Cross-Language Transfer of High-Quality Annotations: Combining Neural Machine Translation with Cross-Linguistic Span Alignment to Apply NER to Clinical Texts in a Low-Resource Language

Henning Schäfer[1,2] **Ahmad Idrissi-Yaghir**[2,3] **Peter A. Horn**[1] **Christoph M. Friedrich**[2,3]
[1]Institute for Transfusion Medicine, University Hospital Essen,
Hufelandstraße 55, 45147 Essen, Germany
[2]Department of Computer Science, University of Applied Sciences and Arts Dortmund
(FH Dortmund), Emil-Figge-Straße 42, 44227 Dortmund, Germany
[3] Institute for Medical Informatics, Biometry and Epidemiology (IMIBE),
University Hospital Essen, Hufelandstraße 55, 45147 Essen, Germany
{henning.schaefer,peter.horn}@uk-essen.de
{ahmad.idrissi-yaghir,christoph.friedrich}@fh-dortmund.de

Abstract

In this work, cross-linguistic span prediction based on contextualized word embedding models is used together with neural machine translation (NMT) to transfer and apply the state-of-the-art models in natural language processing (NLP) to a low-resource language clinical corpus. Two directions are evaluated: (a) English models can be applied to translated texts to subsequently transfer the predicted annotations to the source language and (b) existing high-quality annotations can be transferred beyond translation and then used to train NLP models in the target language. Effectiveness and loss of transmission is evaluated using the German Berlin-Tübingen-Oncology Corpus (BRONCO) dataset with transferred external data from NCBI disease, SemEval-2013 drug-drug interaction (DDI) and i2b2/VA 2010 data. The use of English models for translated clinical texts has always involved attempts to take full advantage of the benefits associated with them (large pre-trained biomedical word embeddings). To improve advances in this area, we provide a general-purpose pipeline to transfer any annotated BRAT or CoNLL format to various target languages. For the entity class medication, good results were obtained with 0.806 $F1$-score after re-alignment. Limited success occurred in the diagnosis and treatment class with results just below 0.5 $F1$-score due to differences in annotation guidelines.

1 Introduction

Clinical texts contain many important buried information that can be accessed through natural language processing (NLP). Systematic analysis of this vast amount of data can improve clinical care and aid in decision making. There are many other applications already in use, such as cohort selection, pharmacovigilance, and quality reporting (Spasić et al., 2020). Clinical text is often available as unstructured texts: Retrospective analysis therefore involves an enormous amount of work (Wu et al., 2019). By using NLP, biomedical concepts can be extracted and processed using named entity recognition (NER), allowing large amounts of text on specific topics of interest to be retrospectively analyzed. While biomedical text is intended for publications, clinical text is written by and aimed at health care professionals. They are written under time pressure and are heterogeneous in terms of abbreviations, omission of words, and medical jargon to keep information density high (Leaman et al., 2015).

Compared to English texts, the processing of non-English clinical texts by NLP is far from what is actually possible by the current state-of-the-art (Névéol et al., 2018; Schneider et al., 2020). This is due to the fact that in the U.S., Health Insurance Portability and Accountability (HIPAA) clearly regulates which 18 different identifiers of protected health information (PHI) must be removed in order for a document to be considered anonymized, creating many facilitators for de-identification of clinical texts (Yogarajan et al., 2020; Ahmed et al., 2020). Based on these rules, large clinical datasets such as Medical Information Mart for Intensive Care III (MIMIC-III) (Johnson et al., 2016) and shared tasks with high-quality annotations have been published, resulting in research and tools for processing English clinical texts being widely developed.

Proceedings of the 4th Clinical Natural Language Processing Workshop, pages 53 - 62
July 14, 2022 ©2022 Association for Computational Linguistics

With regard to the availability of NLP tools for other languages, there are major differences, for example in the processing of German clinical texts: Anonymization is left to individual institutions, data protection officers, and ethics committees, which means that there are no uniform regulations. The state-of-the-art for German texts lags behind and, despite great efforts (Hahn et al., 2018), continues to be limited to rule-based systems (Roller et al., 2020) or is often based on in-house data (Richter-Pechanski et al., 2021), which means that neither the data nor the trained models can be shared (Carlini et al., 2021). Freely available large anonymized datasets with high-quality annotated German clinical texts are therefore non-existent.

In order to bridge this gap, this work provides a general-purpose pipeline to transfer annotated datasets in BRAT or CoNLL format to various target languages[1]. Approaches based on neural machine translation (NMT) have recently been applied to NER tasks (Xie et al., 2018; Mayhew et al., 2017; Yan et al., 2021). Improved translation quality through advances in neural machine translation (Ng et al., 2019; Tran et al., 2021) have reached a level that allows the transfer of predictions or annotated data in combination with word alignments (Jalili Sabet et al., 2020; Dou and Neubig, 2021) to other languages.

In this work, the Berlin-Tübingen-Oncology Corpus (BRONCO) (Kittner et al., 2021) is used and treated as a zero-resource dataset, for which English models and external biomedical and clinical datasets are used instead. The aim is to evaluate whether low-resource languages can benefit from the available English resources. The methodology of this work can be applied to other clinical datasets and languages, as word alignment with contextualized embeddings through multilingual BERT (Devlin et al., 2019) covers 104 languages. Accordingly, multilingual models are available for translation, e.g., the mBART (Tang et al., 2021) many-to-many model covers 50 languages.

2 Data

The BRONCO corpus (Kittner et al., 2021) is the first small, fully anonymized dataset for German clinical texts, that can be accessed via a data usage agreement form request. The dataset contains 200 discharge reports of hepatocellular carcinoma and melanoma, with 50 reports retained by the authors as independent test data. Due to strict data protection regulations and to make de-anonymization more difficult, the discharge summaries were shuffled into sentences so that the clinical context is only preserved at sentence level. It includes three annotated entity classes: diagnosis, medication and treatment (see Table 1). According to Kittner et al. (2021) the annotation process was performed by 2 groups of annotators, group A (2 medical experts) and group B (3 medical experts and 3 medical students). Conflicting annotations were resolved in the final version of BRONCO.

Table 1: Berlin-Tübingen-Oncology Corpus (BRONCO) descriptive statistics.

Entity	BRONCO 150	BRONCO 50	Total
Diagnosis	4,080	1,165	5,245
Treatment	3,050	816	3,866
Medication	1,630	383	2,013
Total	8,760	2,364	11,124
No. of Documents	150	50	200
No. of Sentences	8,976	2,458	11,434
No. of Tokens	70,572	19,370	89,942

For the 3 entity classes in BRONCO, 3 existing English external datasets are used. In order to use external data, the underlying documents and annotation guidelines should match if possible.

2.1 Medication

To fine-tune models for recognizing medication entities in BRONCO, the SemEval-2013 drug-drug interaction (DDI) (Segura-Bedmar et al., 2011) corpus will serve as an external English resource. The corpus is semantically annotated and contains documents describing drug-drug interactions from the DrugBank database and MEDLINE, and includes annotated medication text-spans. It is the only corpus that covers both generic names and brand names.

2.2 Diagnosis

The BRONCO entity class diagnosis is defined by the annotation guidelines as a disease, symptom or medical observation that can be matched with the German modification of the International Classification of Diseases (ICD-10). The NCBI disease corpus (Doğan et al., 2014) is used for this purpose, although it differs in terms of document style and annotation guidelines.

[1] https://github.com/0xhesch/CLAT-cross-lingual-annotation-transfer

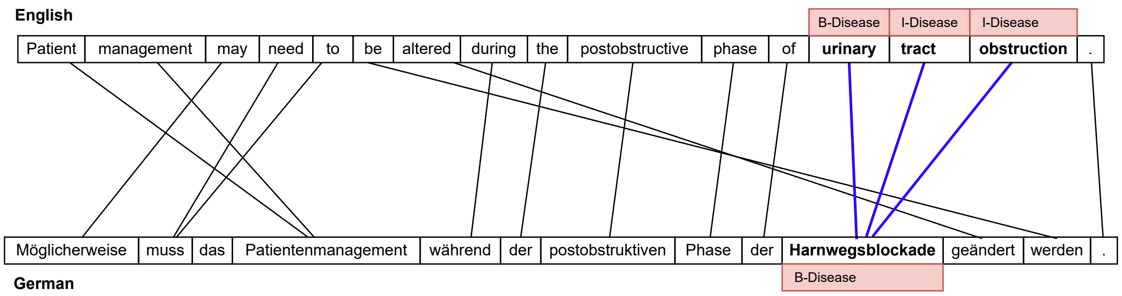

Figure 1: Behaviour of words in relation to the translated target language (here German). There are considerable differences in sentence structure, number of words and required word count for the description of a single medical term. The source to target token positions are to be resolved by word alignment systems to transfer annotations across languages.

Table 2: The table shows a sentence in BIO format from the NCBI dataset, translated into German along with the aligned annotation of the tokens.

Identification	O	Identifizierung	O
of	O	von	O
APC2	O	APC2	O
,	O	,	O
a	O	einem	O
homologue	O	Homologen	O
of	O	des	O
the	O	Tumorsuppressors	B-Disease
adenomatous	B-Disease	der	O
polyposis	I-Disease	adenomatösen	B-Disease
coli	I-Disease	Polyposis	I-Disease
tumour	I-Disease	coli	I-Disease
suppressor	O	.	O
.	O		

2.3 Treatment

Analogous to diagnosis, the BRONCO treatment class is a diagnostic procedure, e.g., surgery or systemic cancer treatment, found in the German Operationen and Prozedurenschlüssel (OPS) coding system. There is no exact match for this, although the treatment class of the i2b2/VA 2010 challenge data (Uzuner et al., 2011) shows overlapping annotation guidelines. Here, the treatment class also comprises medications which has to be taken into account in the methodology.

3 Methods

The experiments are divided into two parts. First, the German clinical dataset is treated as a zero-resource problem. This means that none of the annotated data is used to develop recognition models for the three entity classes diagnosis, medication and treatment. Instead, three existing English high-quality annotated datasets as described in Section 2 are used to train on the entity classes. Inferences are either made based on the translation and are then retroactively aligned to the German text, or models are fine-tuned on the translated form of the English datasets and are directly applied to the German clinical texts.

The second part focuses on the extent to which English pre-trained biomedical language models can be adapted for use in another language. For this purpose, the German dataset is translated and aligned in order to fine-tune large English pre-trained biomedical transformer-based models. The inference is then re-aligned to the German language. This is compared to non-biomedical German and cross-lingual transformer-based language models. In this way, the loss due to translation and subsequent alignment can be determined and weighed against the benefits of large biomedical language models that would not otherwise be available.

Based on current benchmarks (Ng et al., 2019; Tran et al., 2021), the selection for translation models fell on the directional WMT 19 en ↔ de model from Facebook AI Research (FAIR) as well as the multilingual WMT 21 model that covers 7 different languages. Since careful review of the translation quality of some clinical texts did not reveal any relevant deficiencies, the more resource-friendly WMT 19 model was chosen. For the span alignment of the annotations, Simalign (Jalili Sabet et al., 2020) is used without fine-tuning a parallel corpus. The work of Jalili Sabet et al. (2020) has shown that word alignments via contextualized embeddings from multilingual language models achieve good results. Here, the Itermax algorithm is used with contextualized word embeddings from multilingual BERT (Devlin et al., 2019). Itermax aligns two parallel sentences at token level with cosinesimilarity, where for each token the parallel vectors co-represent the context of the token within its sentence. Since for many sentences no mutual argmaxes are available, the suggestion mentioned by the authors to perform this process iteratively is followed. This also allows for token of the source language to be mapped to multiple token in the target language. This seems reasonable for clinical entities. For example, *urinary tract obstruction* is merged to only one token *Harnwegsblockade* in the German language (see Figure 1).

For fine-tuning language models, all experiments use the hyperparameters as described in Table 6. All experiments were conducted on an NVIDIA V100 SXM2 GPU.

3.1 Zero-Resource

Here, two variants seem reasonable. First, datasets with annotations can be translated from en → de (forward-passed), thereby training models directly in the target language. On the other hand, low-resource language texts can be translated into English (de → en) and the prediction subsequently re-aligned (en → de) to the originating language (backward-pass). Both variants are visualized as detailed workflows in Figure 2 (forward-pass) and Figure 3 (backward-pass).

3.1.1 Forward-Pass

For medication, the DDI corpus will be forward-passed to predict medication mentions in German text. The DrugBank, as well as the MEDLINE portion of the dataset, are merged. Except for drugs and brand names, all other entities are omitted. The two entity classes drug and brand name are then merged into a single medication entity class.

For diagnosis, the NCBI data is forward-passed. The general process of translation and word alignment for this class is shown as an example in Figure 1. A sample sentence of the resulting translated German NCBI corpus is shown in Table 2.

For treatment, the i2b2/VA 2010 challenge data is forward-passed. The i2b2 annotation guidelines state, that treatment also covers medication. Prior to training the model on the treatment entity class, drug predictions based on the DDI model that overlap with i2b2 treatment entities are therefore removed.

3.1.2 Backward-Pass

For the backward-pass, the three external resources are used untranslated to directly fine-tune Bio_Discharge_Summary_BERT (Alsentzer et al., 2019), a state-of-the-art biomedical language model that was initialized with BioBERT (Lee et al., 2019) and then further trained on discharge summaries from MIMIC-III.

For prediction, the German BRONCO 150 dataset is then translated into English using FAIR's WMT 19 model de → en, without word alignments. The inference on translated BRONCO 150 sentences are then re-aligned with the original German sentences.

3.2 Fine-Tuning

This experiment aims to determine the loss incurred by translation and re-alignment for named entity recognition within the clinical domain and uses a large pre-trained biomedical language model. Note that this does require available annotations. Since the initial baseline of the authors of the BRONCO dataset does not include transformer-based results, this experiment also covers cross-lingual and German-specific pre-trained experiments. At the same time, these experiments will test whether non-biomedical models are suitable for German clinical texts. For this purpose mBERT (Devlin et al., 2019), GBERT (Chan et al., 2020), GELECTRA (Chan et al., 2020) and XLM-R (Conneau et al., 2020) are used in the base, as well as in the large versions if available.

To take advantage of English biomedical pre-trained language models, Bio_Discharge_Summary_BERT is used as described in Figure 3 which means that the inference takes place on the translation and the annotations are retroactively aligned. BRONCO 150 results are reported through 5-fold cross-validation. For BRONCO 50 evaluation, the models are trained on the full BRONCO 150 data. Results on BRONCO 50 are reported independently by the dataset authors. The evaluation is done by providing the models, as well as the pipeline for translation and retroactive alignments. Since the evaluation on BRONCO 50 must be performed by the curators, the range of models is limited here.

4 Results

4.1 Zero-Resource

The results based on the external data are reported for all 3 entity classes to see if there are differences between translating external datasets into the target language or aligning the inference of the English

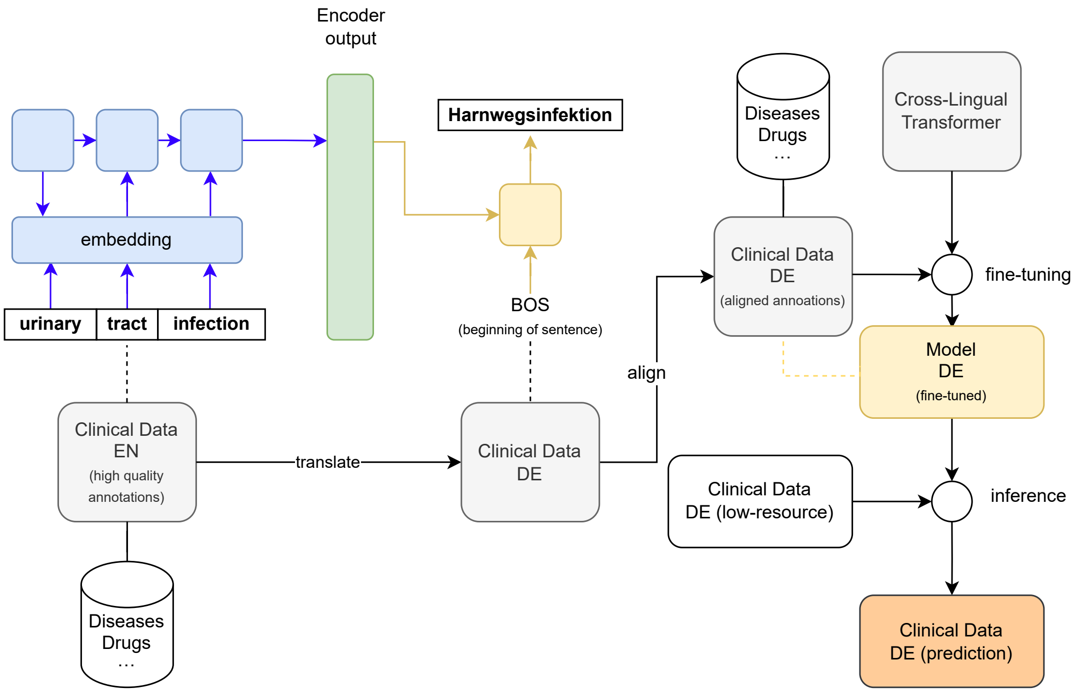

Figure 2: Schematic workflow (forward-pass) to perform prediction for clinical data with few resources. Here, the external English data is translated with annotations and then used to fine-tune cross-lingual language models for the target language. Prediction is then directly applied to the target language.

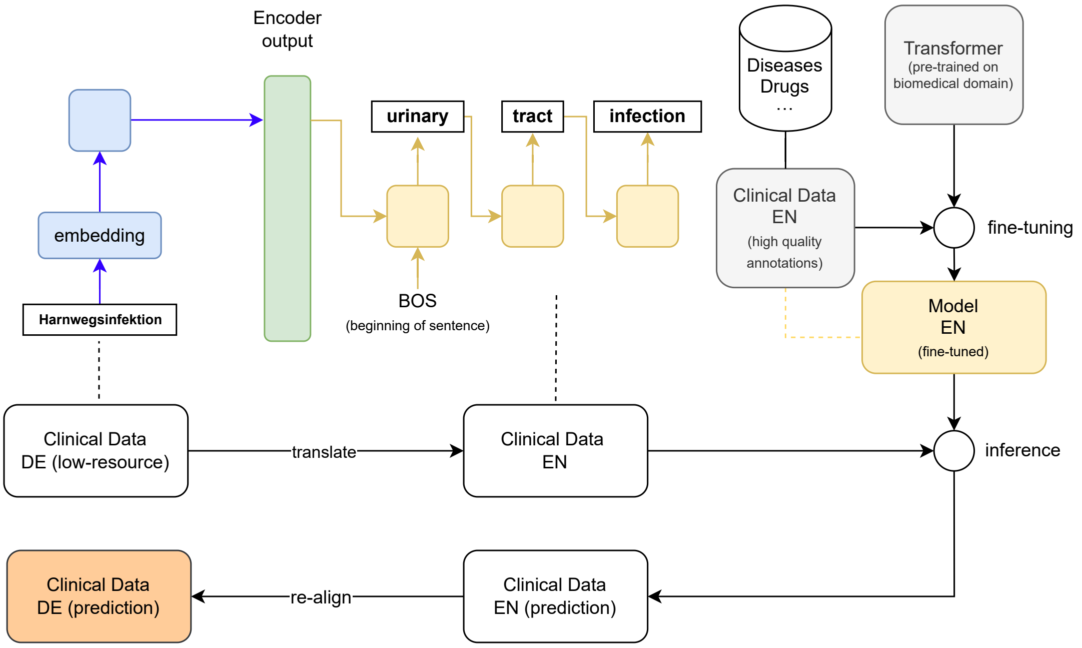

Figure 3: Schematic workflow (backward-pass) to perform prediction for clinical data with few resources. Models trained on external English data are applied to the translation and the prediction is aligned retrospectively.

Table 3: Forward- and backward-pass results on 3 entity classes for BRONCO 150 corpus through external data sources. State denotes if the results were obtained before- or after re-alignment in backward-pass runs.

Target Entity	Method	State	External Data Source	Model	Precision	Recall	F1
Medication	Forward-pass	-	DDI Corpus	deepset/gbert-base	0.637	0.809	0.712
	Forward-pass	-		deepset/gelectra-base	0.605	0.824	0.698
	Forward-pass	-		deepset/gelectra-large	**0.803**	0.769	0.785
	Forward-pass	-		bert-base-multilingual-cased	0.525	0.793	0.631
	Forward-pass	-		xlm-roberta-base	0.600	0.816	0.692
	Forward-pass	-		xlm-roberta-large	0.782	0.798	0.790
	Backward-pass	before re-alignment		Bio_Discharge_Summary_BERT	0.745	0.729	0.737
	Backward-pass	after re-alignment		Bio_Discharge_Summary_BERT	0.788	**0.826**	**0.806**
Diagnosis	Forward-pass	-	NCBI-Disease Corpus	deepset/gbert-base	0.433	0.445	0.439
	Forward-pass	-		deepset/gelectra-base	0.410	0.374	0.391
	Forward-pass	-		deepset/gelectra-large	**0.537**	0.419	0.471
	Forward-pass	-		bert-base-multilingual-cased	0.469	0.370	0.414
	Forward-pass	-		xlm-roberta-base	0.482	0.354	0.408
	Forward-pass	-		xlm-roberta-large	0.476	0.387	0.427
	Backward-pass	before re-alignment		Bio_Discharge_Summary_BERT	0.502	0.378	0.431
	Backward-pass	after re-alignment		Bio_Discharge_Summary_BERT	0.524	**0.474**	**0.498**
Treatment	Forward-pass	-	i2b2/VA 2010	deepset/gbert-base	0.510	0.429	0.466
	Forward-pass	-		deepset/gelectra-base	0.521	0.456	0.486
	Forward-pass	-		deepset/gelectra-large	0.523	0.454	0.486
	Forward-pass	-		bert-base-multilingual-cased	0.473	0.402	0.434
	Forward-pass	-		xlm-roberta-base	0.504	0.411	0.453
	Forward-pass	-		xlm-roberta-large	0.526	0.434	0.475
	Backward-pass	before re-alignment		Bio_Discharge_Summary_BERT	0.476	0.387	0.427
	Backward-pass	after re-alignment		Bio_Discharge_Summary_BERT	**0.536**	**0.463**	**0.497**

Table 4: Average results of 5-fold cross-validation for BRONCO 150 with reported standard deviation. † denotes initial baseline results by Kittner et al. (2021).

Target Entity	Model	Precision	Recall	F1
Medication	CRF†	0.960 (0.008)	0.850 (0.020)	0.900 (0.009)
	CRF+WE†	**0.960 (0.004)**	0.840 (0.009)	0.900 (0.006)
	LSTM†	0.910 (0.050)	0.860 (0.030)	0.880 (0.020)
	LSTM+WE†	0.960 (0.020)	0.870 (0.060)	0.910 (0.040)
	deepset/gbert-base	0.923 (0.019)	0.935 (0.016)	0.929 (0.012)
	deepset/gbert-large	0.929 (0.027)	0.941 (0.018)	0.935 (0.011)
	deepset/gelectra-base	0.850 (0.011)	0.912 (0.013)	0.880 (0.012)
	deepset/gelectra-large	0.951 (0.006)	**0.956 (0.018)**	**0.954 (0.008)**
	bert-base-multilingual-cased	0.926 (0.024)	0.937 (0.009)	0.931 (0.013)
	xlm-roberta-base	0.923 (0.005)	0.932 (0.014)	0.927 (0.006)
	xlm-roberta-large	0.929 (0.011)	0.941 (0.018)	0.935 (0.011)
Diagnosis	CRF†	0.800 (0.010)	0.710 (0.020)	0.750 (0.020)
	CRF+WE†	0.782 (0.006)	0.700 (0.020)	0.740 (0.010)
	LSTM†	0.750 (0.030)	0.690 (0.030)	0.720 (0.010)
	LSTM+WE†	**0.810 (0.080)**	0.740 (0.080)	0.770 (0.080)
	deepset/gbert-base	0.744 (0.012)	0.802 (0.020)	0.772 (0.016)
	deepset/gbert-large	0.769 (0.009)	0.814 (0.015)	0.791 (0.008)
	deepset/gelectra-base	0.692 (0.023)	0.773 (0.026)	0.730 (0.022)
	deepset/gelectra-large	0.789 (0.008)	**0.826 (0.013)**	**0.807 (0.008)**
	bert-base-multilingual-cased	0.740 (0.017)	0.797 (0.022)	0.768 (0.019)
	xlm-roberta-base	0.728 (0.012)	0.792 (0.018)	0.759 (0.013)
	xlm-roberta-large	0.767 (0.012)	0.815 (0.014)	0.790 (0.007)
Treatment	CRF†	**0.860 (0.020)**	0.780 (0.010)	0.820 (0.010)
	CRF+WE†	0.850 (0.020)	0.780 (0.010)	0.810 (0.010)
	LSTM†	0.830 (0.040)	0.790 (0.030)	0.810 (0.020)
	LSTM+WE†	0.850 (0.060)	0.820 (0.070)	0.840 (0.060)
	deepset/gbert-base	0.783 (0.009)	0.830 (0.012)	0.806 (0.009)
	deepset/gbert-large	0.796 (0.023)	0.846 (0.019)	0.820 (0.020)
	deepset/gelectra-base	0.678 (0.015)	0.791 (0.023)	0.730 (0.017)
	deepset/gelectra-large	0.821 (0.009)	0.856 (0.011)	**0.839 (0.010)**
	bert-base-multilingual-cased	0.783 (0.026)	0.839 (0.016)	0.810 (0.022)
	xlm-roberta-base	0.753 (0.005)	0.825 (0.008)	0.788 (0.005)
	xlm-roberta-large	0.821 (0.013)	**0.857 (0.017)**	0.839 (0.014)

Table 5: Results for BRONCO 50. † denotes initial baseline results by Kittner et al. (2021). * denotes that the results are based on the translation and have been re-aligned.

Target Entity	Model	Precision	Recall	F1
Medication	CRF†	0.940	0.870	0.900
	CRF+WE†	**0.950**	0.850	0.900
	LSTM†	**0.950**	0.850	0.890
	LSTM+WE†	0.910	0.890	0.900
	deepset/gbert-base	0.929	**0.958**	**0.943**
	Bio_Discharge_Summary_BERT*	0.921	0.944	0.932
Diagnosis	CRF†	0.790	0.670	0.720
	CRF+WE†	0.770	0.660	0.710
	LSTM†	0.780	0.650	0.710
	LSTM+WE†	0.790	0.650	0.720
	deepset/gbert-base	**0.792**	**0.772**	**0.782**
	Bio_Discharge_Summary_BERT*	0.661	0.689	0.675
Treatment	CRF†	0.830	0.730	0.780
	CRF+WE†	0.810	0.730	0.760
	LSTM†	**0.850**	0.690	0.760
	LSTM+WE†	0.760	0.740	0.750
	deepset/gbert-base	0.782	**0.824**	**0.803**
	Bio_Discharge_Summary_BERT*	0.661	0.742	0.699

models. The results for the forward- and backward-pass are shown in Table 3. For all classes, the backward-pass resulted in better scores, although the difference compared to the forward-pass is not substantial. The results of the German and multilingual models are comparable to the results before the re-alignment step, i.e. on the BRONCO 150 translation. To estimate any loss that may occur due to the translation quality of the WMT 19 en ↔ de model, the case-sensitive SacreBLEU score (Post, 2018) on the re-translation of BRONCO150 is reported, which resulted in a score of 40.41. The medication class achieved the best results after re-alignment with 0.806 $F1$-score. The classes diagnosis and treatment both remained just below 0.5 $F1$-score, also after re-alignment. Aligning the annotations back to German, increases recall in particular, as e.g. in the case of medication by almost 0.1 $F1$-score. The forward-pass results show that large models are superior. A general outperformance of German-specific language models over multilingual language models is not present.

4.2 Fine-Tuning

Table 4 shows the 5-fold cross-validation results. Here, the BRONCO 150 dataset was fine-tuned using multiple German transformer-based language models and multilingual language models. In addition, the results are also compared to those reported in (Kittner et al., 2021). For all target entities, all transformer-based models except GELECTRA$_{base}$ outperformed the models used by Kittner et al. (2021) and achieved a better $F1$-score. Although, the Conditional Random Field (CRF) and Long Short-Term Memory (LSTM) models reported better precision for all classes, their recall scores were outperformed with the transformer-based models. Overall, the large transformer-based models achieved the highest scores, with GELECTRA$_{large}$ performing the best and reaching an $F1$-score of 0.954 ± 0.008 for medication, 0.807 ± 0.008 for diagnosis and 0.839 ± 0.010 for treatment. The model was followed by XLM-R$_{large}$, which was on par with GBERT$_{large}$ for all the target entities. Altogether, the results show that large German-specific language models perform the best, with XLM-R$_{large}$ being a strong multilingual language model that can even compete with task language-specific models.

The results achieved on the BRONCO 50 dataset show similar findings, where the German-specific language model GBERT$_{base}$ reached the best $F1$-score for all classes. Furthermore, the result achieved through translation and alignment was superior to the models reported in (Kittner et al., 2021) for medication, but these models were not as successful for the classes diagnosis and treatment.

5 Discussion

In the zero-resource setting, there is an advantage in the backward-pass approach over the forward-pass models. Good results could only be achieved for the medication class, but this is not necessarily due to translation and word alignment, but to the nature of the data. For the diagnosis and treatment class, there is no equivalent English dataset that fully matches the annotation guidelines of the German clinical text. The medication class seems unproblematic in that medication terms are more easily aligned, one-to-many token constellations due to translation are rare, and medications are often represented similarly in both languages. Nevertheless, the underlying sentence structure is fundamentally different between English and German, which means that the transfer of the results can be considered successful. Further limitations are discussed in Section 6.

These assumptions are also supported by the fine-tuning results, which show that although translation and alignment result in a loss, it is still competitive compared to the initial baseline. Only in the comparison with multilingual and German transformer architectures the disadvantage becomes clear. Provided that annotations are available, a general advantage of English biomedical models over non-

Table 6: Hyperparameters used for fine-tuning transformer-based models on external data and BRONCO 150.

Hyperparameter	Value
Batch size	64 (16 for large models)
Epochs	4
Manual seed	42
Learning rate	4e-5
Max sequence length	512
Optimizer	AdamW (Loshchilov and Hutter, 2019)
Adam epsilon	1e-8

domain language models on German clinical texts can therefore not be confirmed.

6 Conclusion and Future Work

The results of this work show that English language models can in principle be applied to other languages in clinical contexts. Translated training data can serve as a good basis and approach for languages where there are otherwise no resources. In a zero-resource scenario, the approach is limited to the extent that it works for data where the documents and annotation guidelines match across languages. Cross-linguistic differences in the available standards that annotators work with also play a limiting role here. BRONCO corpus is based on German ICD-10 and German OPS standards, which is also reflected in the annotation guidelines, making it difficult to apply external data.

Transfer in the clinical setting was evaluated with only one language pair (en ↔ de). Success with other language pairs depends not only on the annotation standard, but also on the similarity of the languages (grammar and morphology). Transfer can only succeed if the quality of translation and word alignment is sufficient, which can be expected between Indo-European languages, but can be much more difficult when transferring between language families.

Practical applications on other low-resource languages is left for future work. It would be interesting to see the effect of adding a few annotated samples to the external data. In this context, zero-shot and few-shot approaches would be a useful addition as a comparator. For comparison, it would also be helpful to have a non-alignment baseline that is fine-tuned to English data and directly infers German test data.

Acknowledgements

We thank Robert Martin of the Faculty of Mathematics and Natural Sciences at the Humboldt University of Berlin for the independent evaluation of the results on the BRONCO 50 test data.

This work was funded by a PhD grant from the DFG Research Training Group 2535 Knowledge- and data-based personalization of medicine at the point of care (WisPerMed), University of Duisburg-Essen, Germany.

References

Tanbir Ahmed, Md Momin Al Aziz, and Noman Mohammed. 2020. De-identification of electronic health record using neural network. *Scientific Reports*, 10(1):18600.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Udo Hahn, Franz Matthies, Christina Lohr, and Markus Löffler. 2018. 3000pa - towards a national reference corpus of german clinical language. In *Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth - Proceedings of MIE 2018, Medical Informatics Europe, Gothenburg, Sweden, April 24-26, 2018*, volume 247 of *Studies in Health Technology and Informatics*, pages 26–30. IOS Press.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

Madeleine Kittner, Mario Lamping, Damian T Rieke, Julian Götze, Bariya Bajwa, Ivan Jelas, Gina Rüter, Hanjo Hautow, Mario Sänger, Maryam Habibi, Marit Zettwitz, Till de Bortoli, Leonie Ostermann, Jurica Ševa, Johannes Starlinger, Oliver Kohlbacher, Nisar P Malek, Ulrich Keilholz, and Ulf Leser. 2021. Annotation and initial evaluation of a large annotated German oncological corpus. *JAMIA Open*, 4(2).

Robert Leaman, Ritu Khare, and Zhiyong Lu. 2015. Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics*, 57:28–37.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2536–2545, Copenhagen, Denmark. Association for Computational Linguistics.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of biomedical semantics*, 9(1):12.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Phillip Richter-Pechanski, Nicolas A Geis, Christina Kiriakou, Dominic M Schwab, and Christoph Dieterich. 2021. Automatic extraction of 12 cardiovascular concepts from german discharge letters using pre-trained language models. *DIGITAL HEALTH*, 7:20552076211057662. PMID: 34868618.

Roland Roller, Laura Seiffe, Ammer Ayach, Sebastian Möller, Oliver Marten, Michael Mikhailov, Christoph Alt, Danilo Schmidt, Fabian Halleck, Marcel Naik, Wiebke Duettmann, and Klemens Budde. 2020. Information extraction models for german clinical text. In *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–2.

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online. Association for Computational Linguistics.

Isabel Segura-Bedmar, Paloma Martínez, and Cesar de Pablo-Sánchez. 2011. Using a shallow linguistic kernel for drug–drug interaction extraction. *Journal of Biomedical Informatics*, 44(5):789–804.

Irena Spasić, Özlem Uzuner, and Li Zhou. 2020. Emerging clinical applications of text analytics. *International Journal of Medical Informatics*, 134:103974.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook AI's WMT21 news translation task submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 205–215, Online. Association for Computational Linguistics.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Stephen Wu, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, Qiong Wang, Qiang Wei, Yang Xiang, Bo Zhao, and Hua Xu. 2019. Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*, 27(3):457–470.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379, Brussels, Belgium. Association for Computational Linguistics.

Huijiong Yan, Tao Qian, Liang Xie, and Shanguang Chen. 2021. Unsupervised cross-lingual model transfer for named entity recognition with contextualized word representations. *PLOS ONE*, 16(9):1–17.

Vithya Yogarajan, Bernhard Pfahringer, and Michael Mayo. 2020. A review of automatic end-to-end de-identification: Is high accuracy the only metric? *Applied Artificial Intelligence*, 34(3):251–269.