SpanAlign: Efficient Sequence Tagging Annotation Projection into Translated Data applied to Cross-Lingual Opinion Mining

Léo Jacqmin[1] **Gabriel Marzinotto**[1] **Justyna Gromada**[2]

Ewelina Szczekocka[2] **Robert Kołodyński**[2] **Antoine Chauvière**[1] **Géraldine Damnati**[1]

(1) Orange Innovation / Lannion, France
(2) Orange Innovation Poland / Warsaw, Poland
first.last@orange.com

Abstract

Following the increasing performance of neural machine translation systems, the paradigm of using automatically translated data for cross-lingual adaptation is now studied in several applicative domains. The capacity to accurately project annotations remains however an issue for sequence tagging tasks where annotation must be projected with correct spans. Additionally, when the task implies noisy user-generated text, the quality of translation and annotation projection can be affected. In this paper we propose to tackle multilingual sequence tagging with a new span alignment method and apply it to opinion target extraction from customer reviews. We show that provided suitable heuristics, translated data with automatic span-level annotation projection can yield improvements both for cross-lingual adaptation compared to zero-shot transfer, and data augmentation compared to a multilingual baseline.

1 Introduction

Large self-supervised pre-trained models which are fine-tuned on downstream tasks have become the de facto standard in NLP. However, monolingual pre-trained models are only available in some high resource languages due to data limitations. Additionally, in multilingual settings, having a separate model for each language quickly becomes unpractical. Multilingual models which are pre-trained on monolingual corpora in multiple languages such as multilingual BERT (mBERT) and XLM-R provide an alternative. They have been shown to yield on par, if not better performance on various tasks compared to monolingual models and perform particularly well for low-resource languages provided similar languages are represented in the training data (Conneau et al., 2020). What's more, they have shown surprising capacities for generalization with zero-shot cross-lingual transfer, in which a model is fine-tuned for a task using annotated data in a specific language and then evaluated in an unseen language (Radford et al., 2019). Considering that labeled data is more readily available for certain languages, this technique of cross-lingual transfer can be harnessed to process low-resource languages, therefore bypassing the need for a costly annotation process. Some research effort has been done in grasping to what extent these models are language agnostic (Pires et al., 2019), though more probing is necessary in order to fully understand and measure how multilingual these models are.

Another approach to cross-lingual transfer consists in using machine translation (MT) to adapt the annotated training data available in a source language to the target language. The resulting translated data is used to train a model with supervision in the target language. Producing a label for translated data is straightforward for sentence-level tasks. However, little work focuses on adapting datasets annotated at the word level. Such adaptation requires an additional step of label projection which can be error-prone and introduce noise in the training data. A common way of projecting labels consists in obtaining word alignments between source and target utterances and then projecting the labels. In a recent paper, (Li et al., 2020) argued that zero-shot cross-lingual transfer surpassed translation-based adaptations for several sequence tagging tasks, and proposed to first warm up the model's weights on the translated data only and then fine-tuning on the original data. In contrast, we show that translation-based adaptation yields superior performance compared to zero-shot cross-lingual transfer, provided the right annotation projection method is used.

We propose to apply this method in the context of opinion mining, and specifically to aspect-based sentiment analysis (ABSA) (Pontiki et al., 2014). Nowadays, people increasingly rely on reviews and comments, e.g. on social media and review websites, to select which products to buy or which services to use. Companies can also make use of these

Proceedings of the 2021 EMNLP Workshop W-NUT: The Seventh Workshop on Noisy User-generated Text, pages 238–248
November 11, 2021. ©2021 Association for Computational Linguistics

data which provide insightful feedback from their customers. ABSA is concerned with extracting fine-grained information from customer feedback, e.g. which aspects of a product or a service are being referred to, which entities are associated with them, and what is the customer's attitude towards them. One challenge that this task poses is that reviews are user-generated and can thus be very noisy. Online users tend to produce text that derives from the standard form of language with e.g. misspellings, internet slang, abbreviations, phonetic transcriptions, and missing or incorrect punctuation marks. For companies operating in several countries, another challenging aspect is the multiplicity of languages found in these reviews. Accordingly, multilingual approaches to opinion mining are essential to efficiently analyze customer feedback across countries.

We first review previous works related to cross-lingual transfer and ABSA. We then present our translation-based adaptation approach, and show its relevance in two scenarios : cross-lingual adaptation where annotated data is only available in the source language, and data augmentation where annotated data is available for both source and target languages. Lastly, we turn to an analysis of noisy user-generated text and propose a heuristic to filter out noisy utterances before translation.

2 Related Work

2.1 Aspect-based Sentiment Analysis

Due to the difficulties associated with processing user-generated text, opinion mining is commonly formulated as a text classification task concerned with e.g. classifying the overall polarity of a sentence by assigning it a polarity label (Pang and Lee, 2008).

Aspect-based sentiment analysis (ABSA) (Hu and Liu, 2004; Popescu and Etzioni, 2005; Pontiki et al., 2014, 2015, 2016) is a more fine-grained opinion mining task with several sub-tasks associated with it. For example, opinion target extraction (OTE) retrieves the entity on which an opinion is expressed in a review, and aspect sentiment classification (ASC) identifies the polarity of an opinion expressed on a given target entity. Some works focus on one subtask at a time, e.g. OTE (Li and Lam, 2017; Xu et al., 2018) or opinion word extraction (OWE) (Fan et al., 2019; Pouran Ben Veyseh et al., 2020). But this division of subtasks is ill-suited for real-world scenarios as both ASC and OWE assume that the opinion target is given. Moreover, these subtasks aim to extract related information. To facilitate practical applications of ABSA, recent works address OTE and ASC simultaneously (Li et al., 2019a,b). The problem can be formulated as a sequence tagging task with unified labels to simultaneously detect opinion targets and the corresponding aspect sentiments. Some works went even further by addressing OTE, ASC, and OWE simultaneously, a task dubbed as aspect sentiment triplet extraction (Peng et al., 2020; Wang et al., 2021).

Most works in the state of the art focus on SemEval datasets from 2014 to 2016, with data from the restaurant, hotel or laptop domain in English. However, SemEval data are also available in several languages. Jebbara and Cimiano (2019), for instance, evaluated CNN models for OTE with multilingual word embeddings in a zero-shot cross-lingual framework.

2.2 Cross-lingual transfer

Multilingual pre-trained models have shown strong capacities for generalization and have been successfully applied for zero-shot cross-lingual transfer on a variety of natural language understanding tasks (Wu and Dredze, 2019). This enables the application of such models to low-resource languages for which little or no labelled data is available.

Machine Translation (MT) can help learning cross-lingual representations and transfering information across languages. Rather than being at odds, zero-shot cross-lingual transfer and MT-based approaches are complementary : MT can be used to generate synthetic data in languages for which no annotated data is available. Accordingly, MT can improve cross-lingual transfer in two ways : (i) By translating the training data into the target languages and fine-tuning on all languages, e.g. for subjectivity analysis (Banea et al., 2008), sentiment classification (Duh et al., 2011) or semantic role labeling (Fei et al., 2020). (ii) Or by applying a model fine-tuned on the source language to a test set translated from target to source language. (Conneau et al., 2020) refers to the former approach as *translate-train* and to the latter as *translate-test*. In this work, we focus on the former approach. This approach naturally applies to sentence classification tasks and previous work has shown that translation-based adaptation is superior to zero-shot cross-lingual transfer for text classification (Schwenk and Li, 2018) and text pair classification (Conneau et al., 2018; Yang et al., 2019). However, little work focuses on adapting datasets annotated at the word level for sequence tagging. This requires an important additional step of annotation projection as no word-to-word correspondence is available.

One approach to annotation projection for sequence tagging tasks relies on obtaining word alignments to project labels from source to target utterances. Several statistical word alignment tools are available such as `fast_align` (Dyer et al., 2013) which constitutes the usual baseline for this approach. Starting from the assumption that neural machine translation (NMT) models capture word alignment through their attention mechanism, other works (Chen et al., 2020; Zouhar and Pylypenko, 2021) focus on using attention weights for word alignment. In an attempt to alleviate the need for parallel corpora, (Jalili Sabet et al., 2020) leveraged modern multilingual pre-trained models and released `SimAlign`, a tool for unsupervised word alignment based on the similarity of multilingual word representations.

In order to project sequence annotations, one has to leverage word-level alignment towards span alignment. Marzinotto (2020) used attention-based word alignments to project FrameNet annotations (targets and Frame Elements) into a target language. Others have sought to improve label projection through different approaches. Jain et al. (2019) proposed an entity projection method for named entity recognition, in which they obtain potential translations of an entity in the target language and select the best match with the source entity. To make use of task-related information, Xu et al. (2020) proposed an end-to-end model to jointly align and predict target slot labels for cross-lingual NLU. (Li et al., 2020) propose an approach called *span-to-span mapping* which derives span alignment from word alignments. They applied it to several tasks (Opinion Mining, Semantic Role Labeling, Named Entities) but didn't obtain satisfactory results when directly using the translated and aligned data to fine-tune their models.

3 SpanAlign

In the case of sentence-level tasks, the translated data can be used right away using the same labels as in the source language corpus. In contrast, adapting sequence tagging datasets annotated at the word-level requires an additional step of annotation projection. A word-to-word correspondence between source and target utterances can be obtained using word alignment tools. One simple approach to annotation projection is then to project the gold labels in the source utterance to their corresponding word in the target utterance using this mapping. However, as token-level annotations may span several words, this approach can be error-prone, as illustrated in Figure 1. Consequently, we introduce a span-based alignment approach to project gold annotations at the span level. Additionally we make the assumption that translated data with erroneous annotation alignment are more likely to be harmful for the subsequent training process, and propose to introduce a constraint on span alignment to filter out translated annotations and remove them from the candidate pool of translated training data.

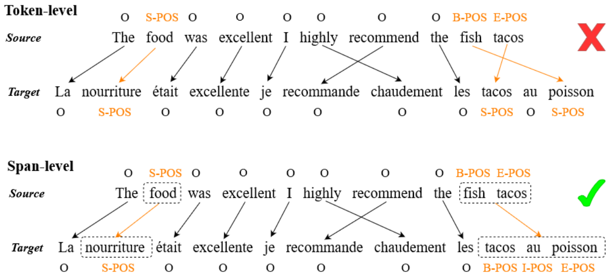
Token-level
O S-POS O O O O O O B-POS E-POS
Source
The food was excellent I highly recommend the fish tacos
✗
Target
La nourriture était excellente je recommande chaudement les tacos au poisson
O S-POS O O O O O O S-POS O S-POS
Span-level
O S-POS O O O O O O B-POS E-POS
Source
The food was excellent I highly recommend the fish tacos
✓
Target
La nourriture était excellente je recommande chaudement les tacos au poisson
O S-POS O O O O O O B-POS I-POS E-POS

FIGURE 1: Different strategies for annotation projection using word alignments.

3.1 Annotation Projection for Sequence Tagging

Let $x_{1:N}^{src} = \{x_1, ..., x_N\}$ be the source utterance and $x_{1:M}^{trg} = \{x_1, ..., x_M\}$ the translated utterance. We obtain word alignments $a_{1:N}^{src} = \{a_1, ..., a_N\}$ from these paired utterances, with $a_i \in [1, M] \cup \{\text{NULL}\}$ indicating which words in the target utterance correspond to source word x_i^{src}. For each annotated span $x_{i:j}^{src}$ in the source utterance from the i-th source word to the j-th source word, we identify the projected span in the target utterance as $x_{p:q}^{trg}$, where

$$\begin{aligned} p &= min(a_{ij}^{src}) \\ q &= max(a_{ij}^{src}) \end{aligned} \tag{1}$$

This procedure is described in pseudo-code in Algorithm 1. We then assign the label without position, e.g. `POS`, to all words in the projected span and reformat the position labels, e.g. `B-{POS}`, `E-{POS}`.

Algorithm 1: Span-based alignment

```
function align ($src\_utt, trg\_utt, src\_spans, \alpha$)
$a^{src}_{1:N}$ = get_word_alignment($src\_utt, trg\_utt$)
$trg\_spans$ = []
for $x^{src} \in src\_spans$ do
    $x^{trg}$ = []
    for $i \in x^{src}$ do
        if $i \in a^{src}_{1:N}$ then
            $x^{trg}$.append($a^{src}_i$)
        end
    end
    $x^{trg}$ = sorted($x^{trg}$)
    $largest\_gap$ = get_largest_gap($x^{trg}$)
    if $largest\_gap > \alpha$ then
        return $False$
    else
        $x^{trg}$ = range(min($x^{trg}$), max($x^{trg}$))
        $trg\_spans$.append($x^{trg}$)
    end
end
return $trg\_spans$
```

3.2 Filtering Ill-formed Projected Spans

The underlying word alignments are not guaranteed to be entirely accurate. First, the translations are machine-generated and can therefore be inadequate. Second, user-generated text is noisy. This noise in source utterances can negatively impact the translations and let errors propagate through the rest of the pipeline. To address these issues, we use a heuristic to filter out pseudo-labeled utterances that are likely to be ill-formed. See Figure 2 for an example.

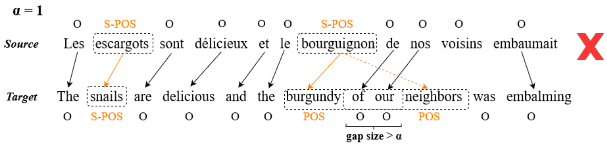
α = 1
O S-POS O O O O S-POS O O O O
Source
Les escargots sont délicieux et le bourguignon de nos voisins embaumait
Target
The snails are delicious and the burgundy of our neighbors was embalming
O S-POS O O O O POS O O POS O O
gap size > α

FIGURE 2: In this case, "bourguignon" which is a dish but also a gentilic was aligned with both "burgundy" and "neighbours", introducing a gap of two words in the proposed translation.

Since our projection method uses the minimum and maximum indices of the aligned words as the offset of the projected span, some gaps can appear in the projected spans where some of the translated words have not been aligned to an opinion target word in the source utterance. We call these words insertions. The hyperparameter α corresponds to the number of allowed insertions within a gap in the projected spans. Utterances that contain a projected span with a gap containing a number of insertions strictly higher than α are considered ill-formed and are filtered out of the translated dataset. Note that this heuristic is well-suited for the task of opinion target extraction because opinion targets are mostly adjacent nominal groups that usually translate into adjacent groups. It should also be relevant for tasks such as Named Entity detection. We are aware that some grammatical structures can translate into non adjacent terms in some languages but we believe that our heuristic approach is reasonable for entity detection tasks.

4 Translation-based Adaptation

Modern neural machine translation (NMT) provides satisfactory results thanks to Transformer models and has become increasingly available. As a result, such models can be used effectively to create synthetic data by translating a source language reference corpus.

We use Marian NMT [1] to translate the source language corpora into the other languages, and vice-versa. Marian NMT is open source and allows the use of pre-trained NMT models from the OPUS-MT project (Tiedemann and Thottingal, 2020) available on this framework. A description of each model and its evaluation on benchmarks can be found online. [2] To obtain word alignments, we use both off-the-shelf `fast_align` (Dyer et al., 2013) and `SimAlign` (Jalili Sabet et al., 2020). For the latter, which doesn't need any parallel data for training but relies on multilingual word representations, we use the Itermax variant, which was shown to outperform other word alignment approaches. Note that we did not adapt the translation models nor the alignment approaches to our specific domains.

4.1 Cross-lingual Adaptation

Cross-lingual adaptation denotes the case where annotated data is available for one language only and we wish to process other languages. The question here is whether the original annotated data is sufficient to fine-tune a pre-trained multilingual model and process unseen languages, or if translating this data into the other languages facilitates

1. https://marian-nmt.github.io
2. https://opus.nlpl.eu/Opus-MT/

cross-lingual transfer. In other words, whether multilingual word representations are truly language agnostic or not. We experiment with the following experiments. **O** and **Tr** respectively denote *original* and *translated* data. **S** and **T** refer to *source* and *target* languages.

O$_S$ **(Zero-shot)** In the context of cross-lingual transfer, zero-shot learning consists in fine-tuning a model for a given task using annotated data in a source language, and then evaluating it for the same task on a test set from an unseen target language. This configuration serves as the baseline method for cross-lingual adaptation.

Tr$_{S\to T}$ The first alternative configuration we consider is to simply adapt the available annotated data to the target language. For a given target language, we fine-tune the model on the translation of the source language data into the target language only.

O$_S$ **+ Tr**$_{S\to T}$ As a step up from the previous configuration, we fine-tune a model on the concatenation of the source data and its translation into the target language.

O$_S$ **+ Tr**$_{S\to all}$ In this last configuration, we use the concatenation of the source data with its translations into all the other languages. This approach results in a single model that can be applied to all languages.

	Original	**Translated**	
Config.	**Source**	**Target**	**Others**
O$_S$ (Zero-shot)	✓		
Tr$_{S\to T}$		✓	
O$_S$ + **Tr**$_{S\to T}$	✓	✓	
O$_S$ + **Tr**$_{S\to all}$	✓	✓	✓

TABLE 1: Data configurations for cross-lingual adaptation.

4.2 Data Augmentation

Data augmentation refers to the case where small amounts of annotated data is available in all the languages we wish to process. We are interested in studying how additional synthetic data will affect the model's performance. We experiment with the following configurations :[3]

O$_T$ **(Monolingual)** As a baseline configuration, we simply fine-tune a model on the original data for a given language and evaluate it on that language.

O$_{all}$ **(Multilingual)** The next configuration uses the concatenation of the original data available in all the languages.

O$_{all}$ **+ Tr**$_{S\to all}$ We fine-tune a model on the concatenation of the original data in all languages along with the translated data from the source language into the other languages. With this configuration, we evaluate the relevance of our approach for data augmentation in a multilingual setting.

O$_{all}$ **+ Tr**$_{all\to S}$ We also experiment with translating the corpora from all the other languages into the source language to understand how the direction of the translation affects our approach and if our assumptions about translation direction are valid.

O$_{all}$ **+ Tr**$_{S\leftrightarrow all}$ This final configuration takes a concatenation of the original data in all languages with all the translated data obtained from both translation directions.

	Original		**Translated**	
Config.	**Target**	**Others**	**S → all**	**all → S**
O$_T$ (Monolingual)	✓			
O$_{all}$ (Multilingual)	✓	✓		
O$_{all}$ + **Tr**$_{S\to all}$	✓	✓	✓	
O$_{all}$ + **Tr**$_{all\to S}$	✓	✓		✓
O$_{all}$ + **Tr**$_{S\leftrightarrow all}$	✓	✓	✓	✓

TABLE 2: Data configurations for data augmentation.

5 Experiments

In this section, we describe the experiments that we conducted to explore the relevance of translation-based adaptation, both for cross-lingual adaptation, where annotated data is available in only one language, and for data augmentation, where multilingual datasets are available and we seek to improve performance from fine-tuning on the plain concatenation of data in all languages.

5.1 Aspect-based Sentiment Analysis

Aspect-based sentiment analysis (ABSA) may refer to several subtasks. We focus on joint opinion target extraction (OTE) and aspect sentiment classification (ASC). Following (Li et al., 2019a), we use a unified tagset with a single sequence tagging

[3] Note that not all possible configurations were reported for the sake of simplicity.

Language	SemEval				In-house					
	Train		Test		Train		Dev		Test	
	# review	# OT	# review	# OT	# review	# OT	# review	# OT	# review	# OT
ar	4802	9612	1227	2371	592	741	198	280	198	237
en	2000	1743	676	612	600	780	200	279	200	269
es	2070	1859	881	713	1200	1080	200	354	200	372
fr	1664	1641	668	650	2047	1984	131	84	131	84
nl	1722	1859	575	713	594	554	198	183	198	190
pl	–	–	–	–	864	1827	200	610	200	581
ro	–	–	–	–	600	1050	200	358	200	357
ru	3665	3078	1209	952	–	–	–	–	–	–

TABLE 3: Number of reviews for each language and corresponding number of Opinion Targets (OT).

model to jointly extract opinion targets and their associated sentiment label.

5.2 Data

We conducted experiments on common benchmark datasets for ABSA, as well as on in-house datasets annotated in the same way.

SemEval-2016 Task 5 (Pontiki et al., 2016) This shared task provides multilingual benchmark datasets for ABSA. Online reviews from several domains have been annotated for various languages. Datasets for OTE are available in English, Spanish, French, Dutch, Russian, Arabic, and Turkish.[4] The annotations include opinion targets on which the opinion is voiced, along with their associated sentiment and the aspect category they belong to. The Arabic dataset pertains to the domain of hotel reviews, while the datasets in all the other languages contain restaurant reviews.

Orange In-house datasets To provide a different view of the same task, we experiment with in-house datasets annotated in a similar fashion. These datasets contain customer reviews for mobile applications. As such they contain noisy text which pose an additional challenge. The mobile applications are related to the activities of the Orange telco operator. Orange being present in several countries, we were able to gather and annotate reviews related to similar apps deployed in several countries. Our corpus covers French, English, Spanish, Arabic,[5] Dutch, Polish, and Romanian. Examples of reviews in English are shown in Table 4.

Cool app, I really like the ux design. Keep up !
Reallu useful app to know ur credit, internet consumption... Etc thnx
nice app all in hand to discover your account
Now it work ! Except Invoice.....Maybe because it's the first one.....For the rest no bad.

TABLE 4: Examples of reviews from the English in-house dataset. Positive and negative opinion targets are shown in blue and red respectively.

Besides the difference in terms of domain, SemEval and in-house datasets differ in the level of segmentation. For SemEval, reviews were segmented into sentences and an input for the model is a single sentence, while for our dataset we kept reviews as a whole (sentence segmentation is not always present nor reliable and we considered it was more related to a real usage). As a consequence, inputs are longer for in-house dataset (for instance 22.8 words on average for English vs 14.1 for English SemEval or 34.8 words on average for Spanish vs 19.6 for Spanish SemEval). Moreover, the mobile applications reviews collected for the in-house datasets are directly typed through a mobile smartphone. As a result, data are more noisy in terms of typographic errors (see for instance the second example in Table 4).

5.3 Source and Target Languages

For the SemEval datasets, we translate the English corpus into all the other languages. Similarly, we use French as the source language for the in-house data. The choice of which language to translate from is based on the following considera-

[4] We do not include the Turkish data in our experiments as no English to Turkish MT model is available within the NMT framework we use.

[5] The application is deployed in several arabo-speaking countries, and the Arabic dataset includes standard Arabic along with 4 dialects (from Morocco, Tunisia, Jordan and Egypt). We are aware that this is a very rough approximation to launch a translation process independently of the dialect but we will try in future work to have more accurate translation systems for the corresponding dialects.

Word alignments		ar	en	es	fr	nl	pl	ro	avg.	% out
`fast_align`	**no filtering**	24.9	38.7	31.7	50.6	53.8	22.0	19.8	34.5	0
	$\alpha = 2$	28.0	41.8	37.8	54.1	55.7	31.0	28.6	39.6	12.9
	$\alpha = 1$	27.9	41.2	35.1	52.5	**56.1**	30.1	26.1	38.4	13.3
	$\alpha = 0$	28.9	40.9	34.3	52.7	54.9	29.3	25.6	38.1	15.0
`SimAlign`	**no filtering**	22.8	31.4	29.5	39.1	37.9	26.2	23.3	30.0	0
	$\alpha = 2$	32.3	43.1	**38.8**	52.4	51.1	36.9	**32.1**	41.0	4.8
	$\alpha = 1$	31.9	42.4	38.7	53.2	54.5	**37.7**	31.5	**41.4**	5.3
	$\alpha = 0$	**33.3**	**43.4**	38.7	**54.5**	49.8	35.8	30.9	40.9	11.4

TABLE 5: Study of the impact of constraining the number of insertions within projected opinion targets with span-based alignment. Results for the $\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow all}$ configuration on the in-house Test datasets.

tions : (i) We expect pre-trained translation models for high-resource languages such as English and French to be more accurate than for languages with fewer resources. (ii) We consider that annotations for English or French data in either case tend to be of higher quality as they will go through a more extensive review process. In the case of the in-house datasets, French data is also more represented.

5.4 Experimental Settings

We use mBERT as our multilingual language model initialized with pre-trained weights[6] available from HuggingFace Transformers (Wolf et al., 2020). For each experiment, a model is fine-tuned with a linear layer for sequence tagging added on top of mBERT's architecture. We train each model for a maximum of 50 epochs using the Adam optimizer (Kingma and Ba, 2015) with a learning rate set to $5e^{-5}$. We use the `BIOES` encoding as our tagging scheme. The evaluation is performed on exact matches of span boundaries and labels. Results are averaged over three runs and are reported using the micro F_1 score. In this study, we did not seek to use the most accurate model for the task and voluntarily chose a simple architecture in order to focus on training data selection and preparation.

5.5 Annotation Projection

To establish what is the optimal value for α on our task, we conduct a preliminary extrinsic evaluation. Using the translations of the French in-house dataset in the other languages, we create different collections of synthetic datasets by projecting the annotations with different values for α. We also evaluate our projection method with no restriction on the number of insertions to assess the relevance of the filtering approach. We fine-tune these models in the cross-lingual adaptation scenario with the $\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow all}$ configuration. We also compare `SimAlign` with `fast_align`. The results are shown in Table 5. We observe a clear gain with our filtering method compared to allowing any insertion. We also obtain better results with `SimAlign`. Removing data where a gap is observed filters out more data when using `fast_align`. Similar experiments on SemEval not reported here yielded similar conclusions. Based on these results, `SimAlign` is used in all the following experiments and α is set to 1.

5.6 Cross-lingual Adaptation

Results are shown in Table 6. We compare our method with the work of Li et al. (2020), even though we were not able to reproduce their results with **Zero-shot**. They use `fast_align` and similar span alignment but without any constraints on allowed gaps. Hence our projection approach differs from theirs in the alignment method (`SimAlign` vs. `fast_align`) and in the filtering process that we introduced. And their training approach consists in warming up the model's parameters using the translated data and then fine-tuning on the source language data only.

Overall, performances vary across languages, with lower results for SemEval in particular in the **Zero-shot** configuration for distant languages (Russian and Arabic). Translation-based adaptation configurations, namely $\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow T}$ and $\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow all}$, provide a significant performance improvement compared to **Zero-shot**. In some cases, $\mathbf{Tr}_{S\rightarrow T}$ performs better than **Zero-shot**, but there is a clear benefit to using the original source language data as well. Results are particularly low for Arabic since the dataset pertains to another domain (restaurant vs. hotel reviews). For SemEval, the gains observed for Arabic despite the domain shift demonstrate the usefulness of fine-tuning on synthetic data in the target language even if the source

[6] `bert-base-multilingual-cased`

language data is from a different domain.

In most cases, $\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow T}$ and $\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow all}$ models come relatively close. However, $\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow all}$ generally outperforms the other configurations. Moreover, it can be considered as superior as a unique model can be applied to all languages which is particularly relevant from an operational point of view, with only one model to maintain. While the results between the approach of Li et al. (2020) and ours are similar, our approach has the advantage of being conceptually simpler and easier to implement.

We also observe that with $\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow all}$, the translated data contribute to improved performances for the source language, i.e. English or French, highlighting the relevance of the translation-based adaptation method for data augmentation as well.

5.7 Data Augmentation

Results for data augmentation are shown in Table 7. When comparing the two non-augmented baselines, i.e. **Monolingual** and **Multilingual**, we observe a significant improvement when fine-tuning a model on the combination of all languages (on average, +11.7 points for SemEval and +16.6 points on the in-house dataset), highlighting the ability of mBERT for cross-lingual transfer learning.

For SemEval, augmenting the training corpus with translated data consistently provides an improvement over the $\mathbf{O}_{all}$ baseline. Using the translations of the non-English corpora into English ($\mathbf{O}_{all}$ + $\mathbf{Tr}_{all\rightarrow S}$) is detrimental to the performance on the English test set, while other languages are not impacted as much by this translation direction. The reason could be that, similarly to the cross-lingual adaptation experiments, it is beneficial to use the translated data in the target language specifically. Overall, $\mathbf{O}_{all}$ + $\mathbf{Tr}_{S\rightarrow all}$ seems to be the most effective configuration for all languages.

Regarding the in-house datasets, the synthetic data is not as beneficial as in the case of SemEval. Results are comparable to $\mathbf{O}_{all}$ for all translation directions, and no data configuration stands out as most effective on average.

6 Analysis

In line with the results from the previous section, we conduct a refined analysis to shed some light on the impact of noisy text.

When applied to the SemEval datasets, data augmentation in the resource-rich setting is beneficial and yields a significant improvement in performance, especially for English. For the in-house datasets however, data augmentation does not affect the performances as much compared to the multilingual baseline. Our hypothesis is that the in-house datasets are more noisy and contain a more specific lexicon, which is what could be perturbing the translations, and thus negatively impacting the rest of the pipeline.

As we did not have gold data to explicitly evaluate the orthographic or grammatical deviation rate, nor the translation quality, nor the alignment quality, we propose an approximation to quantify noise and compare SemEval and in-house datasets. We compute the out-of-dictionary (OOD) rate using reference dictionaries recommended with Hunspell[7] and available from Firefox for each language.[8] The OOD rate is computed as follows :

$$\text{OOD rate} = \frac{O}{N} \times 100 \qquad (2)$$

where O is the number of OOD words and N the number of tokens in the corpus. Table 8 provides the OOD rate for the training sets of each language that is common to both SemEval and in-house data. We observe that the OOD rates are around twice as

[7] http://hunspell.github.io/

[8] https://addons.mozilla.org/en-US/firefox/language-tools/

	SemEval						In-house						
Config.	**en**	**ar**	**es**	**fr**	**nl**	**ru**	**fr**	**ar**	**en**	**es**	**nl**	**pl**	**ro**
$\mathbf{O}_S$ (Zero-shot)	62.7	13.2	51.2	35.1	38.2	34.9	48.6	13.5	23.9	24.8	33.8	19.3	15.4
$\mathbf{Tr}_{S\rightarrow T}$	_	7.6	25.8	37.6	24.0	29.1	_	24.5	30.3	31.5	38.8	31.9	27.2
$\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow T}$	_	31.4	54.3	40.7	49.4	47.7	_	30.0	42.6	38.7	**55.6**	34.2	29.9
$\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow all}$	**64.1**	**34.7**	54.8	41.49	**51.0**	**47.9**	**51.4**	**31.3**	**43.3**	**39.4**	55.5	**35.6**	**31.9**
(Li et al., 2020)	_	_	**58.2**	**46.9**	49.9	44.9	_	_	_	_	_	_	_

TABLE 6: Results for cross-lingual adaptation, i.e. annotated data is available for the source language only.

	SemEval							In-house							
Config.	**ar**	**en**	**es**	**fr**	**nl**	**ru**	**avg.**	**ar**	**en**	**es**	**fr**	**nl**	**pl**	**ro**	**avg.**
$\mathbf{O}_{T}$ (Monolingual)	61.1	62.7	58.9	34.4	52.3	35.5	50.8	10.3	42.6	38.7	48.6	55.6	34.2	29.9	37.1
$\mathbf{O}_{all}$ (Multilingual)	**61.7**	63.0	67.0	62.1	59.3	61.6	62.5	48.0	54.0	54.8	**56.8**	64.5	**56.4**	41.3	**53.7**
$\mathbf{O}_{all}$ + $\mathbf{Tr}_{S\rightarrow all}$	**61.7**	**66.8**	67.7	61.3	62.0	**62.2**	**63.6**	**48.5**	53.2	56.0	55.9	61.6	54.5	41.0	53.0
$\mathbf{O}_{all}$ + $\mathbf{Tr}_{all\rightarrow S}$	**61.7**	56.6	**68.8**	61.8	59.5	61.3	61.6	46.9	54.1	**56.2**	54.1	64.1	55.3	40.8	53.1
$\mathbf{O}_{all}$ + $\mathbf{Tr}_{S\leftrightarrow all}$	61.6	58.7	67.5	**62.3**	**62.1**	61.7	62.3	48.4	**54.9**	54.2	55.2	**64.6**	54.6	**42.0**	53.4

TABLE 7: Results for data augmentation, i.e. annotated data is available for all languages.

high for our in-house datasets compared to SemEval datasets.[9]

	fr	**en**	**nl**	**es**	**ar**	**avg.**
SemEval	1.4	2.5	2.1	4.0	12.1	4.4
In-house	4.3	3.7	4.1	7.0	25.6	8.9

TABLE 8: Average OOD rates in the training sets.

Considering the OOD rate as an indicator of noise in text, we can filter out the most noisy utterances from the source dataset in French before translating it into the other languages. We compute the OOD rate for each utterance and rank them from most noisy to least noisy. The top n percent is then filtered out from the dataset before translation. We experiment with several values for n. Results can be seen in Table 9. We observe a slight increase in performance when filtering out the 5% most noisy utterances compared to not filtering. Beyond that, filtering a larger portion of utterances seems to be reducing the size of the dataset too much and degrades performance.

Filter	**ar**	**en**	**es**	**nl**	**pl**	**ro**	**avg.**
None	**48.5**	53.2	**56.0**	61.6	54.5	41.0	52.5
5%	47.4	**55.3**	55.6	**64.0**	54.7	**41.7**	**53.1**
10%	46.8	53.2	54.5	63.3	**54.8**	40.9	52.2
20%	47.4	54.5	52.6	63.8	53.1	38.3	51.6

TABLE 9: Filtering out the n percent utterances with the largest OOD rate from the Source corpus before translating it into the other languages, impact on Target languages (data augmentation on the in-house dataset with the $\mathbf{O}_{all}$ + $\mathbf{Tr}_{S\rightarrow all}$ configuration)

.

7 Conclusion

We have proposed an efficient yet simple way of generating training data for a sequence tagging task based on translation and label projection. When applied to noisy data such as customer reviews, we propose a way to overcome potentially ill-formed projections by filtering out some translated data thanks to a heuristic. Our `SpanAlign` algorithm, in conjunction with the `SimAlign` word alignment approach, yields interesting results for multilingual opinion mining both in *cross-lingual* configuration where no annotated data is available in the target languages, and in *data augmentation* configuration where annotated data are available in the target languages. We show that it is possible to train a single model for several languages, which is important in an industrial setting for maintenance issues. Finally, a preliminary study shows that selecting data prior to the translation and projection process on the basis of an Out of Dictionary rate increases the process robustness. Future work will consist in characterizing more precisely the level of noise in data and potentially correcting data in order to improve the selection process.

References

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 127–135, Honolulu, Hawaii. Association for Computational Linguistics.

Yun Chen, Yang Liu, Guanhua Chen, Xin Jiang, and Qun Liu. 2020. Accurate word alignment induction from neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 566–576, Online. Association for 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.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli : Evaluating cross-lingual sentence representations. In

[9] Note that the higher average OOD rate in Arabic can be attributed to dialectal variations.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Kevin Duh, Akinori Fujino, and Masaaki Nagata. 2011. Is machine translation ripe for cross-lingual sentiment classification? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, pages 429–433, Portland, Oregon, USA. Association for Computational Linguistics.

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

Zhifang Fan, Zhen Wu, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2019. Target-oriented opinion words extraction with target-fused neural sequence labeling. 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)*. Association for Computational Linguistics.

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.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 1083–1092. Association for Computational Linguistics.

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 *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : Findings*, pages 1627–1643, Online. Association for Computational Linguistics.

Soufian Jebbara and Philipp Cimiano. 2019. Zero-shot cross-lingual opinion target extraction. 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 2486–2495, Minneapolis, Minnesota. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

X. Li, Lidong Bing, Wenxuan Zhang, Zheng Li, and Wai Lam. 2020. Unsupervised cross-lingual adaptation for sequence tagging and beyond. *ArXiv*, abs/2010.12405.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019a. A unified model for opinion target extraction and target sentiment prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33 :6714–6721.

Xin Li, Lidong Bing, Wenxuan Zhang, and W. Lam. 2019b. Exploiting bert for end-to-end aspect-based sentiment analysis. *ArXiv*, abs/1910.00883.

Xin Li and Wai Lam. 2017. Deep multi-task learning for aspect term extraction with memory interaction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2886–2892, Copenhagen, Denmark. Association for Computational Linguistics.

Gabriel Marzinotto. 2020. FrameNet annotations alignment using attention-based machine translation. In *Proceedings of the International FrameNet Workshop 2020 : Towards a Global, Multilingual FrameNet*, pages 41–47, Marseille, France. European Language Resources Association.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1–2) :1–135.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05) :8600–8607.

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.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages

486–495, Denver, Colorado. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, page 339–346, USA. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Nasim Nouri, Franck Dernoncourt, Dejing Dou, and Thien Huu Nguyen. 2020. Introducing syntactic structures into target opinion word extraction with deep learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8947–8956, Online. Association for Computational Linguistics.

Alec Radford, Jeff Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. In *OpenAI blog*.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Fang Wang, Yuncong Li, Wenjun Zhang, and Shengtao Zhong. 2021. A more fine-grained aspect-sentiment-opinion triplet extraction task. *ArXiv*, abs/2103.15255.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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 833–844, Hong Kong, China. Association for Computational Linguistics.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2018. Double embeddings and cnn-based sequence labeling for aspect extraction. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063, Online. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X : A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692. Association for Computational Linguistics.

Vilém Zouhar and Daria Pylypenko. 2021. Leveraging neural machine translation for word alignment.