Model and Data Transfer for Cross-Lingual Sequence Labelling in Zero-Resource Settings

Iker García-Ferrero Rodrigo Agerri German Rigau
HiTZ Basque Center for Language Technologies - Ixa NLP Group
University of the Basque Country UPV/EHU
{ iker.garciaf, rodrigo.agerri, german.rigau }@ehu.eus

Abstract

Zero-resource cross-lingual transfer approaches aim to apply supervised models from a source language to unlabelled target languages. In this paper we perform an in-depth study of the two main techniques employed so far for cross-lingual zero-resource sequence labelling, based either on data or model transfer. Although previous research has proposed translation and annotation projection (data-based cross-lingual transfer) as an effective technique for cross-lingual sequence labelling, in this paper we experimentally demonstrate that high capacity multilingual language models applied in a zero-shot (model-based cross-lingual transfer) setting consistently outperform data-based cross-lingual transfer approaches. A detailed analysis of our results suggests that this might be due to important differences in language use. More specifically, machine translation often generates a textual signal which is different to what the models are exposed to when using gold standard data, which affects both the fine-tuning and evaluation processes. Our results also indicate that data-based cross-lingual transfer approaches remain a competitive option when high-capacity multilingual language models are not available.

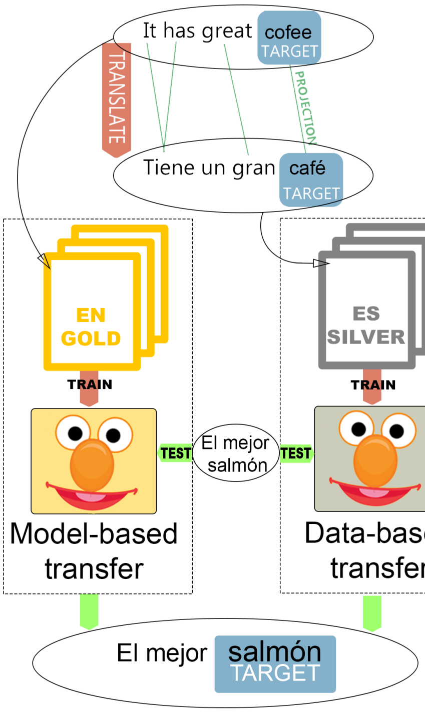


Figure 1: In the data-based transfer approach we translate and project the labels of the gold data into the target language, and use the resulting silver data to train a model for the target language. In the model-based transfer approach we train a model with gold data in English and use it in a zero-shot setting in the target language.

1 Introduction

Sequence labelling is the task of assigning a label to each token in a given input sequence. Sequence labelling is a fundamental process in many downstream NLP tasks. Currently, most successful approaches for this task apply supervised deep-neural networks (Lample et al., 2016; Akbik et al., 2018; Devlin et al., 2019; Conneau et al., 2020). However, as it was the case for supervised statistical approaches (Agerri and Rigau, 2016), their performance still depends on the amount of manually annotated training data. Additionally, deep-neural models still show a significant loss of performance when evaluated in out-of-domain data (Liu et al., 2021). This means that to improvie their performance, it would therefore be necessary to develop very costly manually annotated data for each language and domain of application. Thus, considering that for most of the languages in the world the amount of manually annotated corpora is simply nonexistent (Joshi et al., 2020), then the task of developing sequence labelling models for languages and domain-specific tasks, for which supervised data is not available, remains a challenge of great interest. This task is known as zero-resource cross-lingual sequence labelling.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 6403–6416
December 7-11, 2022 ©2022 Association for Computational Linguistics

Data-based cross-lingual transfer methods aim to automatically generate labelled data for a target language. Previous works on data-based transfer have proposed *translation and annotation projection* as an effective technique for zero-resource cross-lingual sequence labelling (Jain et al., 2019; Fei et al., 2020). In this setting, as illustrated in Figure 1, the idea is to translate gold-labelled text into the target language and then, using automatic word alignments, project the labels from the source into the target language. The result is an automatically generated dataset in the target language that can be used for training a sequence labelling model.

The emergence of multilingual language models (Devlin et al., 2019; Conneau et al., 2020) allows for **model-based** cross-lingual transfer. As Figure 1 illustrates, using labelled data in one source language (usually English), it is possible to fine-tune a pre-trained multilingual model that is directly used to make predictions in any of the languages included in the model. This is also known as *zero-shot* cross-lingual sequence labelling.

In this work we present an in-depth study of both approaches using the latest advancements in machine translation, word aligners and multilingual language models. We focus on two sequence labelling tasks, namely, Named Entity Recognition (NER) and Opinion Target Extraction (OTE). In order to do so, we present a data-based cross-lingual transfer approach consisting of translating gold labeled data between English and 7 other languages using state-of-the-art machine translation systems. Sequence labelling annotations are then automatically projected for every language pair. Additionally, we also produced manual alignments for those 4 languages for which we had expert annotators. After translation and projection, for the data-transfer approach we fine-tune multilingual language models using the automatically generated datasets. We then compare the performance obtained for each of the target languages against the performance of the zero-shot cross-lingual method, consisting of fine-tuning the multilingual language models in the English gold data and generating the predictions in the required target languages.

The main contributions of our work are the following: First, we empirically establish the required conditions for each of these two approaches, data-transfer and zero-shot model-based, to outperform the other. In this sense, our experiments show that, contrary to what previous research suggested (Fei et al., 2020; Li et al., 2021), the zero-shot model-based approach obtains the best results when high-capacity multilingual models including the target language and domain are available. Second, when the performance of the multilingual language model is not optimal for the specific target language or domain (for example when working on a text genre and domain for which available language models have not been trained), or when the required hardware to work with high-capacity language models is not easily accessible, then data-transfer based on *translate and project* constitutes a competitive option. Third, we observe that machine translation data often generates training and test data which is, due to important differences in language use, markedly different to the signal received when using gold standard data in the target language. These discrepancies seem to explain the larger error rate of the translate and project method with respect to the zero-shot technique. Finally, we create manually projected datasets for four languages and automatically projected datasets for seven languages. We use them to train and evaluate cross-lingual sequence labelling models. Additionally, they are also used to extrinsically evaluate machine translation and word alignment systems. These new datasets, together with the code to generate them are publicly available to facilitate the reproducibility of results and its use in future research.[1]

2 Related work

2.1 Data-based cross-lingual transfer

Data-based cross-lingual transfer methods aim to automatically generate labelled data for a target language. Some of these methods exploit parallel data. Ehrmann et al. (2011) automatically annotate the English version of a multi-parallel corpus and projects the annotations into all the other languages using statistical alignments of phrases. Wang and Manning (2014) project model expectations rather than labels, which facilities transfer of model uncertainty across languages. Ni et al. (2017) use a heuristic scheme that effectively selects good-quality projection-labeled data from noisy data. They also project word embeddings from a target language into a source language, so that the

[1] https://github.com/ikergarcia1996/Easy-Label-Projection
https://github.com/ikergarcia1996/Easy-Translate

source-language sequence labelling system can be applied to the target language without re-training. Agerri et al. (2018) use parallel data from multiple languages as source to project the labelled data to a target language, showing that the combination of multiple sources improves the quality of the projections. Li et al. (2021) uses the XLM-R model (Conneau et al., 2020) for labelling sequences in the source part of the parallel data and also for annotation projection.

Instead of relying on parallel data, Jain et al. (2019) and Fei et al. (2020), use machine translation to automatically translate the sentences of a gold-labelled dataset to the target languages. The translated data is then annotated by projecting the gold labels from the source dataset. For this purpose, Jain et al. (2019) first generate a list of projection candidates by orthographic and phonetic similarity. They choose the best matching candidate based on distributional statistics derived from the dataset. Fei et al. (2020) leverages the word alignment probabilities calculated with FastAlign (Dyer et al., 2013) and the POS tag distributions of the source and target words.

High quality parallel data or machine translation systems are not always available. Thus, Xie et al. (2018) proposes to find word translations based on bilingual word-embeddings. Alternatively, Guo and Roth (2021) translate labelled data in a word-by-word manner with a dictionary. Then, they the construct target-language text from the source-language annotations with a constrained pretrained language model.

2.2 Model-based transfer

Language models trained on monolingual corpora in many languages (Devlin et al., 2019; Conneau et al., 2020) allow zero-shot cross-lingual model transfer. Task-specific data in one language is used to fine-tune the model for evaluation in another language (Pires et al., 2019). The zero-shot cross-lingual capability can be improved for the sequence labelling task using different techniques. The approaches of Wang et al. (2019) and Ouyang et al. (2021) use monolingual corpora to improve the alignment of the language representations within a multilingual model. Instead of using a single source model, (Rahimi et al., 2019) propose to use many models from many source languages to improve the zero-shot transfer to a new language. They learn to infer which are the most reliable models in an unsupervised manner. Wu et al. (2020) take advantage of a Teacher-Student learning approach. NER models in the source languages are used as teachers to train a student model on unlabeled data in the target language. Bari et al. (2021) propose an unsupervised data augmentation framework to improve the cross-lingual adaptation of models using self-training. Hu et al. (2021) use the minimum risk training framework to overcome the gap between the source and the target languages/domains. They propose a unified learning algorithm based on the expectation maximization.

Using low-capacity multilingual language models such as mBERT, Fei et al. (2020) finds that their data-based cross-lingual transfer approach is superior to the zero-shot transfer method. However, Li et al. (2021) when using XLM-RoBERTa, a higher capacity multilingual model, obtain the best results for German and Chinese applying the data-based cross-lingual transfer approach, while the zero-shot approach is best for Spanish and Dutch. We extend their research on zero-resource settings with two different Sequence Labelling tasks, seven languages and three multilingual models of different capacity. Our experiments and the error analysis carried out establish the required conditions on which zero-shot and data-transfer approaches outperform each other.

3 Translation and projection method

Our data-based cross-lingual transfer method to perform cross-lingual sequence labelling is the following: we assume our source language to be English, for which we have train and development data. Furthermore, we also assume that the only gold-labelled data available for the target language is the evaluation set. In this setting, we automatically generate data for the target language by translating the gold-labelled English data. Then we project the gold labels from the source sentences to the translated sentences by leveraging automatic word alignments. Given a sentence $x = \langle x_1, ..., x_n \rangle$ with length n in the source language and a translated sentence $y = \langle y_1, ..., y_m \rangle$ with length m in the target language, we use a word aligner to find a set of pairs $A = \{\langle x_i, y_j \rangle : x_i \in \mathbf{x}, y_j \in \mathbf{y}\}$ where for each word pair $\langle x_i, y_j \rangle$ y_i is the lexical translation of x_j. Next, given a sequence $s = \langle x_a, ..., x_b \rangle \in \mathbf{x}$ labeled with a category C we will label the sequence $t = \langle y_c, ..., y_d \rangle \in \mathbf{y}$ with category C if $\{\forall y_j \in t\, \exists x_i \in s : (\langle x_i, y_j \rangle \in A)\}$. In other

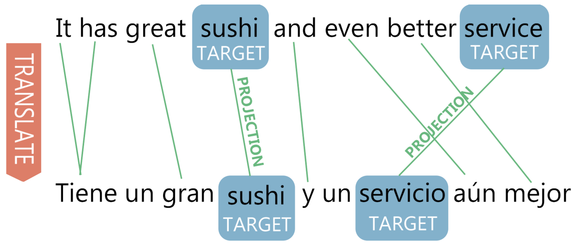
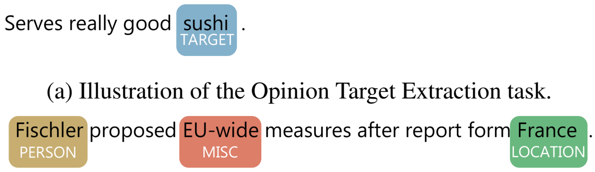

It has great sushi and even better service
TARGET
TARGET
TRANSLATE
PROJECTION
PROJECTION
Tiene un gran sushi y un servicio aún mejor
TARGET
TARGET

Figure 2: Illustration of the translation and annotation projection method for Opinion Target Extraction (OTE).


Serves really good sushi .
TARGET
Fischler proposed EU-wide measures after report form France .
PERSON
MISC
LOCATION

(a) Illustration of the Opinion Target Extraction task.

(b) Illustration of the Named Entity Recognition Task.

Figure 3: Sequence Labelling tasks used in our experiments.

words, if a word labelled with a category in the source sentence is aligned to a word in the target sentence, we label the target word with the category from the word in the source sentence. Figure 2 illustrates our method.

When projecting annotations we find two main problems: *split annotations* and *annotation collision*. In the first case, a labeled sequence in the source sentence is split into multiple sequences in the target sentence. This happens when the alignment for a word is missing. In this case, we merge the sequences in the target sentence if the gap between them is just one word. If we still end up with multiple sequences, we choose the longest one. In the *annotation collision* case, a word in the target sentence is aligned to two different labelled sequences in the source language. If the two sequences are of the same category, we merge them and we label the two sequences as a single one in the target sequence. If they are of different category we just consider the one with the longest length. Finally, if a punctuation symbol in the target sequence is aligned to a labeled word in the source sentence we remove this alignment.

4 Datasets

We conducted experiments in two sequence labelling tasks, namely, Opinion Target Extraction (OTE) and Named Entity Recognition (NER). Figure 3 illustrates both tasks.

Opinion Target Expression (OTE): Given a review, the task is to detect the linguistic expression used to refer to the reviewed entity. We use the SemEval-2016 Task 5 Aspect Based Sentiment Analysis (ABSA) datasets (Pontiki et al., 2016). We experiment with the English, Spanish, Dutch, French, Russian and Turkish datasets from the restaurant domain.

Named Entity Recognition (NER): Given a text, the task is to detect named entities and classify them in pre-defined categories. For Spanish and Dutch we use the CoNLL-2002 datasets (Tjong Kim Sang, 2002). For English and German we use the CoNLL-2003 datasets (Tjong Kim Sang and De Meulder, 2003) and for Italian we use Evalita 2009 data (Speranza, 2009). We map the Geo-Political Entities from Evalita 2009 to *location* labels to make them compatible with the CoNLL data.

5 Experimental Setup

We perform 1-to-1 annotation projection in two directions:

Translate-Train: We translate the English train and development data to the target language. We project the gold labels from the English data to the translated dataset. We then train a sequence labelling model using only the automatically generated dataset for the target language.

Translate-Test: We translate the target language test set to English. We then use a model trained in the English gold-labelled data to label the translated test set. Finally, we project the labelled sequences back to the target language.

These two data-based cross-lingual transfer approaches are compared with the **zero-shot** method in which a fine-tuned model using English gold-labelled data is evaluated by generating predictions in the target language. Finally, we also fine-tuned language models on the **gold**-labelled data, which would constitute the upper-bound in our experimental setting.

5.1 Machine Translation

We tested DeepL[2], MarianMT (Junczys-Dowmunt et al., 2018; Tiedemann and Thottingal, 2020), M2M100 (1.2B) (Fan et al., 2020) and mBART (mbart-large-50) (Tang et al., 2020). A qualitative analysis performed during the projection of the OTE labels established that DeepL produced the more fluent translations. Thus, we decided to

[2] https://www.deepl.com/

use DeepL (web version during the second half of 2021) to perform the machine translation for our data-based cross-lingual transfer experiments. The exception was Turkish, which is not supported by DeepL. In this case we use M2M100.

5.2 Word Alignments

For word alignments, we use the AWESoME (Dou and Neubig, 2021) system. AWESoME leverages multilingual pretrained language models and fine-tune them on parallel text. Unsupervised training objectives over the parallel corpus improve the alignment quality of the models. AWESoME authors claim that the model works best with mBERT (Devlin et al., 2019) as backbone, so we follow their advice. Although we also experimented with GIZA++ (Och and Ney, 2003), FastAlign (Dyer et al., 2013) and SimAlign (Dou and Neubig, 2021), systems based on alignments from AWESoME produced the highest F1 scores when comparing the model projections and manually annotated projections (see Section 7).

To train the alignment models we use the English gold-labelled dataset together with the respective MT system translations as parallel corpora. We augment the training data with 50,000 random parallel sentences from ParaCrawl v8 (Esplà et al., 2019) for all the language pairs except Turkish, for which we use 50,000 random parallel sentences from the raw CCAligned v1 corpus (El-Kishky et al., 2020). CCAligned has received some criticism (Kreutzer et al., 2022), but the available English-Turkish parallel data is very limited. In Section 7 we analyze the performance of the alignment systems, and we show that CCAligned does not hurt the performance of the aligners.

5.3 Sequence Labelling Models

We use three state-of-the-art multilingual pre-trained language models for sequence labelling: multilingual BERT (mBERT) (Devlin et al., 2019) and XLM-RoBERTa (XML-R) base and large (Conneau et al., 2020). For both models, we add a token classification layer (linear layer) on top of each token representation. We use the sequence labelling implementation of the Huggingface open-source (Apache-2.0 License) library (Wolf et al., 2019). F1 scores and standard deviation scores are reported by averaging the results of 5 runs with different random seeds. Details on models sizes, hyper-parameters and datasets are provided in the Appendix (A, B and C).

6 Experiments

6.1 Opinion Target Extraction

Opinion Target Extraction (OTE) results are reported in Table 1. The zero-shot model transfer using mBERT obtains better results for Spanish and French. However, for Dutch, Russian and Turkish the best results are obtained by the data-transfer approaches. The overall picture changes when using XLM-RoBERTa (XLM-R) base. First, the zero-shot baseline is much closer to the gold upper bound than that of mBERT. This shows that XLM-R has better multilingual transfer learning capabilities for this task. In fact, the zero-shot transfer outperforms the translate-train and translate-test approaches for all languages except Turkish. Second, the XLM-R base results on gold-labelled data are substantially better than those of mBERT. Finally, XLM-R large offers the best cross-lingual transfer capabilities, as the zero-shot transfer is clearly superior for every language, including Turkish.

A common trait for all three models in the OTE benchmark is that the translate-train approach is superior to the translate-test approach in the large majority of the cases. As expected, all the approaches achieve a performance significantly lower than the gold upper bound.

6.2 Named Entity Recognition

If we compare the OTE results with those obtained for NER (Table 2), we see a number of different patterns. First, the zero-shot approach using mBERT outperforms the data-based cross-lingual transfer methods (translate-train and translate-test) for the majority of languages . Second, unlike in OTE, the translate-test is systematically better than translate-train. Third, the mBERT performance on CoNLL data is similar to that of XLM-R base. Finally, fine-tuning XLM-R base on translated and projected data obtains better results for German and Italian than the zero-shot method. However, XLM-R large provides obtains the same results as for OTE, obtaining the best results for every language in the zero-shot setting. This validates the findings of the OTE results, namely, that the performance of the zero-shot method heavily depends on the characteristics of the multilingual language model used.

Previous research has demonstrated that cross-lingual transfer with mBERT works best for topologically similar languages (Pires et al., 2019; Wu and Dredze, 2020), which is somewhat coherent with the results obtained for Spanish and

	mBERT			
	Gold	Zero-shot	Trans-Train	Trans-Test
EN	76.2 ± 0.9	-	-	-
ES	75.2 ± 0.5	**68.4** ± 0.6	67.9 ± 0.8	62.2 ± 1.2
FR	74.0 ± 1.1	**62.7** ± 1.2	59.7 ± 1.2	57.6 ± 1.1
NL	69.7 ± 0.9	61.7 ± 0.8	64.3 ± 1.5	**67.0** ± 0.8
RU	72.5 ± 0.5	53.8 ± 2.2	**62.9** ± 0.6	59.7 ± 0.4
TR	62.0 ± 1.2	45.3 ± 4.0	**45.7** ± 2.3	35.5 ± 2.4
	XLM-R base			
	Gold	Zero-shot	Trans-Train	Trans-Test
EN	84.4 ± 0.9	-	-	-
ES	81.1 ± 0.7	**78.2** ± 0.4	72.5 ± 0.7	62.9 ± 0.9
FR	80.2 ± 0.6	**72.7** ± 0.3	64.7 ± 0.8	60.0 ± 0.6
NL	80.8 ± 1.7	**75.5** ± 0.8	70.0 ± 1.6	71.0 ± 1.5
RU	81.5 ± 0.3	**74.9** ± 0.9	69.5 ± 0.3	62.2 ± 1.6
TR	69.0 ± 1.1	58.1 ± 3.5	**58.9** ± 1.8	36.4 ± 1.8
	XLM-R large			
	Gold	Zero-shot	Trans-Train	Trans-Test
EN	86.4 ± 1.1	-	-	-
ES	83.6 ± 0.1	**79.3** ± 0.8	73.7 ± 1.1	64.0 ± 1.4
FR	82.2 ± 0.6	**74.6** ± 1.7	66.1 ± 0.6	60.7 ± 0.6
NL	80.4 ± 2.1	**77.7** ± 1.9	74.0 ± 1.3	72.9 ± 1.8
RU	82.8 ± 0.4	**76.8** ± 1.3	69.3 ± 2.3	62.2 ± 1.3
TR	72.3 ± 2.4	**62.4** ± 1.0	57.8 ± 2.4	33.7 ± 0.9

Table 1: OTE F1 score with models of different capacity.

French, where the zero-shot transfer is superior to the Translate-train and Translate-test approaches, while it is worse for Russian and Turkish. Additionally, it is worth noting that mBERT has been trained using only Wikipedia text for 104 languages.

In contrast, XLM-R (both base and large) have been trained using CommonCrawl (Wenzek et al., 2019), a much larger multilingual corpus with a variety of texts extracted from the Web, perhaps also including texts of similar domain to those in the OTE datasets. This may also account for the large differences in OTE performance between XLM-R base and mBERT. In this sense, the similar performance between mBERT and XLM-R base for NER might be partially due to the fact that the CoNLL and Evalita datasets consist of news stories in which most of the labelled entities may appear in the Wikipedia, the texts used to pre-train mBERT.

The performance of the XLM-R large shows that pretrained models with larger capacity help to obtain strong performance across languages, also for zero-shot cross-lingual methods. Still, data-based cross-lingual transfer (Translate-Train and Translate-Test) approaches remain useful if access to the required hardware for working with such larger language models is not available.

Finally, Table 3 lists the results of previous methods that leverage parallel data and/or annotation projections to perform cross-lingual transfer on the NER CoNLL 2002-2003 data. By comparing previous work with our zero-shot baselines using mBERT, XLM-R base and XLM-R large, we can see that the XLM-R large in the zero-shot setting still outperforms most previous approaches. The only exception being the results obtained by Li et al. (2021) for German.

	mBERT			
	Gold	Zero-shot	Trans-Train	Trans-Test
EN	90.7 ± 0.3	-	-	-
ES	87.4 ± 0.4	**74.6** ± 0.4	69.5 ± 0.4	70.8 ± 0.6
DE	82.0 ± 0.4	**71.0** ± 0.9	70.1 ± 0.3	70.6 ± 0.5
NL	90.8 ± 0.4	**78.5** ± 0.5	74.4 ± 0.6	75.4 ± 0.8
IT	84.7 ± 0.3	68.2 ± 0.5	68.7 ± 0.5	**70.7** ± 0.3
	XLM-R base			
	Gold	Zero-shot	Trans-Train	Trans-Test
EN	90.4 ± 0.2	-	-	-
ES	87.7 ± 0.2	**75.0** ± 0.4	70.1 ± 0.6	72.5 ± 0.2
DE	83.1 ± 0.3	67.9 ± 0.5	**70.5** ± 0.5	70.1 ± 0.8
NL	89.8 ± 0.2	**78.1** ± 0.6	73.3 ± 0.9	74.7 ± 0.4
IT	84.3 ± 0.3	71.2 ± 0.5	71.1 ± 0.4	**71.7** ± 0.3
	XLM-R large			
	Gold	Zero-shot	Trans-Train	Trans-Test
EN	92.4 ± 0.1	-	-	-
ES	88.9 ± 0.2	**79.5** ± 1.0	70.9 ± 0.6	74.0 ± 0.5
DE	85.1 ± 0.6	**74.5** ± 0.7	73.7 ± 0.5	72.9 ± 0.3
NL	92.9 ± 0.7	**82.3** ± 0.6	77.5 ± 0.9	77.2 ± 0.6
IT	87.5 ± 0.2	**76.0** ± 0.5	73.7 ± 0.4	73.5 ± 0.6

Table 2: NER F1 score with models of different capacity.

Models	ES	DE	NL
mBERT (Dou and Neubig, 2021)	64.3	-	-
BiLSTM + CRF (Jain et al., 2019)	73.5	61.5	69.9
BiLSTM + CRF (Guo and Roth, 2021)	77.9	71.4	80.6
XLM-R large (Li et al., 2021)	78.9	**76.9**	79.7
mBERT (Ours - zero-shot)	74.6	71.0	78.5
XLM-R base (Ours - zero-shot)	75.0	67.9	78.1
XLM-R large (Ours - zero-shot)	**79.5**	74.5	**82.3**
XLM-R base (Ours - Translate train)	70.1	70.5	73.3
XLM-R base (Ours - Translate test)	72.5	70.1	74.7
XLM-R large (Ours - Translate train)	70.9	73.7	77.5
XLM-R large (Ours - Translate test)	74.0	72.9	77.2

Table 3: Comparison between the previous research methods that leverage projections, the zero-shot baselines and our annotation projections in the 2002-2003 NER CoNLL datasets. F1 score reported

7 Error Analysis

The experiments described in Section 6 showed that translate-train and translate-test perform worse than the zero-shot approach when using XLM-R large. In this section we will assess the performance of the machine translation and word alignment models. Furthermore, we will undertake an error analysis to better understand the shortcomings of translate-train and translate-test with respect to the zero-shot cross-lingual transfer.

7.1 Evaluating the Projection Method

We start our experiments by analyzing the quality of our automatically projected annotations. In order to do that, human annotators manually projected the labels from the English OTE gold-labelled data to the automatic translations to Spanish, French and Russian using DeepL and M2M100 for Turkish. The annotators are NLP PhD candidates with either native and/or proficient skills in both English and the target language. See Section E for more details.

We compare the projections of the annotations automatically generated by the different word alignment methods with those provided by the human annotators. Table 4 shows that the language model-based methods (SimAlign and AWESoME) outperform the statistically based methods (GIZA++ and FastAlign) by a wide margin in all languages. Furthermore, AWESoME consistently outperforms SimAlign for every language. The performance of the AWESoME system confirms that it is possible to generate high quality annotations close to those generated by human experts. The results also show that for Turkish performance is lower than for the other languages. This is the case for the methods that require parallel data (GIZA++, FastAlign and AWESoME) as well as SimALign that does not require parallel data. So we can attribute the lower performance to the difficulty of projecting annotations for the English-Turkish pair and not the usage of the CCAligned corpus.

	GIZA++	FastAlign	SimAlign	AWESoME
ES	77.0	75.0	86.7	**91.5**
FR	73.3	72.9	86.3	**91.3**
RU	72.4	76.9	87.7	**93.4**
TR	64.0	68.4	81.9	**88.5**

Table 4: OTE F1 score between the human annotation projections vs the automatic projections generated using different alignment models.

While Table 4 shows that we generate high quality annotation projections, the best model, AWESoME, still makes some mistakes. To explore the effect of these mistakes we fine-tune XLM-R large models on the manually projected train datasets and compare their performance on the gold-labelled test sets with the models trained on the AWESoME automatically projected data. Table 5 shows that the models obtained using the manually projected data are sightly better, except for Turkish, which once again acts as outlier. In any case, as the results obtained by fine-tuning on the manually projected data are still worse than the zero-shot method, this experiment proves that the projection of annotations is not responsible for the data-based cross-lingual transfer methods to be inferior to the zero-shot baseline.

XLMR	Trans-Train	Trans-Train (Manual)
ES	73.7 ± 1.1	**75.1** ± 1.2
FR	66.1 ± 0.6	**67.9** ± 1.0
RU	69.3 ± 1.3	**69.4** ± 2.1
TR	**57.8** ± 2.4	50.6 ± 1.4

Table 5: XLM-R large OTE F1 score when training with automatically and manually projected datasets

7.2 Downstream Evaluation of Machine Translation Models

In order to evaluate the influence of the machine translation system used, we translate the English gold-labelled data using four different translation systems. We fixed AWESoME as the word aligner for annotation projection. We fine-tune XLM-R large with each of the generated training data and evaluate it against the gold-labelled test data from OTE. As Table 6 shows, there are no big differences in the final F1 scores when using different translation systems (Turkish is again being the exception), we decided to carry on using DeepL based on the manual assessment mentioned in Section 3.

	MarianMT	Mbart	M2M100	DeepL
ES	**75.6** ± 0.8	75.3 ± 0.7	74.2 ± 0.8	73.7 ± 1.1
FR	64.5 ± 1.6	**66.4** ± 1.1	64.9 ± 1.3	66.1 ± 0.6
NL	70.0 ± 2.0	68.8 ± 4.0	70.1 ± 3.1	**74.0** ± 1.3
RU	66.6 ± 4.4	**69.7** ± 1.4	**69.7** ± 0.7	69.3 ± 2.3
TR	49.5 ± 2.9	56.1 ± 5.2	**57.8** ± 2.4	-

Table 6: OTE F1 score of different XLM-R large models trained using data generated with different translation systems.

7.3 Where do the models fail?

To better understand what is happening we identify the most common false negatives and positives for both OTE and NER tasks. Table 7 shows the most frequent false negatives and positives where there is a big mismatch between methods.

As it has been previously noticed (Agerri and Rigau, 2019), in the ABSA data the words "comida" (food) and "restaurante" (restaurant) are highly ambiguous, so we could expect the models to fail with these words. In addition, we have found out 4 main sources of errors, which are analyzed below.

Many-to-one translations: This is stereotypical of targets such as "trato" and "atención" in Spanish, which, in addition to "servicio", are used to refer to "service" in English. There are 160 sentences in the English gold-labelled data containing the word "service"; in 153 of them "service" is labelled as target. DeepL systematically translates it as "servicio". However, as shown by Table 8, in the Spanish gold-labelled data "service" is also commonly referred as "trato" or "atención", instead of "servicio".

This would result in a training set without any occurrences of "trato" and "atención" which often occur in the gold-labelled test data. Both the zero-shot and the data-based cross-lingual transfer approaches fail to correctly label these words, which shows a problem of using automatically translated data. Interestingly, the zero-shot approach using XLM-R large correctly classifies "trato" (only fails to label 1 of the 19 occurrences). As shown by our experimental results, XLM-R large is more robust than mBERT and XLM-R base.

Something similar happens with the word "place", which in Spanish can be most frequently translated as "lugar" or "sitio". However, DeepL almost always translates it as "lugar" which results in "sitio" being absent in the automatically generated training data while being more frequent than "lugar" in the gold-labelled data. Note that this is not a problem for the "translate-test", given that the translation direction is Spanish to English.

Errors induced by incorrect or missing alignments: For NER we found errors of different nature. Articles and prepositions (i.e. "de", "la") are among the words with higher false positive rate for the translate-test and translate-train approaches. We can attribute it to word alignment errors. Large multi-word named entities such as "Consejo General de la Arquitectura Técnica de España" (General Council of Technical Architecture of Spain) are labelled as entities. Word aligners struggle to correctly align articles in these complex expressions specially when a one-to-many or many-to-one alignment is required. In fact, in this example, the word aligners we tested failed to correctly align "of" with "de la".

Errors induced by dataset inconsistencies: Another issue is the differences across languages in the original gold-labelled annotations. Thus, "Gobierno" (Government) and "Estado" (State) are labelled as organizations in the Spanish gold-labelled data, but they are not considered to be entities in the English gold-labelled data. The opposite occurs with demonym words. They are labelled as miscellaneous entities in the English data but in Spanish they are not annotated. Cross-lingual models are likely to fail labelling these cases.

Lost in Translation: Finally, there is another group of words related to Spanish Government names which are not commonly used in English for the same contexts (i.e. "Economía" to refer to the"Ministry of Economy" or "Ministerio de Economía" in Spanish, "Junta" for "local government", or "Plan" for "government projects"). While these words appear frequently in the Spanish data as part of commonly used named entities, that is not the case in the English data, where it is customary to use "Treasury Department" (or variations thereof) which are correctly translated into Spanish by DeepL as "Departamento del Tesoro". This means that, during fine-tuning on the translated data, the model is not receiving any signal to learn that "Economy" may be part of an organization entity. This may explain why the zero-shot method performs better for cases such as "Economía", "Hacienda", "Plan" and "Junta", listed in Table 7.

Summarizing, we see that machine translation data often generates a signal which is, due to inherent differences in language use, different to the signal received when using gold-labelled data in the target language. This disagreement seems to be the most common reason for the larger number of false positive and negatives of the data-based cross-lingual transfer method with respect to the zero-shot technique.

8 Concluding Remarks

In this paper we described an in-depth and comprehensive evaluation of model-based and data-based zero-resource cross-lingual sequence labelling on two different tasks.

Contrary to what previous research suggests, zero-shot transfer approach is the best performing method when using high capacity multilingual language models such as XLM-R large. However, data-based cross-lingual transfer approaches are still useful when having a model with poor downstream cross-lingual performance. For example, when using a pretrained language model not trained for a specific domain, or when the required hardware for working with such larger language models is not readily available.

	GOLD			Zero-shot			Tr-Train			Tr-Test			Total
	B	Xb	Xl	B	Xb	Xl	B	Xb	Xl	B	Xb	Xl	
OTE False Negatives													
comida	3	3	2	6	2	1	4	1	1	1	9	5	98
restaurante	7	5	7	9	5	6	7	6	6	7	12	10	43
servicio	2	2	2	2	1	1	2	0	1	1	1	2	85
trato	1	1	0	5	6	1	14	10	5	6	8	6	19
atención	2	3	3	8	2	3	7	1	3	7	7	7	13
lugar	0	0	0	2	0	0	1	0	0	0	1	0	12
sitio	1	0	0	5	1	1	3	3	3	2	1	1	14
NER False Negatives													
de	32	29	33	45	51	90	233	252	264	148	146	167	450
la	4	5	3	10	12	16	63	62	62	45	44	45	174
Gobierno	0	0	0	17	53	64	72	70	75	30	45	67	80
Estado	0	0	0	4	4	8	9	8	9	6	6	8	10
Administación	0	0	0	4	8	11	10	11	11	5	5	7	11
Economía	0	0	0	2	6	2	7	8	8	5	6	8	8
Plan	0	0	0	1	2	2	3	5	5	1	4	7	8
Junta	0	0	0	0	0	0	4	10	8	2	3	5	24
Hacienda	0	0	0	1	3	0	4	4	4	4	3	4	5
NER False Positives													
español	0	0	0	16	16	2	16	16	12	13	14	15	0
catalán	0	0	0	8	8	5	7	7	8	8	8	8	0

Table 7: Most common false negatives and positives were there is a big mismatch between methods and the total number of labelled apperances of the word in the test data. B is the acronym for mBERT, Xb for XLM-R base and Xl for XLM-R large.

En.Word	Es.Word	En Gold	Es Gold	Es Translate
Service	Servicio	153	229	133
Treatment	Trato	0	54	0
Attention	Atención	2	35	0
Place	Sitio	120	41	2
Place	Lugar	120	19	91

Table 8: Number of times words appear as target words in the train datasets

A detailed error analysis demonstrates that data-based cross-lingual transfer is hindered by machine translations which, although linguistically sound, do not align with the cultural behaviour of the target language use. Moreover, the results also show that the different word alignments methods (for annotation projection) are of high quality, obtaining comparable results with respect to manually generated alignments.

In any case, our results establish that there is still room for improving the cross-lingual performance of zero-resource sequence labelling.

Acknowledgments

We are grateful to Nayla Escribano, Suna Şeyma Uçar and Olia Toporkov for their help in manually projecting the labels from the English gold-labelled data into the automatically translated dataset. We are also thankful to the anonymous reviewers for their insightful comments. This work has been supported by the following projects: (i) DeepKnowledge (PID2021-127777OB-C21) funded by MCIN/AEI/10.13039/501100011033 and FEDER Una manera de hacer Europa; (ii) Disargue (TED2021-130810B-C21), MCIN/AEI/10.13039/501100011033 and European Union NextGenerationEU/PRTR; (iii) DeepR3 (TED2021-130295B-C31) funded by MCIN/AEI/10.13039/501100011033 and EU NextGeneration programme EU/PRTR; (iv) Antidote (PCI2020-120717-2) funded by MCIN/AEI/10.13039/501100011033 and by European Union NextGenerationEU/PRTR. Iker García-Ferrero is supported by a doctoral grant from the Basque Government (PRE_2021_2_0219). Rodrigo Agerri's work is also supported by the RYC-2017-23647 fellowship (MCIN/AEI/10.13039/501100011033 y por El FSE invierte en tu futuro).

Limitations

We compare baseline cross-lingual zero-shot model transfer with machine translation and annotation projection. We do not explore alternative cross-lingual data-based methods, such as the usage of available parallel corpora instead of a machine translated corpus. We also skip evaluating methods to improve model-transfer approaches such as the ones described in Section 2.2. We may also consider that our annotation projection approach and zero-shot model transfer approach work for Indo-European languages, while their performance for other language families remains unknown. Finally, the error analysis was performed for the EN-ES language pair only.

In any case, we believe that our main claim still holds. Even though MT quality has substantially improved over the last few years, our results indicate the current optimal solution to perform cross-lingual transfer is by using large multilingual language models such as XLM-RoBERTa-large. Thus, our error analysis suggests that this might be due to important differences in language use. More specifically, MT often generates a textual signal which is different to what the models are exposed to when using gold standard data, which affects both the fine-tuning and evaluation processes. This is confirmed by our error analysis which shows that mistranslations are not the main source of errors in the data-transfer method.

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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Rodrigo Agerri and German Rigau. 2016. Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence*, 238:63–82.

Rodrigo Agerri and German Rigau. 2019. Language independent sequence labelling for opinion target extraction. *Artificial Intelligence*, 268:85–95.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

M Saiful Bari, Tasnim Mohiuddin, and Shafiq Joty. 2021. Uxla: A robust unsupervised data augmentation framework for zero-resource cross-lingual nlp.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. *CoRR*, abs/2101.08231.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 644–648. The Association for Computational Linguistics.

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, Hissar, Bulgaria. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 5960–5969, Online. Association for Computational Linguistics.

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125.

Hao Fei, Meishan Zhang, and Donghong Ji. 2020. Cross-lingual semantic role labeling with high-quality translated training corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7014–7026, Online. Association for Computational Linguistics.

Ruohao Guo and Dan Roth. 2021. Constrained labeled data generation for low-resource named entity recognition. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4519–4533, Online. Association for Computational Linguistics.

Zechuan Hu, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Risk minimization for zero-shot sequence labeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4909–4920, Online. Association for Computational Linguistics.

Alankar Jain, Bhargavi Paranjape, and Zachary C. Lipton. 2019. Entity projection via machine translation for cross-lingual NER. 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 1083–1092, Hong Kong, China. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

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

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13452–13460. AAAI Press.

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, Vancouver, Canada. Association for Computational Linguistics.

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

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-m: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora.

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.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for*

Computational Linguistics, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Manuela Speranza. 2009. The named entity recognition task at evalita 2009. In *EVALITA 2009*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Mengqiu Wang and Christopher D. Manning. 2014. Cross-lingual projected expectation regularization for weakly supervised learning. *Transactions of the Association for Computational Linguistics*, 2:55–66.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime Carbonell. 2019. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. *arXiv preprint arXiv:1910.04708*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data. *CoRR*, abs/1911.00359.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Qianhui Wu, Zijia Lin, Börje Karlsson, Jian-Guang Lou, and Biqing Huang. 2020. Single-/multi-source cross-lingual NER via teacher-student learning on unlabeled data in target language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6505–6514, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

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.

A Model size

We experiment with multilingual BERT (mBERT) (Devlin et al., 2019) and XLM-RoBERTa (XML-R) base and large (Conneau et al., 2020). We list the number of parameters of each model in Table 9

Model	#params
multilingual BERT	110M
XLM-RoBERTa-base	250M
XLM-RoBERTa-large	560M

Table 9: Number of parameters for the language models that we use in our experiments

B Hyper parameters

B.1 Word alignment models

We train AWESoME with 8 batch size and $2e-5$ learning rate for $40,000$ steps, with all the unsupervised training objectives (mlm,tlm,tlm_full,so,psi) and softmax extraction method. We use mBERT as backbone. For SimAlign we run inference with 0.0 distortion rate, 1.0 null align rate and the "itermax" matching method. We use bpe tokens and mBERT backbone. We use the MGIZA multicore implementation [3] of GIZA++ with the recommended configuration file [4]. We use FastAlign with the default hyper-parameters. For both, GIZA++ and FastAlign we combine the forward and backward directions of the alignments using the grow-diag-final-and algorithm.

B.2 Sequence Labelling models

For OTE we use a batch size of 32, $5e-5$ learning rate, we train the model for 10 epochs and 128 maximum sequence length. Since only a train and test splits are available for the OTE task, we

[3] https://github.com/moses-smt/mgiza

[4] https://pastebin.com/b1ksHtUy

use the train set as both, train and development data. For NER we use a batch size of 32, $2e-5$ learning rate, we train the model for 4 epochs and 256 maximum sequence length. We use the default values (sequence labelling implementation of the Huggingface library [5]) for the remaining hyperparameters. For both tasks we use the BILOU encoding scheme.

C Datasets Size

We list the dataset size (number of sentences) of the datasets we use.

For OTE we use the SemEval-2016 Task 5 Aspect Based Sentiment Analysis (ABSA) datasets (Pontiki et al., 2016). We list the size of the datasets in Table 10.

Lang	Train	Test
EN	2000	676
ES	2070	881
FR	1664	668
NL	1722	575
RU	3655	1209
TR	1232	144

Table 10: Number of sentences in the OTE datasets

For NER we use the Spanish and Dutch data from the CoNLL-2002 datasets (Tjong Kim Sang, 2002). For English and German we use the CoNLL-2003 datasets (Tjong Kim Sang and De Meulder, 2003) and for Italian we use Evalita 2009 data (Speranza, 2009). We list the size of these datasets in Table 11.

	Train	Dev	Test
EN	14987	3466	3684
ES	6871	1914	1516
DE	12705	3068	3160
NL	15806	2895	5195
IT	11227	0	4136

Table 11: Number of sentences in the NER datasets

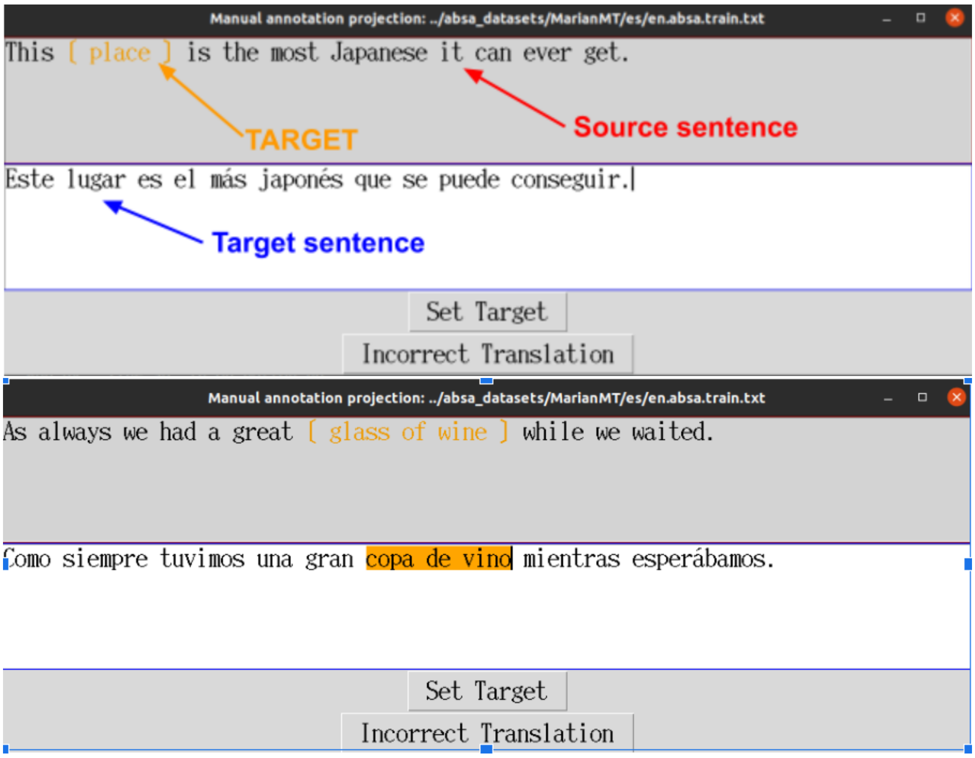
D Computer infrastructure

We perform all our experiments using a single NVIDIA A30 GPU with 24GB memory. The machine used has two Xeon Gold 6226R CPUs and 256GB RAM.

E Manual Projection of the datasets

Human annotators manually projected the labels from the English OTE gold data to the automatic translations to Spanish, French and Russian using DeepL and m2m10 for Turkish The annotators are NLP PhD candidates with either native and/or proficient skills in both English and the target language. We describe the experiment in Section 7.1. For the purpose of this experiment, we developed an application to assist during the annotation process. The annotator sees the sentence in English, where there is a highlighted target and must select the same target in a translated target sentence. Figure 4 shows two screenshots from the application. The full guidelines and the code of the application provided to the annotators are available at https://github.com/ikergarcia1996/Annotation-Projection-App.

[5] https://github.com/huggingface/transformers/tree/main/examples/pytorch/token-classification



Manual annotation projection: ../absa_datasets/MarianMT/es/en.absa.train.txt

This [place] is the most Japanese it can ever get.

Este lugar es el más japonés que se puede conseguir.

Manual annotation projection: ../absa_datasets/MarianMT/es/en.absa.train.txt

As always we had a great [glass of wine] while we waited.

Como siempre tuvimos una gran copa de vino mientras esperábamos.

Figure 4: Application used to manually annotate the projections