Applying Natural Annotation and Curriculum Learning to Named Entity Recognition for Under-Resourced Languages

Valeriy Lobov
MIPT
lobov.va@phystech.edu

Alexandra Ivoylova
RSUH
a.m.ivoylova@gmail.com

Serge Sharoff
School of Languages
University of Leeds, UK
s.sharoff@leeds.ac.uk

Abstract

Current practices in building new NLP models for low-resourced languages rely either on Machine Translation of training sets from better resourced languages or on cross-lingual transfer from them. Still we can see a considerable performance gap between the models originally trained within better resourced languages and the models transferred from them. In this study we test the possibility of (1) using natural annotation to build synthetic training sets from resources not initially designed for the target downstream task and (2) employing curriculum learning methods to select the most suitable examples from synthetic training sets. We test this hypothesis across seven Slavic languages and across three curriculum learning strategies on Named Entity Recognition as the downstream task. We also test the possibility of fine-tuning the synthetic resources to reflect linguistic properties, such as the grammatical case and gender, both of which are important for the Slavic languages. We demonstrate the possibility to achieve the mean F1 score of 0.78 across the three basic entities types for Belarusian starting from zero resources in comparison to the baseline of 0.63 using the zero-shot transfer from English. For comparison, the English model trained on the original set achieves the mean F1-score of 0.75. The experimental results are available from https://github.com/ValeraLobov/SlavNER

1 Introduction

The use of pre-trained language models (PLMs), such as BERT (Devlin et al., 2018), has significantly improved accuracy of many NLP tasks, such as POS tagging and Named Entity Recognition (Tenney et al., 2019). It is also possible to achieve good quality transfer of the classifiers across the languages using multilingual PLMs (Conneau et al., 2020; Hu et al., 2020). However, lesser resourced languages still present a considerable problem. First, the amount of raw text data available for their pretraining is on the order of magnitudes smaller than what is available for bigger languages, such as English. Second, zero- or few-shot multilingual transfer comes with the price of a performance gap, when a model tested on lesser resourced recipient languages is less accurate than the original donor language model. This comes partly because of the linguistic differences between the donor and recipient languages and partly because of the lower quality of pretrained embeddings obtained on smaller corpora for lesser resourced recipient languages (Vulić et al., 2020).

We propose a method to improve the accuracy of models for lesser resourced languages by what we call “natural" annotation, i.e., when some desired linguistic properties are derived from annotations arising as a by-product of a natural activity which is not directly related to the task of the model. Parallel corpora provide an example of natural annotation arising from human translations which are not produced by the their translators for the purpose of Machine Translation or Word Sense Disambiguation. Similarly, in our study we use natural annotation from Wikipedia categories, which provide a sufficient number of Named Entity (NE) examples even for lesser resourced languages. This allows production of synthetic corpora for such languages as Belarusian, which has no native NER resources. As an example, the Belarusian Wikipedia contains more than 130 thousand entries with more than four thousand entries about people (as of April 2022), sufficient for creating synthetic training sets.

Instead of the initial problem with the availability of training data, the use of synthetic corpora leads to the problem with having potentially millions of noisy annotated sentences per language. Therefore, we need to estimate their usefulness for training a model in the recipient languages. In this study we experiment with applying Curriculum

Proceedings of the 29th International Conference on Computational Linguistics, pages 4468–4480
October 12–17, 2022.

Learning methods (Bengio et al., 2009; Zhu et al., 2021) to synthetic corpora with natural annotation.

The contributions of this study are as follows:

- how to build a corpus with natural annotation from the available resources in Wikipedias;
- how to improve it by adjusting its linguistic properties;
- how to choose the curriculum learning strategy for corpora of this kind.

2 Methodology

2.1 Synthetic dataset

In this study we concentrate on building synthetic corpora for seven Slavic languages with three NE categories – PERsons, LOCations and ORGanizations. The main idea behind our experiment is:

1. to select annotated sentences in a better resourced donor language (we use the English WikiNER (Pan et al., 2017) for all experiments),
2. to produce synthetic corpora by Machine Translation of relevant annotated sentences in the recipient languages (we use Google Translate), and
3. to replace their annotations with the relevant Wikipedia entries from the recipient languages.

At the same time, Machine Translation is especially unreliable for NEs. For example, Google Translate renders ***Chain*** *was declared the winner as a result* into Slovene as ***Veriga*** *je bil zaradi tega razglašen za zmagovalca* with a literal translation of *Chain* as 'a series of connected links or things'.

Therefore, in Step 1 we select annotated WikiNER sentences with a single known NE (PER in this sentence), replace it with a pronoun placeholder into ***He*** *succeeded in purifying penicillin* and replace the placeholder to a range of known NEs from Wikipedia, for example, for this context ***Einstein/Romanova/Arhit*** *je bil zaradi tega razglašen za zmagovalca*, see Table 2. This way, each sentence in our synthetic datasets contains at least one NE (and no more than two of them for PER). This makes it quite different from natural datasets: the entity distribution is much more uniform in synthetic data.

The known NEs are determined as those with matching categories in the respective Wikipedias, see Table 1. For example, the English PER NEs are from categories like 1791 BIRTHS → *James Buchanan* → Джэймс Б'юкенен (in Belarusian).

We produced three versions of the synthetic datasets:

S1 Replacements of the parallel set of named entities;

S2 Replacements of the parallel set of named entities, while taking into account the grammatical case, number and gender;

S3 Replacements of a maximum available number of entities from the respective categories

For S1 and S2 we ensured that the NEs are available in all languages, so that the synthetic datasets can be completely parallel. We also used rules to normalize the names of the entries, e.g., Ломоносов, Михаил Васильевич ('Lomonosov, Mikhail Vasilyevich' in Russian) was normalized to Михаил Ломоносов 'Mikhail Lomonosov', which is the form used in texts.

S1 does not take into account the grammatical properties, such as the case, gender and number. Besides, it contains quite a lot of duples or quasi-duples (e.g., 'A.Lukashenko' and 'Alexander Lukashenko') which were deleted during S2 creation: this is the reason why PER number decreased for S2. Thus, S2 is generated by constraining the respective contexts for the respective categories. For example, *Romanova* is a female name, while *je bil razglašen* is a form requiring the male name, so in S2 we only generate synthetic sentences respecting these constraints:

en **Albert Einstein** was declared the winner as a result.
sl.m.sg → **Albert Einstein** je bil zaradi tega razglašen za zmagovalca
en **Anastasia Romanova** warned that the syndicalists aims were in perpetuating syndicalism itself.
sl.f.sg → **Anastazija Romanova** opozorila je, da so sindikalistični cilji ohranjanje samega sindikalizma.

The number and gender for the ORG type was detected from the syntactic properties of the head of each ORG name using udpipe (Straka et al., 2016).

Finally, S3 drops the constraint of having parallel NEs by using all of the NEs detected for a given language through the categories in Table 1 (the contexts remain parallel, though). The number of PER NEs becomes unreasonable, so we constrained it to 20,000 PER NEs per language.

Table 1: Labels of NER-related categories in Wikipedias.

Language	Wiki entries	PER Categories	LOC Categories	ORG Categories
en	6,489,550	Born	Cities|Countries	Organizations
be	217,410	Нарадзіліся	Краіны|Гарады	Арганізацыі
bg	281,108	Родени	Градове|Държави	Организации
cs	502,428	Narozen	Měst|Země	Organizace
pl	1,519,696	Urodzeni	Miasta|Kraje	Organizacje
ru	1,814,092	Родились	Города|Страны	Организации
sl	176,025	Rojeni	Mesta|Države	Organizacije
uk	1,151,062	Народились	Міста|Країни	Організації
#Parallel		4,492	1,709	239
#Total		140,000	22,709	34,329

Table 2: Synthetic dataset contents.

	PER	LOC	ORG
S1			
Entities	7,889	1,709	239
Contexts	7,694	1,535	348
S2			
Entities	4,492	3,178	239
Contexts	7,439	1,535	348
S3			
Entities	20,000	6,178	9,828
Contexts	7,439	1,535	1,087

We compare training on our synthetic corpora against two commonly used baselines with random training dataset ordering:

B1 Zero-shot transfer of a multilingual PLM with training on the original English dataset;

B2 Training on the target language dataset produced by Machine Translation of the original English dataset (The original entities were marked in the text by special symbols like '|0|' so that the labels would not be lost during MT).

Each training dataset consists of 12000 randomly generated examples from the available set of entities and contexts for particular language. The test dataset was produced by manual cleaning of approximately 10,000 tokens taken from WikiNER for each of the eight languages including English. The NE counts for this dataset are listed in Table 3.

2.2 NER setup

To test the contribution of our synthetic corpora with curriculum learning mechanisms, we rely on a competitive NER approach, which is based on XLM-R (Conneau et al., 2019). In the starting stage, we compared the multilingual BERT (Devlin et al., 2018) and XLM-R on our datasets. Subsequently, the XML-R model was used for all predictions as it showed slightly better results in all of our experiments (as also shown in other downstream applications). In addition, XLM-R offered better zero-shot transfer results, which is relevant for our B1 baseline.

2.3 Curriculum learning methods

The main purpose of using curriculum learning (CL) in the proposed method is to ensure better ordering of potentially noisy examples. Since the training dataset may contain samples from incompatible semantic categories within the same NE type (e.g., Greek philosophers doing research on penicillin), the CL model must reduce the negative effect of noisy examples and thereby make learning process more stable.

In total, we implemented three models to determine the order of significance for the examples:

C1 order data by sample size, i.e. sentence length

C2 order data by average confidence in predicted named entities

C3 order data by perplexity value

These models cover our hypotheses regarding aspects of data complexity for the named entity recognition task. C1 reflects the amount of information that the model is gradually learning. Long sentences carry significantly more information and therefore may be more difficult to mark up entities in them. Another complexity measure for NER

Table 3: Counts of NER-related categories in the test dataset.

Entities	en	be	bg	cs	pl	ru	sl	uk
PER	340	479	321	377	414	373	360	298
LOC	336	821	974	530	716	552	690	948
ORG	358	125	87	195	195	292	104	138

task samples is the average confidence of the CL model in predicted named entities (model C2). Inspired by the article (Zhu et al., 2021) we also calculate the probability that the entity word has a label from the gold-label markup. By sorting the training dataset in descending order of such probability, we thereby place incorrect examples at the end and give priority to high-confidence data.

The model needs to be based on a different architecture, for which we used Bi-LSTM-CRF. It calculates the optimal tag probability as:

$$p^*(x) = \max_{y' \in \boldsymbol{y}} \left[CRF_F(x, y') + CRF_B(x, y') \right] \tag{1}$$

where $\boldsymbol{y}$ means the original set of labels, x - input word, CRF_F and CRF_B represent the calculated probabilities from the CRF model for forward and backward pass of the Bi-LSTM network. Then the average probability among the entities words is computed as:

$$\text{C2}_{score}(S) = \frac{1}{|\boldsymbol{w}|} \sum_{i=1}^{|\boldsymbol{w}|} (p^*(w_i)) \tag{2}$$

where $\boldsymbol{w}, \boldsymbol{w} \subset S$ mean subset of words from the sentence S which are named entities.

Finally C3 is perplexity of CL model for a given sentence. Perplexity is a measurement of how well a probability model p predicts a sample. This metric is calculated below:

$$PP(p) = 2^{-\sum_x p(x) \log_2 p(x)} \tag{3}$$

So, we are curious if this metric is suitable for sorting training samples and improving the learning process in our task. As follows from the definition, the training dataset needs to be sorted in the ascending order of this metric.

Several studies in curriculum learning (Wang et al., 2019a,b; Castells et al., 2020) suggest discarding the top-N% of the most complex samples from the dataset. Therefore, in our study, we conducted several experiments to understand whether the model should use all samples or a certain percentage during training.

3 Results

Because of space constraints, here we compare the performance of CL models and of synthetic datasets for two languages – Polish and Belarusian. Belarusian is the most under-resourced language out of those tested in our study, in terms of the amount of (1) pre-trained data, (2) Wikipedia entries and (3) available training resources. Polish is a better resourced language from a different Slavic branch (West vs East for Belarusian), and also it is written in a mix of Latin and Polish-specific characters, which creates non-trivial tokenization problems, as its tokenizer is influenced by major European languages. The results for Polish reported below are similar to those for Belarusian, which emphasizes the universality of the methods for Slavic languages presented in this paper. The applicability of these methods to other language groups is beyond the scope of this article, but nevertheless this is an interesting topic for further research.

Table 4 reports summary F1 metrics on WikiNER evaluation dataset for all models and training datasets for Belarusian. Table 5 - a similar table for Polish. The main overall measure is the mean F1 scores for the three main NE categories (omitting the Other category, which does not indicate the source of errors). The best mean score for Belarusian is 0.78, for Polish 0.79. For comparison, the same English model trained on the same set achieves the mean F1-score of 0.75, while the best zero-shot transfer achieves 0.63 for Belarusian and 0.68 for Polish. The full set of results for 7 Slavic languages is presented in Figure 6 in the Appendix.

3.1 Synthetic datasets

There is considerable variation between the NE types. PER is easy to detect by using any training set, this is followed by LOC, while ORG is more difficult. There is a very considerable improvement from using more linguistic information when moving from S1 to S2. S1 is mostly not better than either of the two baselines, i.e. B1 – zero-shot transfer or B2 – direct MT of the English training

Table 4: Belarusian: F1 scores for all entities, models and datasets. The scores in bold mean the best results for a CL model within a dataset. The underlined scores are the best scores for each NE type across all models and datasets. Last row, *mean score*, represents mean F1 score for three entity types.

Entity	Model	B1	B2	S1	S2	S3
PER	C1	0.89	0.90	0.85	0.91	**0.92**
	C2	0.88	0.91	**0.89**	**0.93**	0.91
	C3	0.88	0.89	0.88	0.92	0.91
LOC	C1	0.60	0.64	0.59	0.78	**0.76**
	C2	0.55	0.58	0.51	**0.81**	0.60
	C3	0.64	0.66	**0.70**	0.79	0.57
ORG	C1	0.40	0.43	0.30	0.56	**0.62**
	C2	0.42	0.44	0.33	0.61	0.58
	C3	0.38	0.45	**0.37**	**0.62**	0.57
Mean score	C1	0.63	0.66	0.58	0.75	**0.77**
	C2	0.62	0.64	0.57	**0.78**	0.70
	C3	0.63	0.67	**0.65**	0.77	0.68

Table 5: Polish: F1 scores for all entities, models and datasets. The scores in bold mean the best results for a CL model within a dataset. The underlined scores are the best scores for each NE type across all models and datasets. Last row, *mean score*, represents mean F1 score for three entity types.

Entity	Model	B1	B2	S1	S2	S3
PER	C1	0.88	0.89	0.86	0.92	0.93
	C2	0.88	0.90	0.82	**0.93**	**0.94**
	C3	0.87	0.89	**0.87**	0.92	0.92
LOC	C1	0.64	0.67	0.54	0.76	**0.81**
	C2	0.66	0.68	0.55	0.75	0.68
	C3	0.68	0.70	**0.64**	**0.80**	0.63
ORG	C1	0.54	0.56	0.50	0.61	0.57
	C2	0.51	0.55	0.48	0.60	**0.59**
	C3	0.43	0.48	**0.55**	**0.66**	0.54
Mean score	C1	0.68	0.70	0.63	0.76	**0.77**
	C2	0.68	0.71	0.62	0.76	0.74
	C3	0.66	0.69	**0.68**	**0.79**	0.70

set. Also, surprisingly, the very simple B2 setup is often better than the popular zero-shot transfer of B1.

S2 is the best dataset overall, so the expected improvements by using larger training sets in S3 did not materialise. The most likely reason is that most frequent location entities in the test dataset are country names, which are well covered by S2, as most of them are translated across the Wikipedias; thus, adding even a much bigger set of examples with rare place names could not improve the scores (though we suppose that S3 could show better results on a test dataset with rare place names). The effect of gender disambiguation used in S2 and S3 (opposite to S1) clearly show the benefits for PER recognition, even though this is an easier task for all of the models and datasets. The most difficult category is ORG. Some organization names are rendered without translation, i.e., *General Motors*, so they are easier to recognize. ,On the other hand, many ORG names are linguistically diverse with a complicated structure, for example, *Międzynarodowe Centrum Badań nad Ochroną i Konserwacją Dziedzictwa Kulturowego* 'International Centre for the Study of the Preservation and Restoration of Cultural Property', with the problem in their detection persisting across any available NER model.

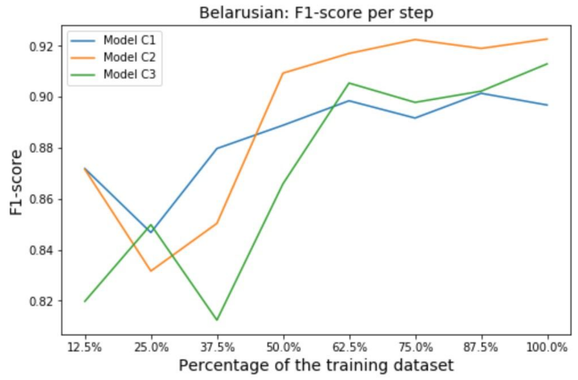

Figure 1: F1 score on evaluation dataset after each training step on belarusian S2 dataset.

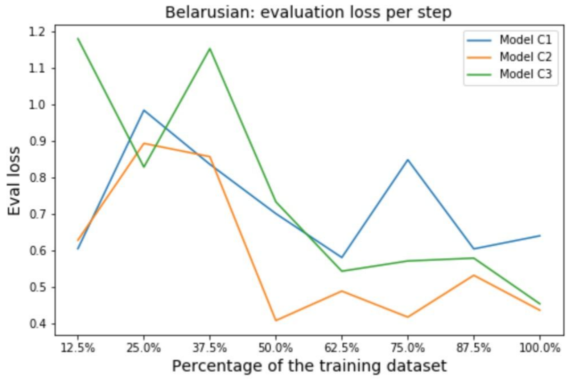

Figure 2: Evaluation loss after each training step on Belarusian S2 dataset.

3.2 CL Models

In our experiments with models, we wanted to evaluate the impact of curriculum learning on the synthetic dataset. First of all, ordering of the training samples improves target metrics compared to baseline, basic XLM models with random sampling. For locations and organizations the difference is substantial: 0.15 and 0.17 of the F1-score points respectively. For persons, the difference is not so big, because initially there are enough entities and context examples for persons, so it is easier to recognize PER entities irrespectively of the order of training.

One of the key goals of using curriculum learning in this study is to reduce the impact of noisy synthetic examples. For example, using the Belarusian dataset S2 and model C2 (average probability of predicting gold-label entity tags), these examples are listed as the highest difficulty for predicting LOC:

1. Шпаер таксама ў спісе з двума. ‘Speyer is also in the list with the two’
2. Рэвалюцыя 1897-1898 гадоў адкрыла дзверы для больш шырокіх ведаў, і пачалося шмат даследаванняў, пра якія гл. Валета ‘The 1987-1898 revolution opened the doors for wider knowledge, and many researches were started, about which see Valeta’

The C2 model reasonably placed these examples at the end of the training dataset. In the first example, the location Шпаер (‘Speyer’) is similar to the surname, so the model is not confident between deciding it is a location and a person. In the second example, the structure is not correct since the end of the sentence contains an artifact from Wikipedia "about which see [entity]". This sentence tells about the revolution, not about the location, so the context of the entity is incomplete. In the examples shown above, the model correctly identified the entities, but was not sure about them, so the curruculum learning method assigned them the higher scores.

Comparing the methods of curriculum learning, we wanted to determine which method makes learning more robust and efficient. For this experiment, we took the S2 dataset, as it provides the best performance among all other datasets. As for the experiment setup, we trained three models C1, C2, C3 on the Belarusian S2 dataset with following configuration: 1 epoch, Adam optimizer, learning rate equals 1e-5, weight decay equals 0.01. The choice of a single epoch and a low learning rate is primarily due to the fact that the NER model trained on synthetic datasets quickly overfits. For example, for Belarusian and Polish S2, the validation loss increases after 12.5% of the training dataset of the second epoch. On the charts 1 and 2 the key evaluation metrics of three CL models are shown. Since we found out that one epoch is enough for training, on the figures each training step is 12.5% of the training dataset. In particular, Figure 2 shows that the ranking of training examples based on the metric of the average probability of predicting named entities (model C2) demonstrates more stable training. Moreover, this model is the fastest gaining more than 90% of the F1 metric and has the minimum evaluation loss among all models.

Also, in the same experiment we found out that throwing out 10% of most complicated data leads to mean decrease in F1 measure by around 3% with all CL models. This means that the model should rather see the most complex examples in order to demonstrate the best quality.

3.3 Errors analysis

We manually analyzed and compared the predictions of models finetuned on S2 and S3 datasets for Polish and Belarusian. The quality for Belarusian is comparable to the other languages in the set. There were no problems found caused by tokenization, though sometimes the models tend to misjudge the boundaries of the entities, including spaces, brackets or commas.

Most errors for S2 are predictably caused by cases: so PER entities which are not in nominative case often do not get a correct label, or case endings may not be labelled (our model labels every BPE-token), e.g., Polish *Artura Rubinsteina* 'Arthur Rubinstein' loses its PER label for endings *-a* in both name and surname and the markup looks like *'Artur, PER'* and *'#a, O'*. The same goes for other entities, for example, Polish *Kanady* 'Canada' (genitive) may not get annotated in certain contexts at all, and ORG entity *Partią Republikańską* 'Republican Party' (instrumental) has an unlabelled case ending *ą*.

Some of other easily explainable errors are caused by entity type: seas, rivers, and mountains were not present in the synthetic datasets, so the models may only partially recognize entities like Belarusian Галілейскага мора 'Sea of Galilee'.

Another quite common, but not so obvious error concerns unlabelled country names: although the models must have seen them, country names are often lost. It is worth mentioning that country names are often unlabelled in real data, such as WikiNER and SlavicNER (Piskorski et al., 2021), as well.

When the models are finetuned on the S3 dataset, the error types generally remain the same, though their quantity is slightly smaller. We noticed that sometimes the context obviously provides a false label, i.e., in Belarusian sentence which contains пад націскам Рима пазбавиў 'under the pressure of Rome [he] deprived' Rome in genitive case gets the PER label, because it is followed by a verb common for PER entities.

4 Related work

4.1 NER and synthetic corpora

The idea of using synthetic data for augmentation was a natural consequence of the development of ML models which need a lot of annotated data in order to learn. Mentions of synthetic data in NLP can be found as early as in 2000s (i.e., (Talbot, 2003)); an obvious solution was to generate training data for ML models using rule-based tools, as it was done to improve machine translation in (Hu et al., 2007). Starting from late 2000s, the idea of using synthetic corpora grew more popular and was applied in various areas of research, for example, for generation of animations for sign languages (Schnepp et al., 2010), or for Implicit Discourse Relation Recognition (Lan et al., 2013). Nowadays, there are plenty of works using synthetic data for improvement, and generating synthetic datasets is considered a common technique. For example, (Kvapilíková et al., 2020) article used unsupervised machine translation to build a synthetic dataset and improve quality on parallel corpus mining task in low-resource languages, which is especially relevant for our research. Some recent works using this approach are, for example, (Li et al., 2021), (Hosseini et al., 2021) or (Whitfield, 2021), the latter introducing GPT-2 model for data generation. Also, method proposed in (Sellam et al., 2020) uses synthetic sentences for BERT pre-training, which contain a wide variety of lexical, syntactic, and semantic diversity. The key goal for synthetic datasets is to provide the maximum variation of text data in order to make the target model more robust.

NER typically involves one of three gold standards: MUC, CoNLL, or BBN, all created by costly manual annotation. One of the first datasets for NER in the CoNLL standard was created in 2003 (Sang and De Meulder, 2003) and covered two languages (English and German); the NER-related categories consisted of PER, LOC, ORG and MISC labels. There were also several datasets in CoNLL standard created around 2010, one of them being WikiGold (Balasuriya et al., 2009) – 40K tokens of Wikipedia articles, manual annotation; the other, Web (Ratinov and Roth, 2009), contained 8K tokens taken from the Web, and the third dataset used Twitter as a resource and contained approximately 34K tokens (Ritter et al., 2011). Later additions are represented by Broad Twitter Corpus (Derczynski et al., 2016) and WiNER (Ghaddar and Langlais, 2017), although the list of NER-annotated corpora isn't exhausted by those.

There are also attempts to create synthetic corpora with NER-annotation, mostly for low-resource languages, i.e., (Jónsson et al., 2021) for Icelandic. Cross-lingual transfer is a viable method as well and it was used for creation of a synthetic Chinese NER corpus in 2014, for example (Fu et al., 2014). There has also been a recent set of experiments on cross-lingual annotation, for example, for African languages (Adelani et al., 2021). There is also an approach to use a small number of examples via triggers (Lin et al., 2020).

4.2 Curriculum learning

Curriculum Learning (CL) is a learning technique where the order of training samples depends on their complexity for the target model. This paradigm resembles human learning: a gradual increase in the complexity of training examples makes learning process more qualitative. The original technique was proposed by (Bengio et al., 2009). But for supervised learning and NLP in particular, (Elman, 1993) article is one of the first where an idea similar to curriculum learning is applied. The author emphasized that the order of training data is important, where the "small" data comes first.

More recently, the curriculum learning approach was used quite extensively in NLP. Different measures of complexity have been proposed, depending on the task. For instance, (Tay et al., 2019) article addresses the problem of reading comprehension of long texts. Curriculum learning based on the answerability and understandability of texts effectively improves training process. In addition, (Platanios et al., 2019) apply curriculum learning to neutral machine translation to learn better and converge faster. The framework presented in the paper shows certain samples to the model at certain times according to their complexity and model competence at that moment.

NER task requires a comprehensive metric to use for curriculum learning. The study closest to our work (Zhu et al., 2021) uses several strategies to organize their training data. All of them use probability of entities from the gold label dataset calculated by the CL model. In one case, the average confidence of the model is calculated from all entities from the input sentence. In another, the averaged confidence of the model is considered only for named entities. This approach is relevant due to the fact that goal is to recognize named entities correctly, other tokens in the sentence are not so important. Furthermore, there are other techniques for filtering and ordering samples for the NER task, especially suitable for generated datasets. (Liu et al., 2021) filter data samples with specific entity based on complex criteria: for example, if this entity is too frequent in the training dataset.

4.3 Cross-lingual transfer

Cross-lingual transfer is relevant when there is sufficient data for training a model for one language and no such data for another language. A model which is trained on the data for one language and applied on the data of the other (zero-shot transfer), usually shows worse results than on donor language material; thus leading to a transfer gap, which can be measured as the difference between the performance of the same model on donor and recipient languages (Hu et al., 2020). In order to use cross-lingual transfer, language spaces must be aligned, and the models which provide higher quality vector spaces perform better. This idea has become popular since 2014, when embedding methods produced high quality spaces which are almost isomorphic across languages and which can be aligned by using small seed dictionaries (Mikolov et al., 2013).

Modern models, such as multilingual BERT or XLM-RoBERTa, are even more efficient at building cross-lingual vector spaces (Conneau et al., 2020). Some recent studies in this area showed that the transfer gap if the donor language is English normally is not bigger than 0.25 (0.14 on average) for a set of recipient languages (Hu et al., 2020; Ruder et al., 2021). For example, the Natural Language Inference (NLI) task for Slavic languages has a transfer gap of 0.07. As for the NER task with the use of cross-lingual transfer, there was a recent analysis of zero-shot transfer between English and Korean (Kim et al., 2021). More specifically on the topic of this paper, cross-lingual transfer for NER on Slavic languages has been discussed in (Sharoff, 2020), and shared tasks such as SlavicNER (Piskorski et al., 2021) are specifically aimed at NER for Slavic languages.

5 Conclusions

In this work we demonstrated how to achieve prediction quality for lesser resourced languages without any performance gap introduced by zero-shot transfer or Machine Translation. A synthetic corpus of about 10,000 sentences produced from a combination of naturally annotated data and machine translation from a better resourced language can produce better results than training on the source dataset for the better resourced language. However, the key to this success is an accurate model of important linguistic phenomena (case, number and gender, as in our S2 and S3 datasets), as without this the synthetic corpus (our version S1) is worse than the zero-shot baseline (B1). The second major contribution is the importance of the curriculum learning strategy. Any strategy for choosing the examples helps, but the average probability (from a different model) and perplexity usually help more than simple ordering by the sentence length. Also discarding the most difficult items does not help, as the models improve when seeing more data.

The negative result of this study is that a bigger collection NEs (S3) did not improve over the smaller set (S2). More research is needed into understanding the reasons for this. Better NE selection can help in matching the test dataset, while this might cause problems in applying the models beyond the test dataset. The study is also limited to a specific set of languages as well as to a single downstream task. In our future research we want to explore Wikipedias and similar resources with natural annotation for building synthetic training sets for more languages and for other downstream tasks, such as NE linking or ontology building.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Dominic Balasuriya, Nicky Ringland, Joel Nothman, Tara Murphy, and James R Curran. 2009. Named entity recognition in wikipedia. In *Proceedings of the 2009 workshop on the people's web meets NLP: Collaboratively constructed semantic resources (People's Web)*, pages 10–18.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

Thibault Castells, Philippe Weinzaepfel, and Jérôme Revaud. 2020. Superloss: A generic loss for robust curriculum learning. In *NeurIPS*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. Broad twitter corpus: A diverse named entity recognition resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179.

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*.

Jeffrey Elman. 1993. Learning and development in neural networks: the importance of starting small. *Cognition*, 48:71–99.

Ruiji Fu, Bing Qin, and Ting Liu. 2014. Generating chinese named entity data from parallel corpora. *Frontiers of Computer Science*, 8(4):629–641.

Abbas Ghaddar and Philippe Langlais. 2017. Winer: A wikipedia annotated corpus for named entity recognition. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 413–422.

Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordoni, and Aaron Courville. 2021. Understanding by understanding not: Modeling negation in language models. *arXiv preprint arXiv:2105.03519*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *arXiv preprint arXiv:2003.11080*.

Xiaoguang Hu, Haifeng Wang, and Hua Wu. 2007. Using rbmt systems to produce bilingual corpus for smt. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 287–295.

Haukur Páll Jónsson, Vésteinn Snæbjarnarson, Haukur Barri Símonarson, and Vilhjálmur Thorsteinsson. 2021. En-is synthetic parallel named entity robustness corpus.

Jongin Kim, Nayoung Choi, Seunghyun Lim, Jungwhan Kim, Soojin Chung, Hyunsoo Woo, Min Song, and Jinho D. Choi. 2021. Analysis of zero-shot crosslingual learning between English and Korean for named entity recognition. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 224–237, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ivana Kvapilíková, Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Ondřej Bojar. 2020. Unsupervised multilingual sentence embeddings for parallel corpus mining. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 255–262, Online. Association for Computational Linguistics.

Man Lan, Yu Xu, and Zheng-Yu Niu. 2013. Leveraging synthetic discourse data via multi-task learning for implicit discourse relation recognition. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 476–485.

Jianfu Li, Yujia Zhou, Xiaoqian Jiang, Karthik Natarajan, Serguei Vs Pakhomov, Hongfang Liu, and Hua Xu. 2021. Are synthetic clinical notes useful for real natural language processing tasks: A case study on clinical entity recognition. *Journal of the American Medical Informatics Association*, 28(10):2193–2201.

Bill Yuchen Lin, Dong-Ho Lee, Ming Shen, Ryan Moreno, Xiao Huang, Prashant Shiralkar, and Xiang Ren. 2020. TriggerNER: Learning with entity triggers as explanations for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8503–8511, Online. Association for Computational Linguistics.

Zihan Liu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. 2021. NER-BERT: A pre-trained model for low-resource entity tagging. *arXiv preprint arXiv:2112.00405*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958.

Jakub Piskorski, Bogdan Babych, Zara Kancheva, Olga Kanishcheva, Maria Lebedeva, Michał Marcińczuk, Preslav Nakov, Petya Osenova, Lidia Pivovarova, Senja Pollak, Pavel Přibáň, Ivaylo Radev, Marko Robnik-Sikonja, Vasyl Starko, Josef Steinberger, and Roman Yangarber. 2021. Slav-NER: the 3rd cross-lingual challenge on recognition, normalization, classification, and linking of named entities across Slavic languages. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 122–133, Kiyv, Ukraine. Association for Computational Linguistics.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom M. Mitchell. 2019. Competence-based curriculum learning for neural machine translation.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 1524–1534.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, et al. 2021. Xtreme-r: Towards more challenging and nuanced multilingual evaluation. *arXiv preprint arXiv:2104.07412*.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Jerry C Schnepp, Rosalee Wolfe, and John McDonald. 2010. Synthetic corpora: a synergy of linguistics and computer animation. *Visual Communication and Technology Education*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Serge Sharoff. 2020. Finding next of kin: Cross-lingual embedding spaces for related languages. *Journal of Natural Language Engineering*, 26:163–182.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia.

David Talbot. 2003. Learning translations from comparable corpora.

Yi Tay, Shuohang Wang, Luu Anh Tuan, Jie Fu, Minh C. Phan, Xingdi Yuan, Jinfeng Rao, Siu Cheung Hui, and Aston Zhang. 2019. Simple and effective curriculum pointer-generator networks for reading comprehension over long narratives.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. Are all good word vector spaces isomorphic? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3178–3192, Online. Association for Computational Linguistics.

Wei Wang, Isaac Caswell, and Ciprian Chelba. 2019a. Dynamically composing domain-data selection with clean-data selection by “co-curricular learning” for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1282–1292, Florence, Italy. Association for Computational Linguistics.

Wei Wang, Ye Tian, Jiquan Ngiam, Yinfei Yang, Isaac Caswell, and Zarana Parekh. 2019b. Learning a multi-domain curriculum for neural machine translation.

Dewayne Whitfield. 2021. Using gpt-2 to create synthetic data to improve the prediction performance of nlp machine learning classification models. *arXiv preprint arXiv:2104.10658*.

Wenjing Zhu, Liu Jian, Xu Jinan, Chen Yufeng, and Zhang Yujie. 2021. Improving low-resource named entity recognition via label-aware data augmentation and curriculum denoising. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1131–1142, Huhhot, China. Chinese Information Processing Society of China.

A Appendix

A.1 Results for 7 Slavic languages

In this section we provide Figure 6 with comprehensive results of our experiments in seven Slavic languages.

Table 6: Mean F1 scores per 3 entities types for all languages, models and datasets. The scores in bold mean the best results for a CL model within a dataset. The underlined scores are the best scores for each NE type across all models and datasets.

Language	Model	B1	B2	S1	S2	S3
Belarussian	C1	0.63	0.66	0.58	0.75	**0.77**
	C2	0.62	0.64	0.57	**0.78**	0.70
	C3	0.63	0.67	**0.65**	0.77	0.68
Ukrainian	C1	0.63	0.67	0.61	0.71	0.69
	C2	0.64	0.68	0.67	**0.75**	0.70
	C3	0.65	0.67	**0.68**	0.73	**0.72**
Russian	C1	0.61	0.63	0.60	0.73	0.70
	C2	0.63	0.62	**0.63**	**0.77**	**0.73**
	C3	0.62	0.63	0.61	0.75	0.72
Slovenian	C1	0.67	0.69	0.60	0.70	0.69
	C2	0.68	0.71	0.64	**0.78**	**0.75**
	C3	0.68	0.70	**0.65**	0.73	0.71
Polish	C1	0.68	0.70	0.63	0.76	**0.77**
	C2	0.68	0.71	0.62	0.76	0.74
	C3	0.66	0.69	**0.68**	**0.79**	0.70
Bulgarian	C1	0.62	0.64	0.61	0.74	0.70
	C2	0.65	0.66	**0.64**	0.76	**0.74**
	C3	0.66	0.67	0.63	**0.78**	0.71
Czech	C1	0.64	0.67	0.65	0.72	0.71
	C2	0.65	0.68	**0.68**	**0.76**	**0.75**
	C3	0.65	0.69	0.66	0.74	0.74