Zero-Shot Cross-Lingual Transfer is a Hard Baseline to Beat in German Fine-Grained Entity Typing

Sabine Weber
School of Informatics
University of Edinburgh, UK
s.weber@sms.ed.ac.uk

Mark Steedman
School of Informatics
University of Edinburgh, UK
steedman@inf.ed.ac.uk

Abstract

The training of NLP models often requires large amounts of labelled training data, which makes it difficult to expand existing models to new languages. While zero-shot cross-lingual transfer relies on multilingual word embeddings to apply a model trained on one language to another, Yarowsky and Ngai (2001) propose the method of annotation projection to generate training data without manual annotation. This method was successfully used for the tasks of named entity recognition and coarse-grained entity typing, but we show that it is outperformed by zero-shot cross-lingual transfer when applied to the similar task of fine-grained entity typing. In our study of fine-grained entity typing with the FIGER type ontology for German, we show that annotation projection amplifies the English model's tendency to underpredict level 2 labels and is beaten by zero-shot cross-lingual transfer on three novel test sets.

1 Introduction

The task of fine-grained entity typing (FET) is to assign a semantic label to a span in a text. The task is distinct from coarse-grained entity typing as done by named entity recognition systems because these systems are restricted to a small set of labels like 'person', 'organization' and 'location' which are not helpful for tasks that require more precise information about the entities. For example, FET assigns the label '/location/city' to the named entity '*Berlin*' in the sentence *'From 1997 to 2000, it had a permanent exhibition in Berlin.'*

Fine-grained entity typing uses a high number of types in a multilevel hierarchy, which can be seen in the level 2 label '/location/city' (see Figure 1). In this work we use the FIGER type hierarchy which consists of two levels with 112 types in total (37 level 1, 75 level 2). FIGER types are derived from the knowledge graph Freebase (Bollacker et al., 2008). They are both interpretable by humans and useful in NLP applications such as relation extraction (Kuang et al., 2020).

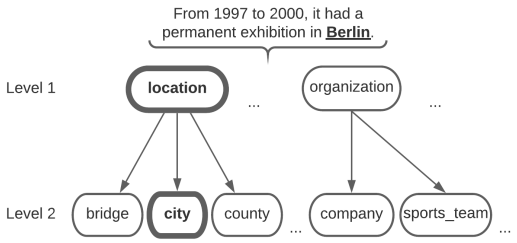

Figure 1: An example of fine-grained entity typing with the FIGER ontology. Correct types are highlighted.

There are systems for named entity recognition and coarse-grained entity typing in languages other than English (e.g. Stanza (Qi et al., 2020)), but systems for FET with FIGER types are only available in English, due to the lack of FIGER annotated data in other languages. Because manual annotation is time consuming and expensive, various methods have been proposed to expand NLP models to other languages without additional manual annotation. The method of **annotation projection** (Yarowsky and Ngai, 2001) uses parallel text to automatically create annotated corpora. Annotations from the resource-rich language are transferred to the resource-poor language using word alignment between translated sentences.

Annotation projection has been used successfully for the task of coarse-grained named entity typing in conjunction with named entity recognition (Agerri et al., 2018; Li et al., 2021; Ni et al., 2017). We follow these examples by using a parallel English-German corpus, automatic named entity recognition and a state of the art English FET model (Chen et al., 2020) to assign FIGER type labels on the English side for transfer. We then project the labels onto the German half of the corpus. The output of this process is a German corpus annotated with FIGER types, which we use to train a German FET model.

Another approach to the same problem is **zero-shot cross-lingual transfer**, in which a model built on multilingual word embeddings and trained on high-resource language data is applied to test data in a different language. Because the English FET model used in this work (Chen et al., 2020) relies on contextualised multilingual word embeddings (XLM-RoBERTa) (Conneau et al., 2019) it is possible to train it on English data and to test it on German.

We compare the two approaches and show that the annotation projection approach amplifies the model's tendency to underpredict level 2 types, which lowers model performance. We also introduce three new test sets for German FET[1] on which zero-shot cross-lingual transfer performs better than models trained with German or a mix of German and English data.

2 Related Work

To the best of our knowledge there is no work that compares annotation projection directly against zero-shot cross-lingual transfer. While annotation projection has been used in a variety of tasks, there has not been a study of a case where this approach fails. Authors admit that the quality of the annotating system plays role (e.g. Ehrmann et al. (2011); Ni et al. (2017)), but they don't specify model properties that are necessary for the approach to work, instead focusing on ways to mitigate noise.

Pires et al. (2019); Hsu et al. (2019) and Artetxe and Schwenk (2019) show the strengths of zero-shot cross-lingual transfer on a variety of different NLP tasks, but they do not address fine-grained entity typing. Zhao et al. (2020) conclude that zero-shot performance can be improved by choosing a small amount of high quality training data from the target language. We test their approach for the FET scenario, but arrive at unclear results.

3 Method

In this work we use the **hierarchical typing model** of Chen et al. (2020) trained on English gold data for the zero-shot approach and also to annotate the English side of the parallel text for annotation projection. We train the model with English silver data to show the amount of noise added by automatic annotation and finally we train it with German data which was produced by annotation projection.

[1]all test data sets and relevant code are available under https://github.com/webersab/german_FET

In the hierarchical typing model the entity and its context are encoded using multilingual XLM-RoBERTa (Conneau et al., 2019). For each type in the FIGER ontology the model learns a type embedding. It passes the concatenated entity and context vector through a 2-layer feed-forward network that maps into the same space as the type embedding. The score is the inner product between the transformed entity and context vector and the type embedding. For further model details refer to Chen et al. (2020).

4 Experimental Setup

Training Data To contrast the zero-shot cross-lingual transfer approach with models trained on automatically annotated and projected data we use three sources of training data. We use the 2M sentences English FIGER corpus as described by (Ling and Weld, 2012) as a source of **English human annotated data**, which we will refer to as *EN gold*. The data set consists of English Wikipedia articles and we use it to train the *zero-shot gold* model.

Second, we use **English machine annotated data** (*EN automatic*). Annotating English data using a model is the first step of the annotation projection. We use this data to train the *zero-shot automatic* model to examine the amount of noise added by automatic annotation. We generate *EN automatic* from English sentences from the WikiMatrix corpus Schwenk et al. (2019) [2], using the hierarchical typing model trained on 2 M sentences *EN gold*. Lastly, we use **German annotation projected data** (*DE projected*) that was generated by projecting the labels from the *EN automatic* onto German. For the details of our annotation projection pipeline please refer to appendix A. We use this data to train the *annotation projected* model.

We portion each training corpus into slices of 100, 200, 300 and 400 K sentences to compare the influence of data size. For *DE projected* only 300 K sentences are available, because only part of the parallel sentences in the WikiMatrix corpus are of high enough quality for annotation projection. For details of the selection process refer to appendix A.

An important point for our experiments is the **label distribution** in the training corpora (see table 1). The hierarchical typing model has the tendency

[2]While higher quality parallel data sets are available, this is the only one in the domain of Wikipedia articles. Preliminary experiments have shown that domain is an important factor for the quality of automatic FET, which is why we chose domain consistency over data quality for our experiments.

	EN gold	EN aut.	DE proj.
Lvl1 labels	60%	78%	77%
Lvl2 labels	40%	22%	23%
Lvl1 labels	155679	148571	150166
Lvl2 labels	104807	42008	43604

Table 1: Percentage and total numbers of level 1 and level 2 labels in 100 K sentences of the training corpora. Data created by annotation with the hierarchical-typing model contains fewer level 2 labels than human annotated gold data.

to underpredict the finer-grained level 2 labels (e.g. /person/actor, as opposed to level 1 label /person), which leads to a different distribution of labels in *EN gold* and the other corpora. Compared to approximately 100 K level 2 labels per 100 K sentences in the gold data, we only see about 50 K level 2 labels in the silver data. This tendency does not depend on the different input data: If we use a model trained on 100 K *EN gold* to predict labels on an unseen portion of *EN gold*, only 25% of the resulting annotations are level 2 labels.

Metrics Following previous FET literature we evaluate the results of our model using strict accuracy (Acc). The strict accuracy is the ratio of instances where the predicted type set is exactly the same as the gold type set. We also evaluate per hierarchy level.

Test sets We compare performance using the following test corpora: **1)** a German machine translation of the test split of the English FIGER corpus (Ling and Weld, 2012) using DeepL, which was manually corrected to eliminate translation and labelling errors (*DE-FIGER*); **2)** 500 manually annotated German sentences from the WikiMatrix corpus (*DE-Wiki*), which we consider to be more challenging than DE-FIGER, because it contains a wider range of type labels; **3)** a small challenge set of 135 sentences taken from DE-Wiki, in which we replaced entities with close string matches to English (e.g. 'Präsident Nixon') with specifically German entities of the same type (e.g. 'Bundeskanzler Kohl'), which we call *DE-GermEnt*; and **4)** for experiments where we mix German and English data, we also compare against test split of the English FIGER corpus (Ling and Weld, 2012) (*EN-FIGER*). Data set statistics can be seen in table 2.

5 Results

Monolingual training Figure 2 compares the performance of the models *zero-shot gold*, *zero-shot automatic* and *annotation projected* at different training data sizes on the *DE-FIGER* and the *DE-Wiki* test sets. *Zero-shot gold* outperforms *zero-shot automatic* and *annotation projected* on both test sets and in all training data sizes. *Zero-shot gold* trained on the full *EN gold* data set of 2 M sentences performs only 1% better on level 1 labels and 3% better on level 2 labels than a model trained with 400 K sentences, which shows that smaller data slices are sufficient to reach most of the possible performance with this data set.

	size	unique lab.	t. lab.
EN-FIGER	563 sent.	42	624
DE-FIGER	563 sent.	42	624
DE-Wiki	500 sent.	57	771
DE-GermEnt	135 sent.	34	213

Table 2: Statistics of the different test sets listing size, number of unique labels and total number of labels. While DE-FIGER is parallel to the commonly used English FIGER test set, DE-Wiki contains more unique labels and more labels in total.

While for level 1 type labels *annotation projected* gets close to the performance of *zero-shot gold* on both test sets, on level 2 type labels the system falls behind *zero-shot gold*, with a wider gap on *DE-Wiki*. The comparison between *zero-shot automatic* and *annotation projected* is less clear. On the *DE-Wiki* test set *annotation projected* consistently outperforms *zero-shot automatic*, while on *DE-FIGER* both systems perform very similarly.

The high performance of *zero-shot gold* and the noisier *zero-shot automatic* might be due to the quality of English and German embeddings in XLM-RoBERTa, as both are high resource languages from the same language family. This confirms Lauscher et al. (2020) who show that this method works especially well for close high resource language pairs and low level semantic tasks. The noise introduced by annotation projection affects level 2 label performance the most (see appendix B and table 1). But the amount of level 2 labels in the training data can not be the only reason for this. The total number of labels in the silver corpora (see table 1) shows that 200 K of silver training data contain approximately the same amount of level 2 labels as 100 K of gold data. Nevertheless, the level 2 performance of systems trained on 200 K of silver data lies behind the model trained on 100 K of *EN gold*. This points towards the possibility, that not only the amount of level 2 labels in the training data, but also their quality and their

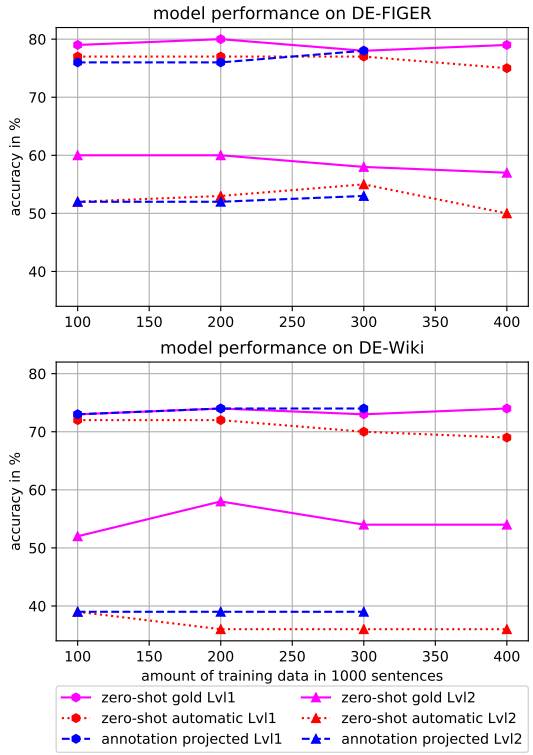

Figure 2: Zero-shot cross-lingual transfer performs best on both German data sets. *EN automatic* and *DE projected* perform similar on both data sets, with a wider gap in level 2 performance on *DE-Wiki*

proportion with level 1 labels play a role here.

Multilingual training The underlying XLM-RoBERTa embeddings allow to train a model with both German and English data. For this we combine slices from *DE projected* with *EN automatic*, because these data sets have the same distribution of labels. Table 3 shows the performance of a model trained with evenly mixed data (*EN+DE*) in comparison with monolingually trained models of the same size tested on *DE-FIGER* and *EN-FIGER*. German performance benefits from using both German and English training data, while performance in English is best with only English data. The mixed model does not outperform *zero-shot gold* on these test sets.

The low performance in the data mixing scenario compared to *zero-shot gold* can be explained with the distribution of labels in the silver corpora. Due to the noise added when labels are projected from English to German, the mixed model tested in German profits from the addition of higher quality English data, but not vice versa.

Few-shot training Zhao et al. (2020) suggest that few-shot learning improves zero-shot performance. To test this we take a model trained on 100 K sentences *EN gold* and fine-tune it by training on the 135 sentence manually annotated *DE-GermEnt* data set. We evaluate the resulting model's performance on *DE-FIGER*. In comparison with the model trained on 100 K *EN gold* only, the performance of the resulting model is 10% lower in accuracy of level 1 labels and 12% lower on level 2 labels. We did not specifically select which sentences to use like Zhao et al. (2020), which is an avenue for future work. The low performance of the few-shot model could be due to the high number of different labels, only a few of which can be observed during few-shot training, but further work is needed to confirm this.

	DE-FIGER		EN-FIGER	
name, size	Lvl1	Lvl2	Lvl1	Lvl2
DE proj., 200K	75	52	79	54
DE+EN, 200K	**77**	**54**	78	54
EN aut., 200K	76	53	**79**	**55**

Table 3: Performance of a model trained with both English and German data, in comparison with monolingual data tested on parallel test sets. While performance in German is best with mixed data, performance in English is best with only English training data.

German entities To challenge *zero-shot gold*, we test a model trained 2 M sentences *EN gold* on the test set *DE-GermEnt*. Surprisingly, we find that the model performs better on *DE-GermEnt* than on its English entity counter part, with 1% higher performance on level 1 labels and 3% higher performance on level 2 labels. It is unclear why *zero-shot gold* behaves this way, and examining this with larger challenge data sets it an avenue for future work.

6 Discussion and Conclusions

Our results show that zero-shot cross-lingual transfer building upon XLM-RoBERTa is a strong baseline for the task of FET and the language pair of English and German. It outperforms annotation projection on three new test sets. We also show that in our specific scenario annotation projection using the hierarchical typing model amplifies the models tendency to underpredict level 2 types.

One way to mitigate these shortcomings would be to sample level 1 and level 2 labels in a training corpus so that they have the same distribution as in the gold data, although this would not control

for data quality. Another way could be to machine translate the manually annotated English corpus into German and then use annotation projection, as suggested by Ehrmann et al. (2011). This way the label distribution of the human annotated data could be preserved as well. Lastly, improving the few-shot approach and designing more challenging test sets are other avenues to explore.

7 Acknowledgements

This work was funded by the ERC H2020 Advanced Fellowship GA 742137 SEMANTAX.

References

Rodrigo Agerri, Yiling Chung, Itziar Aldabe, Nora Aranberri, Gorka Labaka, and German Rigau. 2018. Building named entity recognition taggers via parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

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.

Tongfei Chen, Yunmo Chen, and Benjamin Van Durme. 2020. Hierarchical entity typing via multi-level learning to rank. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8465–8475.

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. *arXiv preprint arXiv:1911.02116*.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Maud Ehrmann, Marco Turchi, and Ralf Steinberger. 2011. Building a multilingual named entity-annotated corpus using annotation projection. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 118–124.

Tsung-Yuan Hsu, Chi-Liang Liu, and Hung-yi Lee. 2019. Zero-shot reading comprehension by cross-lingual transfer learning with multi-lingual language representation model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5933–5940, Hong Kong, China. Association for Computational Linguistics.

Jun Kuang, Yixin Cao, Jianbing Zheng, Xiangnan He, Ming Gao, and Aoying Zhou. 2020. Improving neural relation extraction with implicit mutual relations. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1021–1032. IEEE.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499.

Bing Li, Yujie He, and Wenjin Xu. 2021. Cross-lingual named entity recognition using parallel corpus: A new approach using xlm-roberta alignment. *arXiv preprint arXiv:2101.11112*.

Xiao Ling and Daniel S Weld. 2012. Fine-grained entity recognition. In *AAAI*, volume 12, pages 94–100.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1470–1480.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual pos taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings*

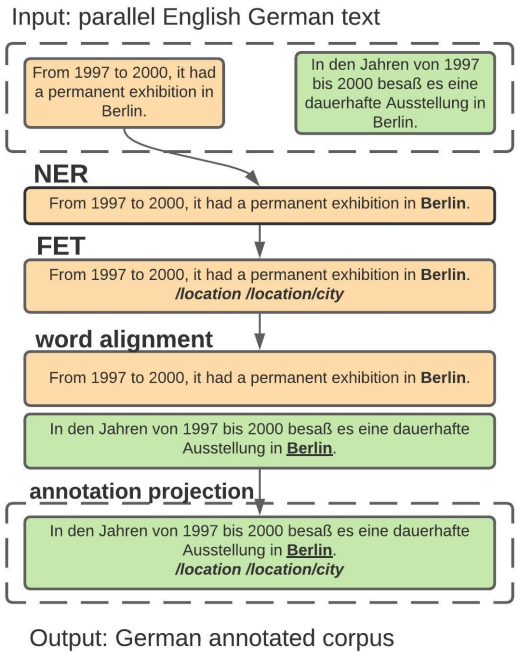

Figure 3: Our annotation projection setup uses parallel text and an automatic named entity recognition component to generate an annotated corpus in German.

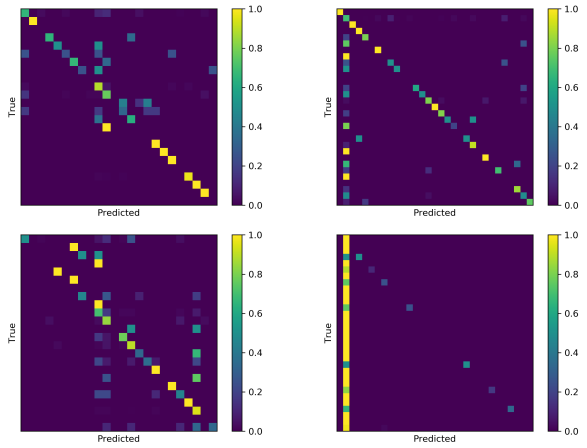

Figure 4: The upper two matrices show level 1 and level 2 performance of EN gold 2 M, the lower two matrices show the same for DE silver 200K. While the English model has a slight tendency to predict no label for level 2 label, this tendency is stronger in the German model. The yellow vertical line shows this effect.

of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics, pages XXX–XXX. ACL.

Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2020. A closer look at few-shot crosslingual transfer: Variance, benchmarks and baselines. *arXiv preprint arXiv:2012.15682*.

A German Data Creation

A.1 Preprocessing

A diagram of our pipeline can be seen in figure 3. To annotate the English halves of our parallel corpora with FIGER types preprocessing is necessary. Due to its automatic creation the WikiMatrix corpus contains a small amount of German sentences in its English half and English sentences in its German half, the translations of which are assigned very high confidence. We remove these by discarding the 5000 highest-confidence sentences.

To enable annotation by the English FET system, we run a named entity recognition system over the English input sentences (see the second box of Figure 3). We used the NER component of Stanza (Qi et al., 2020) for this task. We then use the English FET model to assign FIGER types to the named entities (see the third box of Figure 3). The FET model only annotates one entity per sentence. Sentences that contain more than one named entity occur multiple times in the English input, so that each entity receives an annotation.

A.2 Annotation projection and training with noise mitigation

We use ZAP to obtain a word alignment between the English and German halves of our parallel corpora (see the fourth box of Figure 3). While the similar tools fast_align (Dyer et al., 2013) and Giza++ (Och and Ney, 2003) are language agnostic, ZAP's model for English-German word alignment uses probabilities computed from large-scale parallel corpora. We then use our own code to project the fine grained entity type labels from the annotated English text to its German translation. We use static rules to filter out misalignments, e.g. discarding all cases where not all words of an entity were aligned.

We then use the resulting German FET annotated corpus to train our FET model. Because of the ordering by alignment quality in the machine-aligned WikiMatrix corpus, we introduce a preprocessing epoch to the training to mitigate noisy input. During training the model receives the sentences in exactly the order that they occur in the corpus. In the WikiMatrix corpus the sentences are sorted by the confidence of the alignment algorithm. This means that the sentences towards the bottom of the corpus are more likely to be incorrectly aligned. Incorrectly aligned sentences are more likely to have incorrectly projected labels. Therefor the quality of FIGER type annotations in the resulting German

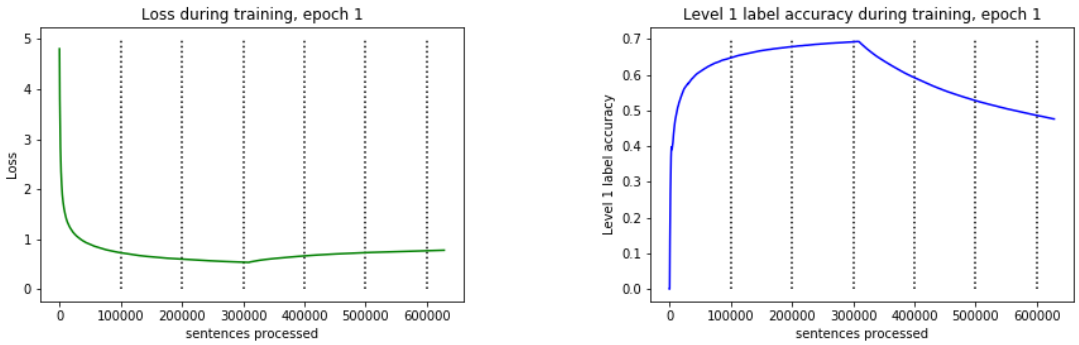

Figure 5: Loss rises and level 1 label accuracy deteriorates as the quality of samples gets worse towards the end of the automatically aligned and sorted corpus. The graphs show a cut-off point at approximately 300 thousand sentences. We use this information to select a high-quality slice of the corpus to train our system.

data is higher towards the beginning of the corpus and lower towards its end.

During the first epoch of training this drop in quality can be observed in the change of learning rate and the accuracy of predictions after approximately 300,000 sentences (see Figure 5). These curves give us important information about what portion of the data is clean enough to be used in the following epochs of training. It gives us a possible cut off point for our data set at 300,000 sentences, so that in the next epochs we only train on a slice of the corpus before this point.

To show the effect of increasing training data size we select for our experiments 3 slices of data that were processed before the cut off point: the first 100K, 200K and 300K sentences of the corpus.

B Confusion matrices

The hierarchical typing model of Chen et al. (2020)'s model tends to under-predict level 2 labels. This property of the model is exacerbated by our annotation projection approach, because we use data generated by this model in English to train the same model in German. Figure 4 shows the confusion matrices for level 1 and level 2 labels for EN gold 2 Mil and DE silver 200K. While for level 1 labels in EN gold there is no dominant class that labels are misclassified to, the most common misprediction for level 2 labels is to assign no label at all, which can be seen as the dotted vertical line in the second upper confusion matrix. When comparing the upper confusion matrices to the lower ones, it becomes clear that this trend to under-predict level 2 labels is even stronger in the German model.

This makes sense because the German model is trained on output from the English model, which contains fewer level 2 labels than the human annotated data used to train the English model in he first place. The German model sees less level 2 labels in its training data and therefore doesn't learn to predict them.

C System and model specifications

In keeping with the EMNLP reproducibility guidelines we report the specifications of the systems that our models where trained on. We trained all models using a single GeForce RTX 2080 Ti GPU. Running the largest model (EN zero-shot trained on 2 million sentences) took approximately 8 hours. Training the other models took under an hour per model. The number of model parameters is 50484362. All hyperparameters of the model were taken from the implementation of Chen et al. (2020). For the few-shot experiment we increased the training epoch number to stop if there was no more improvement on the development set.