 81.49 for English on the presented dataset.

The SMiLER corpus and the source code is available at https://github.com/samsungnlp/smiler/.

2 Related Work

Available Datasets The datasets that are commonly used for the task of Joint Entity and Relation Extraction are still insufficient in size and diversification; furthermore, they are all monolingual English corpora. For instance, CONLL04 (Roth and Yih, 2004) has only 1.7k annotated sentences and 5 relations. ADE (Gurulingappa et al., 2012) is larger, with almost 21k sentences, but it distinguishes only 2 relations (plus no_relation). Yet another such dataset, SciERC (Luan et al., 2018), contains 500 annotated abstracts and 7 relations; however, all of the relations belong to the scientific domain.

We believe that presenting a new, large and diversified dataset will be a valuable contribution to the joint task of Entity and Relation Extraction – both for English and for the multilingual setting.

Information Extraction Information extraction systems collect knowledge that is locked in unstructured text, such as in our Wikipedia articles. A nota-

[*] Work done while at Samsung R&D Institute Poland.

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 1946–1955
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

ble example is the Never-Ending Language Learner (NELL), which was reading the Web for almost 10 years, 2010-2019 (Mitchell et al., 2018). This semi-supervised system was retrained continuously, with the use of the current knowledge, to collect new instances of a pre-defined set of entity types and relations, which also constitutes the final goal of our system. Another well-known knowledge base is Knowledge Vault (Dong et al., 2014). The system extracts information from the Internet (text, tabular data, page structure, human annotations) and combines it with information from Freebase (Bollacker et al., 2008). A probabilistic inference system computes calibrated probabilities of fact correctness.

Instead of utilizing pre-defined entity and relation types, some systems use syntactic analysis: ReVerb (Fader et al., 2011), MinIE (Gashteovski et al., 2017). Another approach is to create a taxonomy with just one relation: isA, as in Probase (Wu et al., 2012).

Relation Extraction Current models (HEBERTa included) for Relation Extraction are based on BERT (Devlin et al., 2018), a multi-layer bidirectional Transformer encoder (Vaswani et al., 2017). Its main contribution is pre-training deep bidirectional representations from unlabeled text by joint conditioning on both left and right context, in all the layers.

The current state-of-the-art in Relation Extraction are REDN (on SemEval-2010 Task 8, NYT, and WebNLG) (Li and Tian, 2020) and Matching the Blanks (on FewRel) (Soares et al., 2019). REDN extracts the token embeddings (BERT) from two different layers to represent the head and tail entities separately, in order to enhance learning reversible relations. Next, it calculates a parameterized asymmetric kernel inner product matrix between all the head and tail embeddings of each token in a sequence. On the other hand, the model constructed by Soares et al. (2019) combines BERT with extensions of Harris' distributional hypothesis to relations, to build task-agnostic relation representations solely from entity-linked text.

Joint Entity and Relation Extraction A more robust approach to Relation Extraction is to combine it with Entity Extraction. Three such joint solutions have achieved the current state-of-the-art: SciBERT (on SciERC) (Beltagy et al., 2019), SpERT (for RE on CONLL04) (Eberts and Ulges, 2019), and End2End Joint NER & RE (for NER on CONLL04) (Giorgi et al., 2019).

SciBERT leverages unsupervised pretraining on a multi-domain corpus of scientific publications, as this domain differs significantly from the general domain used by the original BERT. The second model, SpERT, is an attention model for Span-Based Joint Entity and Relation Extraction. It is trained using within-sentence negative samples, which are extracted in a single BERT pass. Finally, End2End Joint NER & RE combines BERT with biaffine attention.

Distant Supervision *Distant supervision* proposed by Mintz et al. (2009) for Relation Extraction task is a training paradigm, based on the assumption that if there is a relation between entities, then every sentence containing them may also express that relation. In the experiments, multiple sentences containing the entities were used to create their feature vectors. Thus, this approach allows a prediction of the relation between two entities, but does not identify sentences containing useful and correct cues for this relation in the corpus.

Data collection for SMiLER was based on the same assumption, but thanks to the subset of manually validated examples, the dataset could be used as the sole source of supervision during training, where the algorithm was supervised by the database.

Multilingual BERT M-BERT (Pires et al., 2019) is a model that is particularly relevant to our multilingual approach, and one that we tested repeatedly. This is a single language model, pre-trained from monolingual corpora in 104 languages, which performs zero-shot cross-lingual model transfer. Task-specific annotations in one language are used to fine-tune the model for evaluation in another language. Such transfer is possible even to languages in different scripts, such as English and Korean.

M-BERT has become the basis for building other multilingual models, as well as monolingual ones for languages other than English. We have experimented with three such models: German BERT, Italian BERT, and KoBERT for Korean (see section 4 for details).

3 Data

Our SMiLER corpus consists of 1.1 M annotated Wikipedia sentences, and was created specifically for the task of Entity and Relation Extraction. It

	EN-full	EN-mid	EN-small	KO	IT	FR	DE	PT	NL	PL	ES	AR	RU	SV	FA	UK	Total
sentences	748k	269k	35k	20k	76k	62k	53k	45k	40k	17k	12k	9k	7k	5k	3k	1k	1.1M
relations	36	36	32	28	22	22	22	22	22	22	22	9	8	22	8	7	36

Table 1: SMiLER: number of sentences and relations for each language. EN-mid includes EN-small, and EN-full includes EN-mid.

consists of annotated sentences in 14 languages, with each language representing a subset of 36 relations. As far as we know, this is both the biggest and the most comprehensive corpus for these NLP tasks. Table 1 provides the number of sentences and relations for each language, while Table 2 shows a few examples from the corpus.

The relations belong to 9 rough domains: Person (e.g. has-child, has-occupation), Organization (e.g. org-has-member, headquarters), Location (e.g. loc-leader, has-tourist-attraction), Animal (e.g. has-lifespan, eats), Art (e.g. starring, has-author), Device (e.g. invented-by), Event (event-year), Measurement (e.g. has-length), and no_relation. The no_relation sentences do not contain any of the 35 "positive" relations.

Figure 1 shows the number of sentences for each relation, for each language.

Relation	Sentence
has-child	**[Bill]**$_{e1}$ married Hillary on October 11, 1975, and their only child, **[Chelsea]**$_{e2}$, was born on February 27, 1980.
head-quarters	**[AMC Airlines]**$_{e1}$ è una compagnia aerea egiziana con sede al **[Cairo]**$_{e2}$, esegue voli charter da Sharm el-Sheikh, Hurghada, Il Cairo verso le maggiori capitali europee dal maggio 2006.
movie-has-director	**[Lili]**$_{e1}$ ist ein US-amerikanischer Spielfilm des Regisseurs **[Charles Walters]**$_{e2}$ aus dem Jahr 1953.

Table 2: Examples from the SMiLER corpus.

3.1 Dataset Building

To collect the English dataset, we queried DBpedia for (entity$_1$, entity$_2$, relation) triples. The articles about entity$_1$ that would also contain entity$_2$ were obtained from a Wikipedia dump. They were parsed and automatically selected. The complete English dataset is referred to as EN-full in the following discussion.

During this process it is possible that the same sentence will appear multiple times containing:

1. different entity annotations for the same relation, or

2. different entity annotations for another relation.

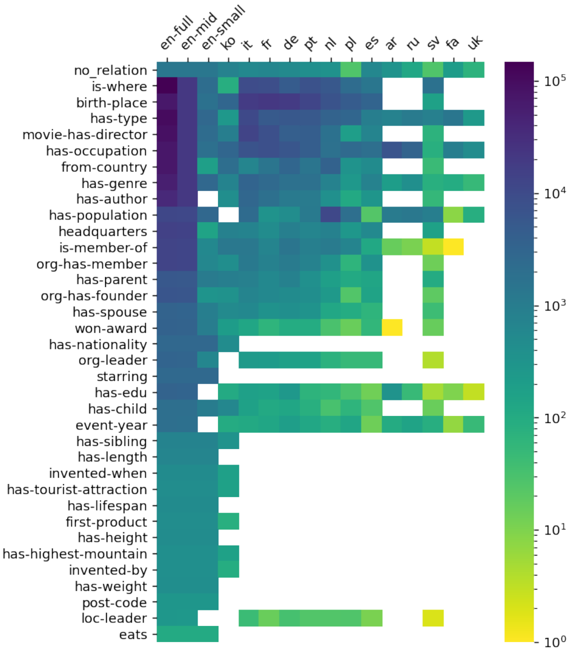

Figure 1: SMiLER: the number of sentences for each relation and for each language.

The first case is unlikely, because sentences are taken from the Wikipedia article of the "main" entity. The second case is possible, though.

Theoretically, it is possible to write grammatical rules in order to do a "cross-match" for augmenting a number of annotated sentences, and it has been done previously (e.g. Wu et al. (2012)). The reason why we did not use grammatical rules is that we wanted to scale the approach to multiple (and diverse) languages and the rules that could help with this task are not universal.

A part of EN-full was manually validated by linguists. The linguistic validation was more in-depth for English than for other languages, as for English the task was to correct the entity annotation whenever possible. Hence, passable annotations were corrected for English, while they were simply assessed as correct for the other languages. Overall, the linguists assessed 58.2% of the sample as fully correct, corrected the annotations in further 18.4%, and assessed the remaining 23.4% as incorrect. The correct 58.2% plus the corrected 18.4% of the sample together form the EN-small dataset.

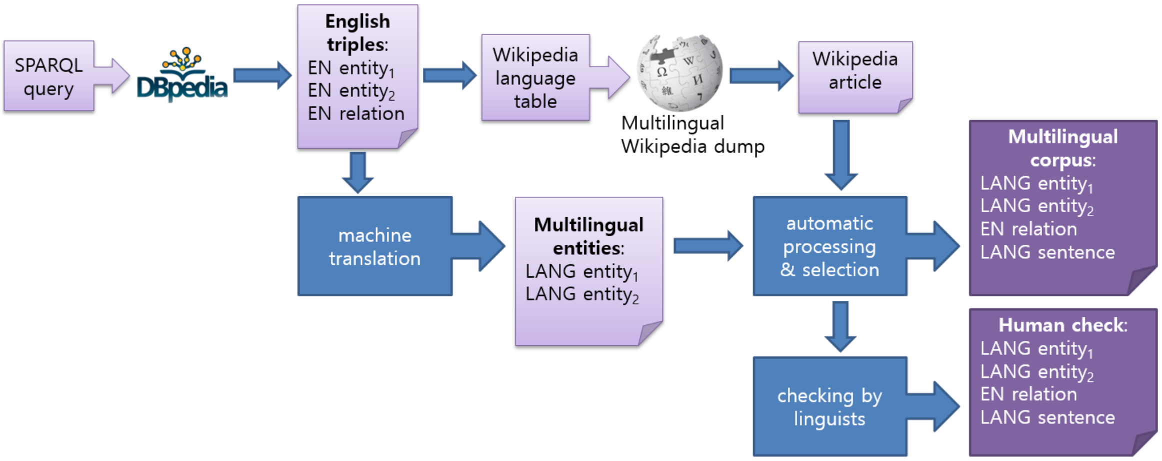

Figure 2: The process of creating the multilingual parts of SMiLER.

The corpus-building process was adapted for the other languages. First, the entities (as well as the entire no_relation sentences) were translated from English. The translation was carried out by an in-house multilingual system similar to (Przybysz et al., 2017; Williams et al., 2018; Wetesko et al., 2019). After translating the entities, the relevant article was copied from the Wikipedia dump in the other language. Finally, samples from the corpus were checked by linguists. Figure 2 illustrates the process.

The no_relation sentences in English have been selected by linguists from Wikipedia articles. The sentences contained entities that were in a relation in another sentence in the data set but were not in a relation in the selected sentence. Since the process is so strict, the translation to other languages was done automatically. All no_relation sentences needed to be hand-checked, because they could contain some hidden relations, and for this reason preparing original sentences for all the languages would take too much time. This is not the case for the 'positive' relations; e.g. if 'Bill' is the father of 'Chelsea', then a sentence containing both 'Bill' and 'Chelsea' – and coming from a Wikipedia article on Bill Clinton – would probably also express the relation has-child. However, if a sentence contains 'Bill', but not 'Chelsea', it could still express another of the "positive" relations.

We also tried other automated approaches for finding no_relation sentences. For example, given a Wikipedia article with a relation found, we could find other sentences with the same entities and assume all of them are no_relation sentences. Unfortunately this approach caused too much noise and was not used in the end.

The linguists verified a random sample of 50 sentences for 16 largest relations + no_relation for each of the following languages: IT, FR, DE, PT, ES, KO. The results are shown in Table 3. The overall percentage of correctness reached 79-80%, which is very high, considering the level of automation involved in the process. The selected random sample was small, because the verification needed to be cost- and time-effective. Still, the results were thoroughly checked and they were largely consistent.

	IT	FR	DE	PT	ES	KO	Overall
positive relations (scraped)	79	70	76	84	84	78	79
no_relation (translated)	88	94	86	86	74	53	80

Table 3: The percentage of correct sentences for selected languages in SMiLER.

The most common errors found by the linguists belong to three groups: (1) unexpected DBpedia query results, (2) wrong sentence parsing (deleting the second half of the sentence or including HTML), (3) selecting random sentences, which either do not express the relation, or have the entities marked in wrong places. Other typical errors are related to missing translations of some English words and *no_relation* sentences that contain one of the 35 positive relations. Table 4 shows one example of each error.

3.2 Datasets for the Model

For each language in our multilingual dataset, we automatically extracted 2% of the sentences to create a dev set and another 2% for a test set. This corpus split maintained the distribution of the rela-

Relation	Error	Sentence with an error
has-type	unexpected type "cardinal" obtained from DBpedia	[**Gerhard Ludwig Müller**]$_{e1}$ (Finthen, 31 dicembre 1947) è un [**cardinale**]$_{e2}$, arcivescovo cattolico e teologo tedesco, prefetto emerito della Congregazione per la dottrina della fede dal 1^{o} luglio 2017.
headquarters	random sentences and deleted words	Temple des Martyrs à [**Taipei**]$_{e2}$. Les [**Forces armées de la république de Chine**]$_{e1}$ sont constituées d'une force d'active d'environ et de .
has-type	HTML included	vignette|gauche|250px|La salle de concerts de Raanana [**Ra'anana**]$_{e1}$ est une [**ville**]$_{e2}$ israélienne d'environ habitants au nord-est de Tel Aviv, dans le sud de la région de Sharon.
no_relation	untranslated "ranks"	[**iSuppli**]$_{e1}$ ranks [**Kingston**]$_{e2}$ como el fabricante de módulos de memoria número uno del mundo para el mercado de memoria de terceros por el décimo año consecutivo.
no_relation	sentence contains relation has-genre	Nach der Registrierung seines aktuellen YouTube-Kanals 2010 veröffentlichte [**Kjellberg**]$_{e1}$ vor allem [**Let's Play**]$_{e2}$ Videos von Horror- und Action-Videospielen.

Table 4: Most common error types found by linguists.

tions in the original data, whenever possible. For the languages where the label distribution was impossible to preserve, the number of relations in the test set may be smaller than in the train set.[1]

For English, we used a single test set for both EN-mid and EN-small. The only difference was in the number of labels: EN-mid (both train and dev) had 36 labels, while EN-small (albo both train and dev) – 32 (see Table 1).

For training the models, we created several combinations of languages, in order to examine the effect of adding/removing languages from the training set. We experimented with the following combinations:

1. **EURO**: IT, FR, PT, DE, ES, EN
2. **SVO**: **EURO**, RU, SV, NL, PL, UK
3. **ALL**: **SVO**, AR, KO, FA
4. Each individual language combined with EN (e.g. IT+EN)
5. Each language alone (e.g. IT)

Pires et al. (2019) showed that the language structure had an impact on the performance of the model. Therefore, we treat Korean, Farsi and Arabic as a special case because they are non-SVO languages (Korean and Farsi are SOV, while Arabic is VSO).[2]

4 Model

Our architecture HERBERTa (Figure 3) uses two pre-trained BERT models. It solves the problem of Entity and Relation Extraction with the use of an unconventional pipeline. In the first step it is trained for Relation Extraction, while the entities are retrieved in the second stage.

The first model is BERT, fine-tuned for the Sequence Classification task. Its input is a tokenized sequence, while its output consists of a sequence output and a pooled output that represents the overall sentence context (from the [CLS] token). The latter is passed to a softmax for classifying the relation.

The second model is our implementation of BERT for Entity Tagging. It is based on BERT for Question Answering, and its inputs are:

- the same tokenized sequence as the one used for Relation Extraction,
- the relation outputted in the first step – encoded as a BERT [unused] token.

As a result, the input sequence is as follows: [CLS] [relation] [SEP] [E0] [E1] ... [En] [SEP]. The output of the model is a set of four indices that correspond to the spans of the two entities having the relation.

At inference time, the model will return the start and end of the two entities separately and this is used to mark the entities. Using our model, it is also possible to select N best predictions (one prediction = one entity pair) for the same relation. This has not been used for the result of the paper because we wanted to find a single best entity pair for each sentence.

[1] Train/test set number of relations for the languages in which the numbers differ: ES (train: 21, test: 16), RU (train: 8, test: 7), SV (train: 22, test: 14), FA (train: 8, test: 4), AR (train: 9, test: 7), NL (train: 22, test: 21), PL (train: 21, test: 20), UK (train: 7, test: 6), KO (train: 28, test: 26).

[2] SVO, SOV, and VSO stand for the relative position of the Subject, Verb, and Object in the typical affirmative sentence.

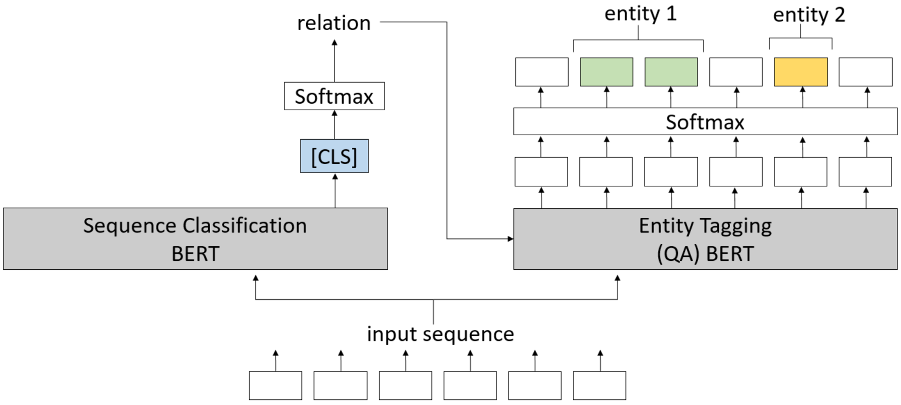

Figure 3: Model architecture.

We used the following types of BERT, depending on the language(s):

- bert-base-cased (Devlin et al., 2018),
- bert-base-multilingual-cased (Pires et al., 2019),
- bert-base-german-cased from the HuggingFace Transformer library (Wolf et al., 2019),
- bert-base-korean-cased – the monologg/kobert model from HuggingFace,
- bert-base-italian-cased – the dbmdz/bert-base-italian-cased model from HuggingFace.

Below we call our model LANG(B) if it is a fine-tuning of bert-base, e.g. DE(B) is the bert-base-german model fine-tuned on the German corpus. We call our model just LANG if it is a fine-tuning of bert-base-multilingual, e.g. DE-EN is the English and German model trained using bert-base-multilingual. The English dataset has two versions: EN (EN-mid) and EN_S (EN-small). We did not use EN-full for the training.

4.1 Training Procedure

The two models presented in Section 4 are trained independently, each one with a different loss function. The first model uses standard cross-entropy as its loss for relation classification. The loss function for the second model is:

$$\mathcal{L} = \frac{1}{4}(\mathcal{L}^{1,start} + \mathcal{L}^{1,end} + \mathcal{L}^{2,start} + \mathcal{L}^{2,end}) \quad (1)$$

Where $\mathcal{L}^{j,start}$ is the cross-entropy loss for the prediction of j-th entity first token index and $\mathcal{L}^{j,end}$ is the cross-entropy of the predictions of its last token index.

The result of the first model becomes the input to the second one, during inference. We conclude that this works well because the performance of the relation classification model is very high (please refer to Table 5 and the Relation column).

The number of epochs is established using an early stopping mechanism. On average, each model (Relation Classifier and Entity Tagger) is trained for around 6 epochs for models with EN-mid set and 3-4 epochs for others. We fine-tune the whole BERT.

The model with all languages (ALL_EN) runs on 6 GPU GeForce RTX 2080 (8GB) for 31h.

5 Results

For evaluating the models, we use the micro F_1 score on the entities and relations together, ignoring no_relation. Below, this measure is referred to as Combined. When comparing with the state-of-the-art, we also show the relation-only micro F_1, as well as the entity-only micro F_1 (for a single entity and for the pair of entities). We use such metrics because they are typically calculated for these tasks; see for instance Eberts and Ulges (2019).

5.1 Comparison with the State-of-the-Art

Given that our dataset is new, there is no state-of-the-art for it. Nevertheless, we decided to compare the results of HERBERTa with SpERT (Eberts and Ulges, 2019), the currently best model on CoNLL (Roth and Yih, 2004). Table 5 presents the results that we obtained by training SpERT on our Wikipedia EN-mid (denoted as EN(SpERT)) and EN-small (denoted as EN_S(SpERT)). Additionally, Table 5 also shows the results of our model trained on the same datasets. In terms of combined F_1, HERBERTa is close, the difference being about 0.2

percentage point with SpERT, the current SOTA on CONLL.

Model	Relation	Entities (Pair)	Combined
EN_S(SpERT)	N/A	80.71	59.24
EN(SpERT)	N/A	**92.89**	**81.71**
EN_S(B)	80.94	61.94	58.31
EN(B)	**94.94**	82.55	81.49
ALL-EN_S	81.64	57.46	53.76
ALL-EN	93.85	78.26	76.97

Table 5: Micro F_1 results on the Wikipedia EN-mid dataset. Combined – both the relation and the entity pair are correct.

5.2 Language and Model Comparison

Single Language As the first step, we trained HERBERTa for each language separately. The results (F_1 measure) are presented in Figure 4 (left). Our models trained with the use of multilingual BERT are named after language, e.g. IT, FR, PT. Our non-multilingual BERT models are called LANG plus (B), e.g. IT(B), DE(B).

We obtained the best result for EN(B). BERT trained on EN-small (EN_S(B)) achieved lower results, 81 vs. 58, because EN-small contains over 7 times fewer sentences. If we compare this with the other languages, the relationship between F_1 and the logarithm of the number of sentences holds in general. That is, the higher number of sentences, the higher F_1 (see Figure 5). EN-small (35k sentences, F_1=58) falls between NL (40k sentences, F_1=60) and PL (17k sentences, F_1=50). RU (7k) and UK (1k, the smallest set) achieve F_1=29 and 10 respectively.

Surprisingly, FA (3k) and SV (5k) – despite small sets – yield F_1=65 and 58 respectively; as a result, they are the outliers in Figure 5. AR (9k) achieves a higher F_1 score than ES (12k); however, the difference is just 2 percentage points. FA, AR (and KO) have different word orders than all the other languages, which might be one of the reasons why they achieve high F_1 despite small datasets. Another reason might be their small number of relations. The high result for SV, which is SVO, could be explained by a smaller number of relations in the test set than in the train set. UK and RU, in spite of their small number of relations, obtain the lowest results, because they have small datasets. Furthermore, contrary to SV, UK and RU have rich inflection which might also be the case.

Comparing the results of our models trained with multilingual BERT versus non-multilingual BERT, we observe that all three possibilities are present in the pairs we checked. For DE we get the same results F_1=58, for IT the non-multilingual BERT gives lower results 64 vs. 66, while for KO the non-multilingual BERT achieves significantly higher results, 51 vs. 42.

Multilingual In the second step, we added models trained with the multilingual BERT for several languages simultaneously, e.g. ALL-EN for all the languages available in the corpus (see Figure 4 (left)). There are four main groups of models: (A) a model with EN-mid and all other languages, (B) models with EN-small and at least 6 languages, (C) models with EN-small and one other language, (D) other models with about 3-4 similar languages.

Comparing the models for single languages with group A and B models, we observe that the multilingual models typically yield similar or better results. The largest increase was observed for UK: F_1=10 on its own and even 26 in ALL-EN_S. We tried to group similar languages RU, PL, and UK to boost UK. However, we obtained just F_1=20 for UK, which is still significantly higher compared to the UK single language model, but lower compared to UK-EN and UK-EN_S.

One exception is EN-mid, which works significantly better in the single model (81 vs. 77); for EN-small, the difference is less pronounced (58 vs. 54-57). Another exception is KO, for which the single KO(B) model obtains 51 vs. multilingual 43-47. However, if we compare the multilingual KO, F_1=42, then the multilanguage models increase the results. Finally, the model that groups IT, FR, PT, and ES achieves similar results to ALL-EN_S.

Relations Figure 4 (right) shows F_1 for relations. The results differ widely between the languages and the relations. For instance, for EN some relations achieve F_1=100 (e.g. has-lifespan, eats), while one relation gets just F_1=22 (from-country). However, the same relation achieves F_1=73 for PL and even 0 for SV. We conclude that the results depend on the number of training examples for each relation.

Entity$_1$ vs Entity$_2$ Figure 6 demonstrates a significant difference in results between the two entities in the sentence. We obtain far higher results for entity$_1$ (left) than for entity$_2$ (right). This is because entity$_1$ usually occurs at the beginning of the sentence, while the position of entity$_2$ is not so deterministic.

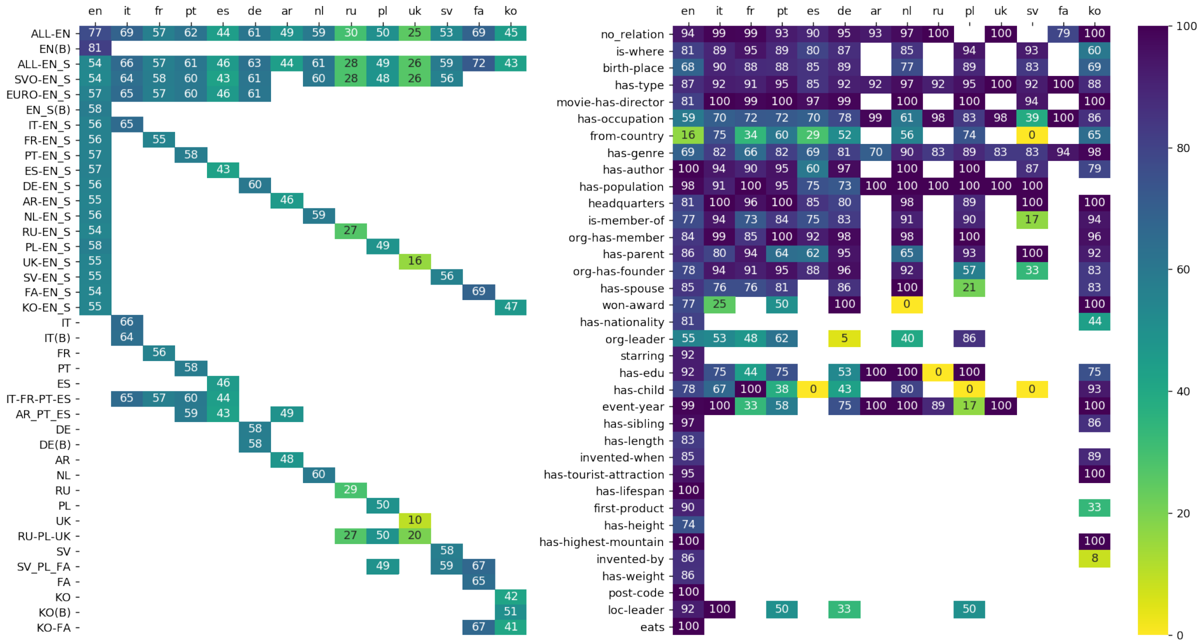

Figure 4: Left: F_1 (label and both entities correct) for all languages and models. Right: F_1 (label correct) for all languages and labels, averaged over the models available for each language.

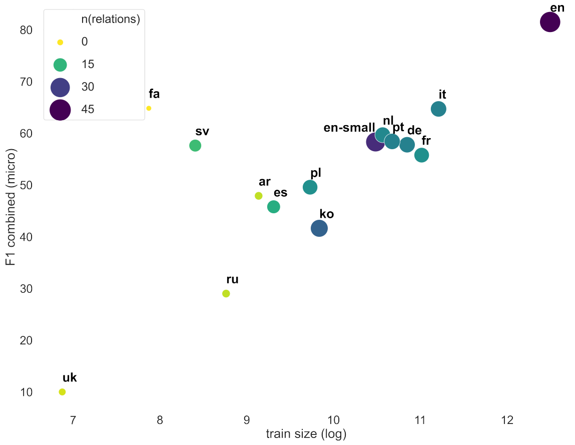

Figure 5: F_1 Combined for single language models versus the number of sentences in the train set. Dot size is proportional to the number of relations in the test set.

It is important to note though that typically sentences begin with their linguistic subject. As a result, for the sentence 'Bill has a child called Chelsea.', the relation has-child is far more plausible than has-parent, even though both are logically correct. For this reason, the fact that entity$_1$ has a more deterministic behavior is not a problem.

6 Conclusions

We have described our approach to solving the task of multilingual Joint Entity and Relation Extraction, by training our novel, HERBERTa model on SMiLER – our large, comprehensive dataset. The model combines two independent BERT models: one for Sequence Classification, and the other for Entity Tagging. Regarding F1 measure our model is close to SpERT (the current state-of-the-art for CONLL04), the difference being 0.2 percentage point. What is more, the SMiLER dataset appears to be currently the largest (1.1 M annotated sentences) dataset for this task, as well as the most comprehensive one (14 languages, 36 relation types).

We observe that our multilingual models achieve higher or similar results, compared to the models trained for each language separately. Languages with less data can benefit the most from such multilingual models. The victim here is English, as it seems to be a giver of F1 to other languages, especially when the number of sentences for this language is significantly higher than for other languages. On the one hand, due to the large amount of data for English, a model is well trained for patterns existing in this language and thus other languages with less data can benefit from it (because of some similarities between languages). Therefore, we observe increased results for less resourced languages. On the other hand, a model tries to accommodate the nuances of the less resourced languages. Their features are noticeable for the model, which reduces the dominant role of English.

As we can see from the results, each non-English language follows a slightly different (error) path.

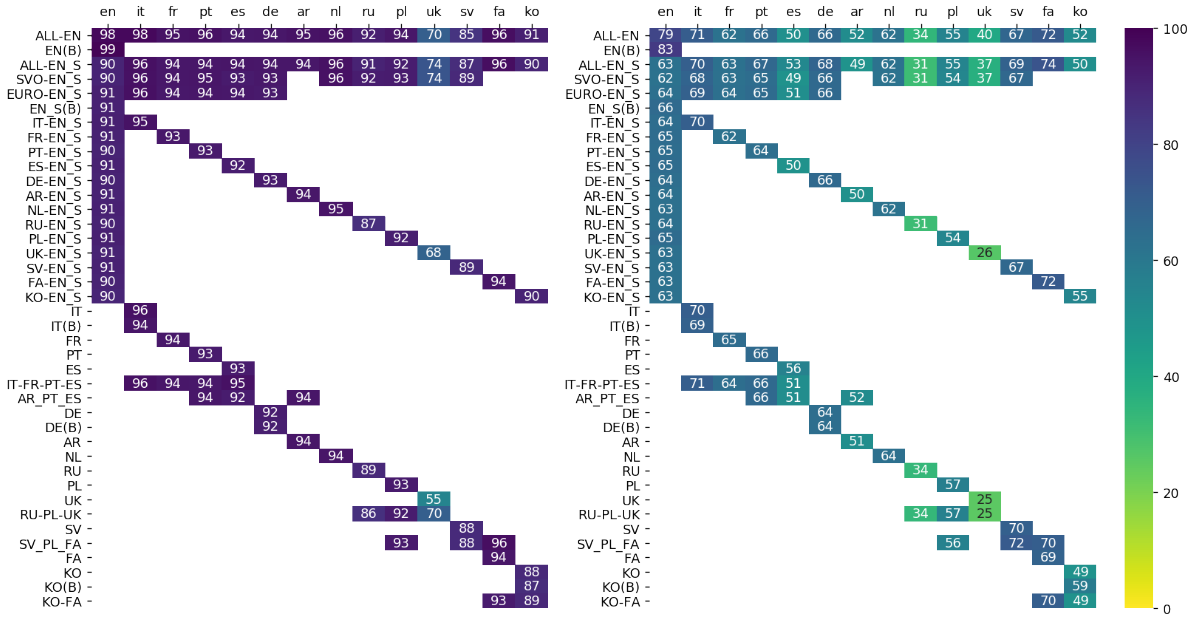

Figure 6: F_1 (left: entity$_1$ correct, right: entity$_2$ correct) for all languages and models.

This does not seem to be a general rule that can be applied. We would like to point out though that this is exactly the reason why we have created this dataset in the first place. We wanted to observe the performance of models on different languages on the relation extraction task and now, thanks to our dataset, this is possible. The fact that there is no simple explanation for the difference in model performance shows that deeper analysis for each language is necessary.

Another observation is that we obtain significantly higher F_1 for entity$_1$ than for entity$_2$, which suggests that entity$_1$ is simpler. This seems to be true, because entity$_1$ typically occurs at the beginning of the sentence, while entity$_2$ does not have any consistent location.

In the future, we plan to train our model on the EN-full dataset and to predict multiple plausible entity pairs for the same sentence. We would also like to extend the dataset to include entity types and sentences containing multiple relations.

Another promising direction is data augmentation by "cross-matching" entities and relations in the dataset with sentences in the dataset. This cross-match could search for two cases.

1. Sentences that contain multiple relations between the same 2 entities.

2. Sentences that contain more than 2 entities (e.g. 3), with different relations between them.

In both cases, the sentence could be added to the dataset multiple times.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *EMNLP/IJCNLP*.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 1247–1250.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 601–610.

Markus Eberts and Adrian Ulges. 2019. Span-Based Joint Entity and Relation Extraction with Transformer Pre-training. In *24th European Conference on Artificial Intelligence*.

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.

Kiril Gashteovski, Rainer Gemulla, and Luciano del Corro. 2017. MinIE: Minimizing Facts in Open Information Extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2630–2640.

John Giorgi, Xindi Wang, Nicola Sahar, Won Young Shin, Gary D. Bader, and Bo Wang. 2019. End-to-End Named Entity Recognition and Relation Extraction Using Pre-trained Language Models. arXiv:1912.13415.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a Benchmark Corpus to Support the Automatic Extraction of Drug-Related Adverse Effects from Medical Case Reports. *Journal of Biomedical Informatics*, 45(5):885 – 892.

Cheng Li and Ye Tian. 2020. Downstream Model Design of Pre-trained Language Model for Relation Extraction Task. arXiv:2004.03786.

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.

Mike Mintz, Steven Bills, Rion Snow, and Daniel 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, Suntec, Singapore. Association for Computational Linguistics.

T. Mitchell, W. Cohen, and E. Hruschka. 2018. Never-Ending Learning. *Commun. ACM*, 61(5):103–115.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT? In *ACL*.

Paweł Przybysz, Marcin Chochowski, Rico Sennrich, Barry Haddow, and Alexandra Birch-Mayne. 2017. The Samsung and University of Edinburgh's submission to IWSLT17. In *Proceedings of the* 14^{th} *International Workshop on Spoken Language Translation*, pages 23–28.

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.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the Blanks: Distributional Similarity for Relation Learning. In *Proceedings of the* 57^{th} *Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In 31^{st} *Conference on Neural Information Processing Systems (NIPS 2017)*.

Joanna Wetesko, Marcin Chochowski, Paweł Przybysz, Philip Williams, Roman Grundkiewicz, Rico Sennrich, Barry Haddow, Antonio Valerio Miceli Barone, and Alexandra Birch. 2019. Samsung and University of Edinburgh's System for the IWSLT 2019. In 16^{th} *International Workshop on Spoken Language Translation 2019*.

Philip Williams, Marcin Chochowski, Paweł Przybysz, Rico Sennrich, Barry Haddow, and Alexandra Birch. 2018. Samsung and University of Edinburgh's System for the IWSLT 2018 Low Resource MT Task. In *Proceedings of the* 15^{th} *International Workshop on Spoken Language Translation*, pages 118–123.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Zhu. 2012. Probase: A Probabilistic Taxonomy for Text Understanding. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*.