UNIVERSITÉ DE LORRAINE

MASTER'S THESIS

Multilingual and Multi-domain Opinion Mining

Author:
Léo JACQMIN

Supervisor:
Gabriel MARZINOTTO

Examiners:
Irina ILLINA
Miguel COUCEIRO
Mathieu CONSTANT
Kamel SMAÏLI
Karën FORT

A thesis submitted in fulfillment of the requirements for the degree of
MSc Natural Language Processing

Based on work done during an internship with the

Deskiñ team
Orange Labs
1 March - 31 August 2021

January 9, 2023

Declaration of Authorship

I, Léo JACQMIN, declare that this thesis titled, "Multilingual and Multi-domain Opinion Mining" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

UNIVERSITÉ DE LORRAINE

Abstract

Institut des sciences du Digital, Management & Cognition
Orange Labs

MSc Natural Language Processing

Multilingual and Multi-domain Opinion Mining

by Léo JACQMIN

Everyday, multinational companies receive large amounts of customer feedback. Valuable insights can be extracted from it to address customers' needs. This textual feedback can span various languages and domains. Due to this variability, generic opinion mining approaches that make best use of annotated data are crucial. A particularly interesting task is aspect-based sentiment analysis (ABSA), which aims to extract fine-grained information about specific entities and their aspects. We show how different transfer learning techniques can be combined to adapt a pre-trained model (PTM) and produce a single ABSA model that is capable of addressing various domains, tasks, and languages at once. Large multilingual PTMs have shown capacities for zero-shot cross-lingual transfer. We found that this approach was in fact limited and that these models benefited from being fine-tuned on the target language specifically. To facilitate such approaches, we introduce a translation-based adaptation method to generate synthetic data in a target language and facilitate cross-lingual transfer for sequence tagging tasks. Thanks to a span-based projection technique, we can preserve token-level annotations. Our method provides an improvement of up to 20 points over zero-shot cross-lingual transfer. With translation-based adaptation, the gap between high- and low-resource multilingual settings can be reduced, and annotating datasets in multiple languages may not be entirely necessary depending on the needs.

Keywords: multilingual NLP, opinion mining, aspect-based sentiment analysis, cross-lingual transfer, translation-based adaptation, multi-domain, multi-task learning

Acknowledgements

First and foremost, I would like to express my gratitude for the invaluable support and guidance that my supervisor Gabriel offered throughout this work. Thank you to all Deskiñ members for welcoming me and integrating me into the team. Special thanks to Bénédicte for making sure I could work in the right conditions and to Géraldine for her insightful comments on my work.

On the personal side, I am especially indebted to my family for their love and support.

Contents

List of Figures

List of Tables

List of Abbreviations

ABSA	**A**spect-**b**ased **S**entiment **A**nalysis
ACD	**A**spect **C**ategory **D**etection
ASC	**A**spect **S**entiment **C**lassification
CNN	**C**onvolutional **N**eural **N**etwork
CRF	**C**onditional **R**andom **F**ield
E2E	**e**nd-**to**-**e**nd
GEC	**G**rammatical **E**rror **C**orrection
GRU	**G**ated **R**ecurrent **U**nit
LSTM	**L**ong **S**hort-**t**erm **M**emory
mBERT	**m**ultilingual **BERT**
MLM	**M**asked **L**anguage **M**odeling
MT	**M**achine **T**ranslation
NLI	**N**atural **L**anguage **I**nference
NLP	**N**atural **L**anguage **P**rocessing
NMT	**N**eural **M**achine **T**ranslation
OTE	**O**pinion **T**arget **E**xtraction
OWE	**O**pinion **W**ord **E**xtraction
POS	**P**art-**o**f-**S**peech
PTM	**P**re-**t**rained **M**odel
RNN	**R**ecurrent **N**eural **N**etwork
seq2seq	**s**equence-**to**-**s**equence
SER	**S**pelling **E**rror **R**ate

Chapter 1

Introduction

This work took place in an industrial research setting at Orange. We first provide some background information on the company and the context within which this work was carried out. We then turn to the research question and the problems associated with this work. Lastly, we describe the outline of this thesis.

1.1 Context

Orange is a multinational telecommunication corporation with 267 million customers worldwide. Orange employs around 147 thousand people and is the leading operator in a majority of the countries in which it is established. The company claims a model of engaged leadership regarding sustainability and social impact. Among other objectives for 2025, it wants to place data and artificial intelligence at the heart of its model. This internship took place at Orange Innovation, Orange's R&D department.

This work was conducted in the Deskiñ team which is in charge of research and development work in natural language processing (NLP). Its mission is to improve the analysis of customer feedback by researching new extraction and classification algorithms, and then developing technologies and services for Orange. The research work covers a wide range of NLP tasks (semantic analysis, information extraction, document structuring, knowledge management, etc.). The team also brings its expertise for the evaluation and continuous improvement of spoken language understanding systems to operational teams that develop chatbots, callbots or personal assistant solutions for Orange. The research work is also fueled by collaborations with academic laboratories and contributions to collaborative projects. The team is composed of 30 people including researchers, data scientists, software engineers, PhD students, and working students.

In order to address customers' needs, Orange is interested in what is said about its products and services. Large amounts of textual feedback are collected through surveys sent to Orange's customers which include both closed and open-ended questions. These surveys cover various domains such as mobile application reviews, feedback on customer care, or internal feedback intended for human resources. The manual analysis of text collected in the open questions is an expensive and time-consuming operation. Deskiñ develops and maintains VisualCRM, a customer relationship management tool dedicated to this issue. VisualCRM facilitates the analysis of customer feedback through opinion mining techniques designed to identify the evoked aspects. This tool also offers visualization interfaces to see these aspects evolve over time and compare them with other criteria, allowing end users to deduce trends and plans. The output of this work is to be integrated into VisualCRM.

1.2 Motivation

Opinion mining techniques allow us to automatically extract valuable information from subjective text such as reviews and answers to surveys. A basic opinion mining task consists in classifying the overall polarity of a given text. In this work, we are interested in going beyond polarity detection by extracting fine-grained information from customer feedback, e.g. which aspects of a product are being referred to, which entities are associated with them, and what is the customer's attitude towards them. This process refers to a more recent task called aspect-based sentiment analysis (ABSA, Pontiki et al., 2014). This task requires the extraction of token-level information from text and can be formulated as a sequence tagging task.

The wide variety of services (telephony, television, banking, B2C, B2B, etc.) and countries covered by Orange results in heterogeneous data spanning many different domains and languages. This variety demands generic approaches: In a production setting such as VisualCRM, deploying specialized models for each single use case is not feasible. The goal of this work is to research generic approaches capable of better understanding customers' opinions regardless of the service or the language. The next section presents this research question in more details.

1.3 Research Question

Deep neural networks require large amounts of annotated data to be effective. To be able to train such models for ABSA, several in-house datasets have been annotated according to standardized annotation guidelines. Such annotation process, however, is costly and its quality is difficult to control in a multilingual setting. Therefore, it is important to make use of as much supervision as available.

Transfer learning, in which knowledge obtained from a source problem is applied to a target problem, can be used to fully exploit the available annotated datasets. The various subtasks associated with ABSA aim to extract related information. Similarly, customers' reviews tend to follow similar syntactic structures regardless of the domain. Accordingly, multi-task and multi-domain learning can prove to be particularly useful to entice generalizable representations and produce generic models capable of tackling various tasks and domains simultaneously.

Large models pre-trained with a self-supervised task and fine-tuned with supervision on a downstream task have been widely adopted for various NLP tasks. During this pre-training phase, these models seem to capture high-level linguistic properties that benefit many tasks. Recent work has shown that such models can be pre-trained on many languages at once, and that they are then able to generalize across multiple languages (Conneau et al., 2020). More impressively, they have been applied effectively for zero-shot cross-lingual transfer, where a model is fine-tuned for a task on a given language and is then applied to an unseen language for the same task (Radford et al., 2019).

This work explores ways of improving cross-lingual transfer through translation-based adaptation. We introduce a method to adapt datasets annotated for sequence tagging from a source language to a target language. This method relies on word alignments to project spans effectively and adapt token-level annotations. We study the effectiveness of this method 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. While the ability of

large pre-trained models (PTMs) to generalize across different languages is impressive, this work suggests that multilingual representations obtained from such models are in fact limited and that fine-tuning on synthetic data in the target language is beneficial. Moreover, our results show that when extending a model to new target languages with annotated data available in a source language only, the annotation process for the target languages could be bypassed thanks to our adaptation method. We compare different resource settings to establish how original annotated data in the target languages affects performance and find that low amounts of annotated data can be combined effectively with the translated data.

Another aspect that we address is noisy user-generated text, particularly with regard to our adaptation method. Noisy text contains spelling and grammatical errors, abbreviations, non-words, etc. This poses a challenge to PTMs, which may only have been trained on clean text and therefore have a hard time dealing with these deviations from standard forms. Though such models are able to handle noisy text to some extent through their subword tokenization, we found that our in-house datasets were particularly challenging. We address this issue by automatically correcting utterances in the source language before translating them into a target language. We also design a heuristic to identify the most noisy utterances. We find that it can be beneficial to keep cleaner examples only when translating a source language corpus or when sampling examples to annotate.

1.4 Contribution

Our contributions are as follows:

- We introduce a novel translation-based adaptation method to project span-level annotations from source to target languages and generate synthetic datasets for sequence tagging tasks.
- We conduct experiments on the ABSA task with various data configurations to demonstrate the effectiveness of this method for both cross-lingual adaptation and data augmentation, and how to make best use of annotated data.
- We show how different tasks, domains, and languages can be combined into one model thanks to transfer learning.
- We propose several approaches to improve robustness to noisy text with regard to our translation-based adaptation method.

1.5 Outline

Chapter 2 provides some background information useful to the reader on topics related to this work. In Chapter 3, the datasets that were used to conduct this research are presented along with a common benchmark for ABSA. Chapter 4 turns to the model architecture that we adopted to address this task. Chapter 5 describes our annotation projection method for sequence tagging, along with the various data configurations that can be derived from this method, both for cross-lingual adaptation and data augmentation. In Chapter 6, we give some details on the experimental setup and describe the results that we obtained. Drawing on these results, Chapter 7 introduces an analysis of several aspects of this work, such as how much annotation is needed and how to deal with noisy text. Lastly, Chapter 8 concludes this work with a discussion.

Chapter 2

Background

2.1 Deep Learning for NLP

In the last decade, increasing access to large-scale annotated datasets and computational resources led to the widespread adoption of machine learning algorithms for NLP. The success of these methods is based on probabilistic models learned through representation learning, i.e. ways of representing the data and extracting features from it (Bengio, Courville, and Vincent, 2013). Probabilistic models attempt to recover latent random variables to describe a distribution over the observed data, and can be expressed as $P(h, x)$ where h and x refer to latent variables and observed data respectively.

More recently, deep learning architectures, which do not require explicit feature engineering, have enabled large performance gains across a variety of supervised NLP tasks. These gains can be attributed to a great extent to the efficient use of transfer learning and adaptation of large PTMs such as BERT (Devlin et al., 2019). In this section, we provide some background information on transfer learning and the Transformer model which is the backbone architecture of many of the recent PTMs.

2.1.1 Transfer Learning

Transfer learning consists in retrieving information from a source problem to apply it to a target problem. This principle is illustrated in Figure 2.2. Typically, information acquired from a source task is adapted to a target task. This principle can also be applied to adapt information from one domain to another (domain adaptation), or even from one language to another (cross-lingual transfer).

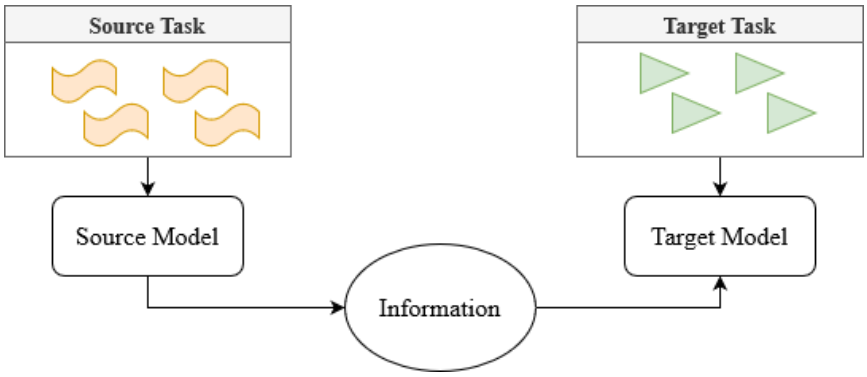


FIGURE 2.1: An example application of transfer learning.

Transfer learning can be applied in various scenarios depending on the available annotated data. Ruder, 2019 defined a taxonomy of transfer learning for NLP, dividing it into two main scenarios:

- Transductive transfer learning: Source and target tasks are the same, but there is a discrepancy between the distributions of source and target data. This scenario includes domain adaptation and cross-lingual transfer.
- Inductive transfer learning: The source task is different from the target task, but labeled data is available in the target domain. Tasks can be learned one after the other (sequential transfer learning) or simultaneously (multi-task learning).

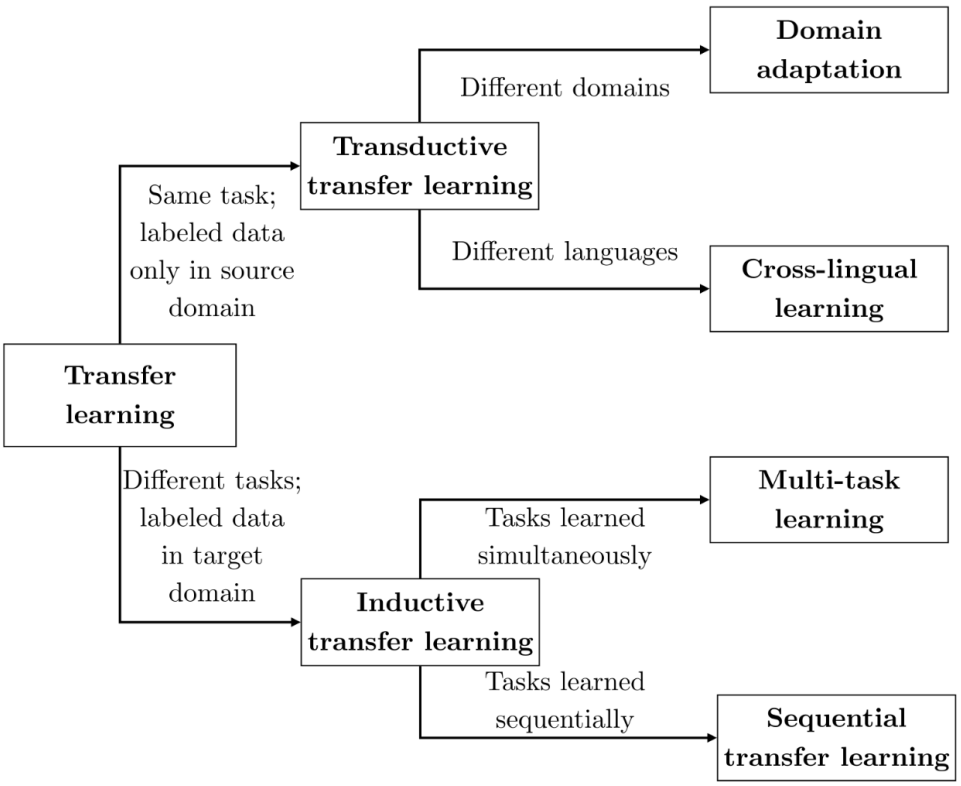


FIGURE 2.2: A taxonomy of transfer learning for NLP (Ruder, 2019).

A sequential transfer learning approach in NLP that has now become ubiquitous is the use of PTMs to obtain word embeddings. Word embeddings, map words from a vocabulary to continuous representations in a shared vector space. These representations are learned in such a way as to capture the meaning of words. A sequence of tokens $(x_1, ..., x_n)$ can be encoded as a sequence of word embeddings $(h_1, ..., h_n)$ to enrich the representation of a sentence for a supervised downstream task, a process illustrated in Figure 2.3.

According to the distributional hypothesis, word occurring in similar contexts tend to have similar meanings (Firth, 1957). Based on this assumption and the observation that annotated data is rare, several approaches have been introduced to learn word representations by leveraging large amounts of unlabeled text which is readily available using self-supervised pre-training objectives.

The first generation of PTMs were used to obtain pre-trained word embeddings. Formally, the goal is to map each word x in a vocabulary $\mathcal{V}$ to an embedding $e_x \in \mathbb{R}^{d_e}$ in a static way using a lookup table $E \in \mathbb{R}^{|\mathcal{V}| \times d_e}$, where d_e is the embedding size. Only the lookup table is needed and the resulting pre-trained models are discarded. In consequence, static word embedding algorithms such as Skip-Gram (Mikolov et

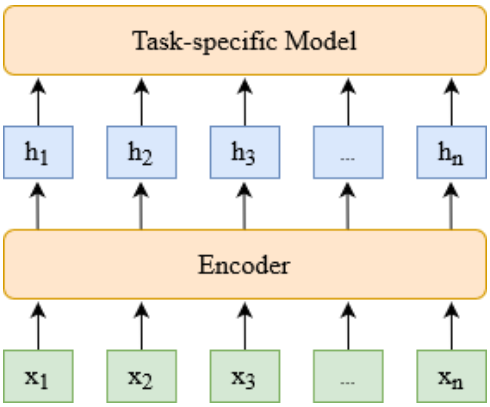


FIGURE 2.3: A common approach to transfer information from a PTM.

al., 2013) and GloVe (Pennington, Socher, and Manning, 2014) were designed as shallow architectures for computational efficiency.

Discrete language symbols are mapped to continuous vector representations using a fixed lookup table. Hence, this approach fails to capture polysemy or homonymy. Moreover, out-of-vocabulary words are not handled. To address this second issue, character-level representations or subword representations were introduced such as fastText (Bojanowski et al., 2017) and Byte-Pair Encoding (Sennrich, Haddow, and Birch, 2016).

Contextualized embeddings account for polysemy by encoding words dynamically based on their context. Given a sequence of words or subwords $(x_1, ..., x_n)$, the objective is to learn a function that encodes each token $x_t \in \mathcal{V}$ into a contextualized representation $h_i \in \mathbb{R}^{d_e}$.

$$(h_1, ..., h_n) = f_{enc}(x_1, ..., x_n) \tag{2.1}$$

where f_{enc} is a neural contextual encoder.

The initial approaches to contextualized representations were also used as feature extractors. The first generation of contextual encoders usually consisted of LSTM-based language models that were used to extract features and obtain word embeddings only. The task-specific models were then trained from scratch but benefited greatly from these contextual word embeddings (Peters et al., 2018; Akbik, Blythe, and Vollgraf, 2018).

Building on these approaches, modern PTMs are directly fine-tuned on a downstream task to update all of the model weights, making a more effective use of transfer learning. To learn universal representations, large PTMs such as BERT (Devlin et al., 2019) and GPT (Radford and Narasimhan, 2018) are trained on large-scale corpora using a language modeling objective or a variation of it. They adopt deeper or more powerful architectures, e.g. the Transformer, along with new self-supervised tasks. Devlin et al., 2019 adopted a masked language modeling (MLM) objective to effectively learn bidirectional representations. MLM randomly masks tokens in the input and the model is trained to predict the masked tokens using the surrounding context. Lewis et al., 2020 introduced BART, a denoising autoencoder to pre-train a sequence-to-sequence model (seq2seq). The model is trained by corrupting text with a noising function and learning to reconstruct the original input.

Empirically, transfer learning has led to state-of-the-art performance across many supervised NLP tasks, e.g. classification, question answering, and information extraction (Devlin et al., 2019). This success led to the development of ever larger

models (Raffel et al., 2020). These models are so large that they have been shown to perform well on a downstream task even with no parameter update, an approach called zero-shot learning (Brown et al., 2020).

Most modern PTMs such as BERT are based on the Transformer architecture which will be described in the next subsection.

2.1.2 Transformer

The Transformer is an encoder-decoder architecture designed for sequence transduction tasks such as language modeling or machine translation. Before the Transformer, the recurrent neural network (RNN) architecture was commonly used to address such tasks. To process sequences of variable length, RNNs encode the inner state at each time step t by combining the memory of the preceding context up to $t-1$ with the information of the current word.

This architecture naturally lends itself to modeling text which is inherently sequential. But it struggles to make use of distant information and is difficult to train because of the vanishing gradient problem. Gated RNNs such as LSTM networks (Hochreiter and Schmidhuber, 1997) are commonly adopted to mitigate these issues by using a gate mechanism that controls the flow of information, but the underlying problem still remains. Moreover, the sequential nature of these networks prevents the use of parallel computational resources.

Based on these considerations, the Transformer model was introduced as a fully parallelizable architecture for efficient sequence processing. This model maps a sequence of input vectors $(x_1, ..., x_n)$ to a sequence of output vectors $(y_1, ..., y_n)$. It consists of stacks of linear layers, feedforward networks, and specific connections around them. Inspired by the observation that successful LSTMs relied on the attention mechanism, the original Transformer work (Vaswani et al., 2017) also incorporated novel self-attention layers which are a key component of this model.

At its core, the attention mechanism compares an item x_i with other items (e.g. using the dot product) to find their relevance in the current context. With self-attention, the compared elements are in the same sequence. This comparison can be expressed as a score:

$$score(x_i, x_j) = x_i \cdot x_j \tag{2.2}$$

We can use these scores by normalizing them with a softmax function to create a weight vector α_{ij} which indicates how relevant an item x_j is to the input item x_i, i.e. the current focus of attention.

$$\begin{aligned} \alpha_{ij} &= softmax(score(x_i, x_j)) \quad \forall j \\ &= \frac{exp(score(x_i, x_j))}{\sum_{k=1}^{n} exp(score(x_i, x_k))} \quad \forall j \end{aligned} \tag{2.3}$$

We then take the sum of the inputs weighted by their respective α value to generate an output value y_i.

$$y_i = \sum_j \alpha_{ij} x_j \tag{2.4}$$

With this mechanism, there is no opportunity for learning as the result is based on the input values x only. To learn how words contribute to the representation of an input sequence, the Transformer introduces a query matrix $W^Q \in \mathbb{R}^{d_k \times d_{model}}$, a key matrix $W^K \in \mathbb{R}^{d_k \times d_{model}}$, and a value matrix $W^V \in \mathbb{R}^{d_v \times d_{model}}$ as additional parameters that operate over the input embeddings. Intuitively, the roles of these matrices can be interpreted as follows: the query is the current focus of attention, the key is what

is being compared to the focus of attention, and the value is used to compute the output of the focus of attention. The self-attention mechanism leverages efficient matrix multiplication to process an input sequence X in parallel.

$$Q = W^Q X;\ K = W^K X;\ V = W^V X \tag{2.5}$$

We can compute the comparisons between query and key by multiplying Q and K, and take their softmax directly. Note that the comparison scores are scaled according to the size of the embeddings to avoid exponentiating large values, since it can lead to a loss of gradients during training. The result is then mutliplied by V, reducing self-attention to the following expression:

$$SelfAttention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.6}$$

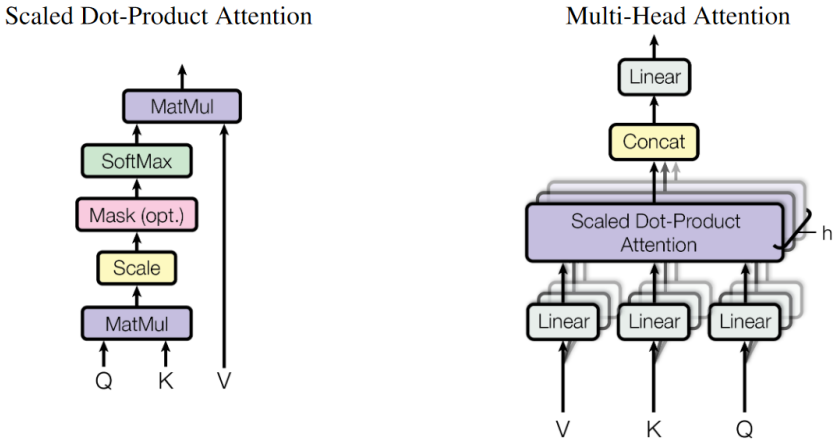


FIGURE 2.4: The two main components of the Transformer model (Vaswani et al., 2017).

To capture the various relationships words can have with each other in a sentence, e.g. syntactic, semantic, and discourse relationships, the Transformer introduces multi-head attention layers. Different sets of key, query, and value matrices are attributed to each attention head i, which attends to information from a specific representation subspace. The outputs from all heads are concatenated and projected to the output dimension with a linear layer $W^O \in \mathbb{R}^{hd_v \times d_{model}}$.

$$\begin{aligned} MultiHeadAttn(Q, K, V) &= W^O(head_1 \oplus head_2 \oplus ... \oplus head_h) \\ head_i &= SelfAttention(W_i^Q X, W_i^K X, W_i^V X) \end{aligned} \tag{2.7}$$

Figure 2.4 illustrates the scaled dot-product attention and multi-head attention concepts that were introduced with the Transformer model.

A final consideration about the Transformer architecture is the use of positional encodings. Since words are processed in parallel, all information about their absolute and relative position is lost. To capture the relationships among the positions of words, a function maps integer inputs to real-valued vectors of size d_{model}, which are then summed with the input embeddings.

The complete Transformer architecture is shown in Figure 2.5. A transformer block consists of multi-head attention layer and a feedforward layer, each with a

residual connection around it followed by a normalization step. Since this architecture is designed for sequence transduction tasks, it includes an encoder and a decoder, which consist in stacks of n transformer blocks. Using attention, the decoder attends to i) earlier positions in the output sequence, and ii) the encoded representation of the input. The BERT model which is used in this work relies on the encoder only.

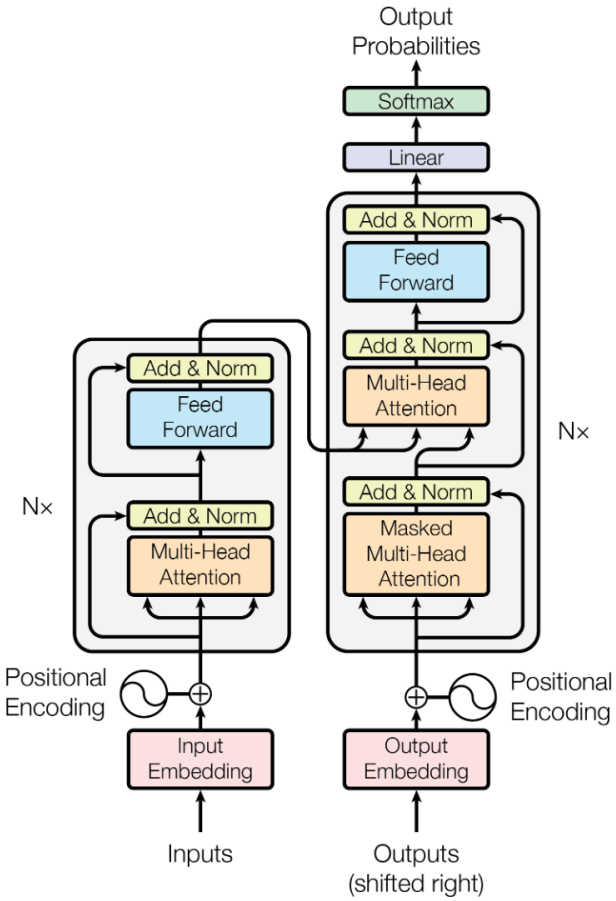


FIGURE 2.5: Overview of the Transformer model (Vaswani et al., 2017).

2.1.3 Sequence Tagging

Many NLP tasks that involve span recognition are treated as sequence tagging tasks, e.g. named entity recognition and, in the case of this work, aspect-based sentiment analysis (ABSA). Formally, the goal is to assign a label y_i to each word x_i in an input word sequence X, such that the output label sequence Y has the same length as X. The label is predicted from a fixed set of labels.

The standard approach to sequence tagging for span recognition is `BIO` tagging, short for beginning, inside, and outside. Thank to this tagging scheme, span recognition problems can be formulated as sequence tagging tasks by capturing span boundaries. Tokens that begin a span are labeled `B`, tokens that occur within a span are labeled `I` , and tokens outside of any span of interest are labeled `O`. This results in a set $\mathcal{Y}$ of $2n + 1$ tags, where n is the number of classes. Other tagging schemes are shown in Table 2.1. `IO` loses some information by removing the `B` tag, while `BIOES` is

more refined with tokens ending a span labeled as `E` and single-token spans labeled as `S`.

Scheme	The	app	is	slow	but	I	like	the	new	interface	.
`BIO`	O	B-target	O	O	O	O	O	O	B-target	I-target	O
`IO`	O	I-target	O	O	O	O	O	O	I-target	I-target	O
`BIOES`	O	S-target	O	O	O	O	O	O	B-target	E-target	O

TABLE 2.1: ABSA as a sequence model with various tagging schemes ("target" refers to an opinion target, see Subsection 2.2.1).

A sequence tagger is trained to assign a label to each token in a text to indicate the presence or absence of a span of interest. The token-level predictions are computed with a softmax activation function over the given tagset C.

$$\begin{aligned} P(y_i = c|x_i) &= softmax(W_c h_i + b_c) \\ &= \frac{exp(W_c h_i + b_c)}{\sum_{j=1}^{C} exp(W_j h_i + b_j)} \end{aligned} \tag{2.8}$$

where W and b are the parameters for the softmax layer and h_i is a word representation of x_i obtained with e.g. a Transformer encoder or an RNN. The softmax provides a probability distribution over the possible classes.

2.2 Opinion Mining

Opinion mining is a task concerned with extracting opinions, attitudes, and emotions expressed in user-generated content (Liu and Zhang, 2012). This task is technically challenging: Online communication can be ambiguous, and identifying what a user had in mind when writing a piece of text is not trivial, even for humans. But the information extracted from this task is potentially very useful. For example, businesses can apply opinion mining techniques to automatically analyze customer feedback on a large scale and extract valuable insights from it to improve their services and meet their customers' needs. With the proliferation of user-generated content online, e.g. on social media and review websites, there has been growing interest in mining opinions from text in both academia and industry.

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). This can be done with varying degrees of granularity depending on the defined classes which can express polarity, e.g. {`negative`, `positive`}, {`negative`, `positive`, `neutral`, `mixed`}, etc. or both polarity and intensity, e.g. {`-5`, ..., `0`, ..., `+5`} (Tian, Lai, and Moore, 2018).

When analyzing online reviews of a product, it can be important to know which aspects or features are evoked in order to address potential flaws. Most approaches however focus on classifying the overall tone of a sentence, with no regard to what entities and aspects may be mentioned. Aspect-based sentiment analysis (ABSA) aims to extract fine-grained information such as entities, their attributes, and their sentiment (Hu and Liu, 2004; Popescu and Etzioni, 2005; Pontiki et al., 2016; Zhou et al., 2019). The focus of this work is ABSA. The next subsection provides an overview of current works dealing with this task.

2.2.1 Aspect-based Sentiment Analysis

Several subtasks are associated with ABSA, for example:

- Opinion target extraction (OTE) retrieves the entity on which an opinion is expressed.
- Aspect category detection (ACD) identifies the category on which an opinion is expressed. The categories are usually predefined.
- Aspect sentiment classification (ASC) identifies the polarity of an opinion expressed on a given target entity, e.g. {`negative`, `positive`, `neutral`}.
- Opinion word extraction (OWE) retrieves the word that explains an opinion for a given target entity.

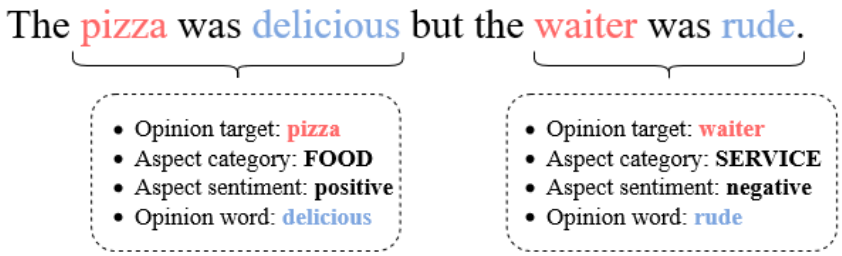


FIGURE 2.6: An example review with values for various ABSA subtasks.

Figure 2.6 illustrates these subtasks with an example sentence.

Some works focus on one subtask at a time, e.g. OTE (Li and Lam, 2017; Xu et al., 2018) or OWE (Fan et al., 2019; Wu et al., 2020b; 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. This approach poses some challenges: OTE and ASC are distinct tasks (extraction vs. classification), and the number of aspect term-polarity pairs in a sentence is arbitrary. Based on this observation, the problem is commonly formulated as a sequence tagging task with unified labels to simultaneously detect opinion targets and the corresponding aspect sentiments, as exemplified in Table 2.2. We refer to this approach as E2E-ABSA.

Input	The	app	is	slow	but	I	like	the	new	interface	.
Unified	O	S-NEG	O	O	O	O	O	O	B-POS	E-POS	O

TABLE 2.2: Example sequence of labels for E2E-ABSA.

Li et al., 2019a proposed an end-to-end system composed of two stacked RNNs. The first layer predicts the boundaries {`B`,`I`,`E`,`S`,`O`} to guide the second layer that predicts the unified tags. Their system also includes i) a gate mechanism to maintain sentiment consistency by predicting a label based on the previous one, and ii) an opinion-enhanced target detection module that checks if there is at least one opinion word in a fixed window using an opinion lexicon. Luo et al., 2019 used a dual RNN to extract representations for OTE and ASC. A cross-shared unit is then used to promote interactions between the OTE and ASC components, with two attention scores

matrices used to quantify the correlation between both subtasks. Li et al., 2019b used BERT to obtain contextual word embeddings followed by a classification layer dedicated to E2E-ABSA (linear, GRU, self-attentive, or CRF), setting a new state-of-the-art performance with a simple linear classification layer.

Some works went even further by addressing OTE, ASC, and OWE simultaneously, a task dubbed as aspect sentiment triplet extraction. Peng et al., 2020 used a two stage framework to i) extract triplets, and ii) generate candidate pairs and classify valid pairs with distance embeddings. Xu et al., 2020b introduced a position aware tagging scheme to jointly extract the triplets with a neural CRF. Wu et al., 2020a proposed a grid tagging scheme for end-to-end sentiment triplet extraction.

To sum up, different ABSA subtasks benefit from informing each other rather than being learned in isolation since these subtasks are correlated to each other (Wang et al., 2017; Dai and Song, 2019). For instance, the aspect sentiment is determined by the opinion word and opinion target, and an opinion word depends on the associated opinion target, e.g. "fresh" is collocated with "food". Pipeline approaches propagate errors and the different components do not work together. In consequence, there has been a tendency to favor end-to-end approaches to jointly tackle multiple ABSA subtasks. In this work, we focus on E2E-ABSA, i.e. joint OTE and ASC.

2.2.2 Processing Noisy Text

Opinion mining techniques are typically applied to user-generated text. Because of the informal nature of computer mediated discourses from which it is collected, this textual data is inherently 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. Such variety results in a data distribution that is difficult to model, and calls for opinion mining techniques that are robust to noisy text.

Several approaches have been proposed to normalize noisy text into standard form before processing it further. Desai and Narvekar, 2015 sought to replace out-of-vocabulary words with the best possible correction using the edit distance. Other works have focused on adapting word representations to noisy text, for both static (Sumbler et al., 2018; Malykh, Logacheva, and Khakhulin, 2018), and contextualized (Muller, Sagot, and Seddah, 2019; Sun and Jiang, 2019) word embeddings.

Kumar, Makhija, and Gupta, 2020 conducted a study on how noisy text affects the performance of BERT. By introducing synthetic noise in a corpus, they showed that as noise increases, BERT's performance drops drastically. As a solution, the authors suggest correcting misspelling mistakes in a dataset before fine-tuning BERT on it.

Grammatical error correction (GEC) is the task of correcting erroneous text. Most modern approaches to GEC rely on the encoder-decoder architecture to encode an erroneous input text and generate a corrected version (Chollampatt and Ng, 2018). The task can be formulated as a low-resource translation problem, in which synthetic data generation is used to alleviate the lack of resources. This generation can be done by applying modifications to clean text that mimic realistic errors (Grundkiewicz, Junczys-Dowmunt, and Heafield, 2019). Edits from Wikipedia revision history have been successfully used as parallel training examples, provided only those containing grammatical corrections are selected (Boyd, 2018). Pre-trained seq2seq models such as BART have been adapted successfully for GEC, thus reducing the amount of parallel data needed (Katsumata and Komachi, 2020).

2.3 Multilingual NLP

For multinational companies, a multilingual approach to opinion mining is essential to account for customer feedback in many different languages. Due to the availability of resources, much work in NLP previously focused on English and other high-resource languages. In recent years, the gap with other languages has been reduced thanks to transfer learning. Information acquired from a language with abundant resources can be transferred to mid- and low-resource languages for various tasks. This section provides an overview of recent work on multilingual representations and cross-lingual transfer.

2.3.1 Multilingual Models

As we have seen in Section 2.1, large PTMs fine-tuned on a downstream task have become the de facto standard in NLP. However, monolingual PTMs are only available for some high-resource languages due to the lack of resources. Additionally, in a multilingual production setting, having a separate model for each language is not possible.

Modern multilingual models such as multilingual BERT (mBERT, Devlin et al., 2019 and XLM-R (Conneau et al., 2020) extend their monolingual counterparts by learning from large multilingual corpora which cover up to a hundred different languages. No indication about the language of the data is available to them during pre-training. They simply share a large vocabulary across all languages. They have been shown to yield on par, if not better performance on various tasks compared to monolingual models, and they 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 on an unseen language (Radford et al., 2019).

As promising as this approach might be, one might wonder to what extent these models can encode similar representations regardless of the language, and if they truly are language agnostic. Previous works have sought to provide some insight into why these models succeed at cross-lingual transfer, using e.g. probing methods (Choenni and Shutova, 2020). Pires, Schlinger, and Garrette, 2019 claimed that such models leverage shared word-pieces and that cross-lingual transfer is thus most effective between typologically similar languages. K et al., 2020 argued that, on the contrary, lexical overlap plays a negligible role in cross-lingual transfer and that network depth is a key factor for it. All in all, more work remains to be done in order to fully understand and measure how multilingual these models are.

2.3.2 Cross-lingual Transfer

With strong capacities for generalization, multilingual PTMs have been successfully applied for zero-shot cross-lingual transfer on a variety of natural language understanding tasks (Wu and Dredze, 2019). Considering that labeled data is more readily available for certain languages, this technique can be harnessed to process low-resource languages, therefore bypassing the need for a costly annotation process. Jebbara and Cimiano, 2019 have explored cross-lingual transfer for opinion target extraction (OTE) using a CNN coupled with multilingual word embeddings that have been aligned in a shared vector space. Adversarial training has also been

shown to improve cross-lingual transfer, in which a discriminative network is used to incite language-agnostic representations (Joty et al., 2017; Keung, Lu, and Bhardwaj, 2019).

2.3.3 Translation-based Adaptation

In recent years, the quality of machine translation (MT) has improved considerably with the adoption of neural neural machine translation (NMT) and more specifically Transformer models. Additionally, off-the-shelf machine translation systems are available for a wide range of language pairs[1] thanks to initiatives such as OPUS-MT (Tiedemann and Thottingal, 2020).

MT can help learn cross-lingual representations and transfer information across languages (Zhou, Wan, and Xiao, 2016; Eger et al., 2018). Rather than being at odds, zero-shot cross-lingual transfer and translation-based adaptation approaches are complementary: MT can be used to generate synthetic data in languages for which no annotated data is available. Accordingly, MT can enable cross-lingual transfer in two ways: either by translating the training data into the target languages and fine-tuning on all languages (Banea et al., 2008; Duh, Fujino, and Nagata, 2011), 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). In this work, we focus on the former approach. This approach naturally lends itself to sentence classification tasks and previous work has shown that translation-based adaptations are 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 token level for sequence tagging tasks. As no word-to-word correspondence is available, an important additional step of annotation projectionis required.

One approach to annotation projection for sequence tagging tasks relies on word alignments to project labels from source to target utterances. Several statistical word alignment tools are available such as `fast_align` (Dyer, Chahuneau, and Smith, 2013) and `efmaral` (Östling and Tiedemann, 2016), and can serve as baseline methods for this approach. Jalili Sabet et al., 2020 leveraged modern multilingual PTMs and released a tool called SimAlign for unsupervised word alignment based on the similarity of multilingual word representations. Starting from the assumption that NMT models capture word alignments through their attention mechanism, other works focus on using attention weights for word alignment (Chen et al., 2020; Zouhar and Pylypenko, 2021).

Others have sought to improve label projection through more refined approaches. Jain, Paranjape, and Lipton, 2019 proposed an entity projection method for named entity recognition. 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, Haider, and Mansour, 2020 proposed an end-to-end model to jointly align and predict target slot labels for cross-lingual transfer. In a recent paper, Li et al., 2020 argued that zero-shot cross-lingual transfer surpassed translation-based adaptations on several sequence tagging tasks, and proposed a warm-up adaptation method to optimize the use of the translated data. In contrast, we show that a plain translation-based adaptation yields superior performance compared to zero-shot cross-lingual transfer, provided the right annotation projection method is used.

[1] https://github.com/Helsinki-NLP/Opus-MT-train/tree/master/models

Chapter 3

Data

In this chapter, we describe the datasets from the SemEval shared task dedicated to ABSA, as well as the in-house datasets we worked on. We then provide a contrastive analysis of these two sets of corpora to highlight their similarities and differences, and the challenges that they pose.

3.1 SemEval-2016 Task 5

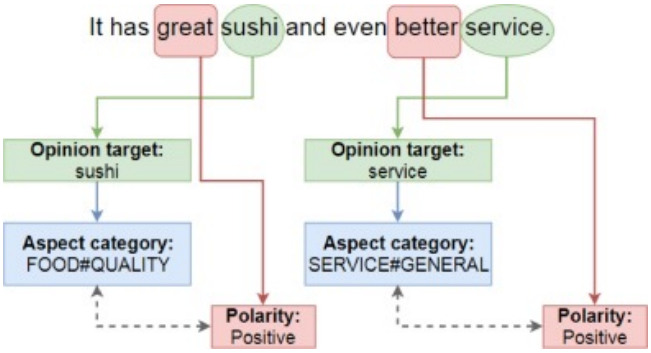


FIGURE 3.1: An example annotated review for ABSA from SemEval-2016 Task 5 (Do et al., 2019).

SemEval-2016 Task 5 (Pontiki et al., 2016) is an evaluation campaign for multilingual ABSA. This shared task is a continuation of the respective tasks in 2014 and 2015 which dealt with English data in two domains (Pontiki et al., 2014; Pontiki et al., 2015). For the 2016 edition, online reviews from several domains were annotated under common guidelines by several research groups in their respective language. The datasets include both sentence-level and token-level annotations and are available across eight languages and seven domains. The main subtask consists in identifying any of the following opinion tuple slots given a review about a target entity:

- Slot 1: Identification of the aspect category, i.e. a pair E#A of entity E and attribute A, e.g. FOOD#QUALITY (ACD).
- Slot 2: Extraction of the linguistic expression referring to the entity E, referred to as an opinion target defined by its starting and ending offsets (OTE).
- Slot 3: Identification of the sentiment polarity for a given entity-attribute pair E#A. Polarity labels include `positive`, `negative`, and `neutral` (ASC).

Datasets annotated for Slot 2 are only available for the restaurant and hotel domains. Since our focus is E2E-ABSA, i.e. joint OTE and ASC, we only consider

those datasets in our experiments. Table 3.1 provides an overview of these datasets, which are available in Arabic (ar), English (en), Spanish (es), French (fr), Dutch (nl), and Russian (ru). The annotations include opinion targets, along with their associated sentiment label and aspect category. The Arabic dataset contains hotel reviews, while the datasets in all the other languages contain restaurant reviews.

Language	Train		Test	
	# utter.	# aspects	# utter.	# aspects
ar	4802	9612	1227	2371
en	2000	1743	676	612
es	2070	1859	881	713
fr	1664	1641	668	650
nl	1722	1859	575	713
ru	3665	3078	1209	952

TABLE 3.1: Statistics for the SemEval datasets.

3.2 In-house Datasets

As a large company operating in several countries, Orange receives a large flow of customer feedback in various languages. One goal therefore is to extract valuable information from this data. To that end, several in-house datasets have been annotated in a similar manner as the SemEval ABSA data for domains and languages that the company handles. These datasets cover the following languages: Arabic, English, Spanish, French, Dutch, Polish (pl), and Romanian (ro). One corpus composed of 1000 annotated reviews for the mobile application MyOrange is common to all these languages.[1] Reviews in French, Spanish, and Polish from other applications have been annotated using the same guidelines. We aggregate all data from the application domain, and only use the test sets from MyOrange data for a comparable evaluation. Table 3.2 provides an overview of the datasets that were used in our experiments.

Language	Train		Dev		Test	
	# utter.	# aspects	# utter.	# aspects	# utter.	# aspects
ar	592	741	198	280	198	237
en	600	780	200	279	200	269
es	1200	1080	200	354	200	372
fr	2047	1984	131	84	131	84
nl	594	554	198	183	198	190
pl	864	1827	200	610	200	581
ro	600	1050	200	358	200	357

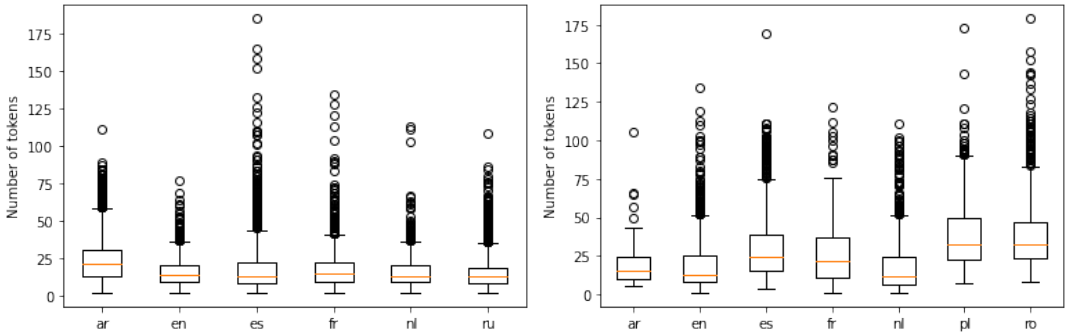
TABLE 3.2: Statistics for the in-house datasets.

3.3 Data Analysis

One particular interest of this work is to establish how well can PTMs generalize across domains, and whether it would be useful to combine ABSA datasets from

[1] The Arabic corpus is composed of four subsets in different dialects which we aggregate.

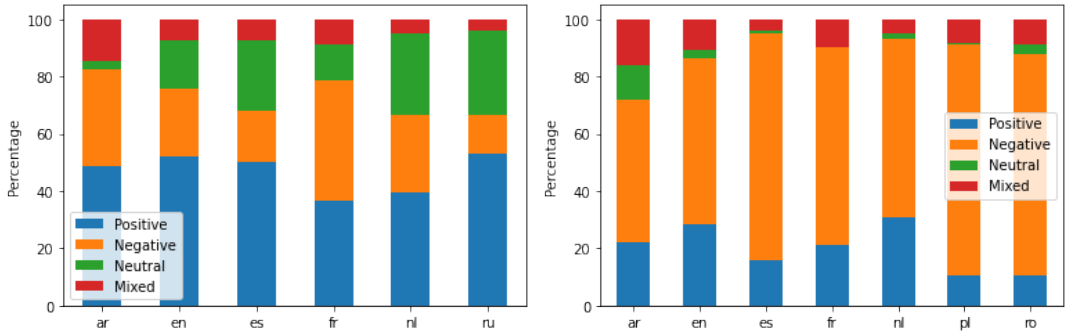
different domains. In this section, we provide a contrastive analysis of the SemEval and in-house datasets to highlight their similarities and differences, and the challenges that combining them may pose.



(A) SemEval datasets (B) In-house datasets

FIGURE 3.2: Distribution of utterances length.

While both sets of corpora deal with customer feedback, they come from distinct domains and opinions have been voiced through different media. As such, both collections come with their own peculiarities. Figure 3.2 shows that, while somewhat similar in sizes, utterances in the in-house datasets tend to be longer on average. The SemEval datasets are comprised of whole reviews of restaurants with gold sentence segmentation, and sentences are processed individually by the model. In contrast, the in-house datasets are processed by whole reviews since no gold segmentation is available. Automatic sentence segmentation is error-prone because of the noisy nature of these reviews, e.g. no punctuation mark in a review, or conversely, excessive use of punctuation marks.



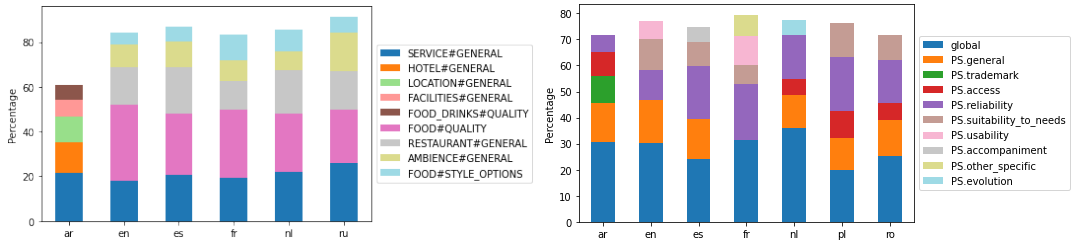
(A) SemEval datasets (B) In-house datasets

FIGURE 3.3: Global polarity distribution. The `mixed` label denotes a review which contains conflicting polarities for different aspects, e.g. "interface" → `positive`, "performance" → `negative`.

A user that writes a comment for a restaurant or a hotel on a review website will tend to describe her experience in detail, highlighting different aspects such as food quality or atmosphere. Some may be positive, some negative. The overarching goal being the sharing of valuable information to other users. As a result the overall polarity distribution in the SemEval datasets is balanced. In contrast, the in-house datasets tend to contain more polarized opinions. Being asked to provide a review of an application, a satisfied user will tend to give a rating along with a short message such as "good" or "not bad". However, such reviews are the minority. Most users

that do comment tend to be unsatisfied with the app and therefore write comments explaining at length what their problem may be. This poses a challenge as, on the one hand, short positive comments do not provide much information to train the model, and on the other hand, longer, mostly negative comments are challenging to process accurately. This contrast in polarities is best illustrated by Figure 3.3. The in-house datasets are marked by an overwhelming majority of negative opinions while the SemEval datasets contain more balanced sentiment labels.

Figure 3.4 shows another area where SemEval and in-house datasets differ. As both sets of corpora come from different domains, the aspects are quite distinct. While the most common aspects in both domains focus on the global picture, other aspects are domain-specific. This suggests that combining both domains will introduce a higher variance in the data distribution.



(A) SemEval datasets (B) In-house datasets

FIGURE 3.4: Most common aspect categories distribution.

One final aspect that we wish to highlight is the amount of noise in the data. User-generated text is notoriously noisy in the sense that it may contain spelling and grammatical errors, abbreviations, etc. Such data is therefore challenging to process. To quantify noise in the data and compare SemEval and in-house datasets, we compute what we call the spelling error rate (SER) using Hunspell[2] with dictionaries from Firefox for each language.[3] The spelling error rate is computed as follows:

$$SER = \frac{E}{N} \times 100 \tag{3.1}$$

where E is the number of spelling errors and N the number of tokens in the corpus. We define a spelling error as any token that (i) does not belong to the spellchecker's dictionary and (ii) is not a punctuation mark. Results can be seen in Table 3.3. We observe that, on average, spelling errors are more than twice as likely to occur in the in-house datasets compared to SemEval datasets, highlighting another difficulty of the in-house datasets.[4]

	fr	**en**	**nl**	**es**	**ar**	**avg.**
SemEval	1.38	2.53	2.07	4.02	12.07	4.41
In-house	4.26	3.67	4.12	6.96	25.61	8.92

TABLE 3.3: Spelling error rates in the training sets.

In this chapter, we have seen that SemEval and in-house show marked dissimilarities. As a consequence combining them may not be straightforward.

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

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

[4] Note that the higher average spelling error rate in Arabic can be attributed to dialectical variations.

Chapter 4

Model

4.1 Model Description

As we have seen in Chapter 2, ABSA may refer to several subtasks. In this work, we focus on E2E-ABSA, i.e. joint opinion target extraction (OTE) and aspect sentiment classification (ASC). Following Li et al., 2019a, we use a unified approach with a single sequence tagging model to jointly extract opinion targets and their associated sentiment label. An overview of the model's architecture is shown in Figure 4.1.

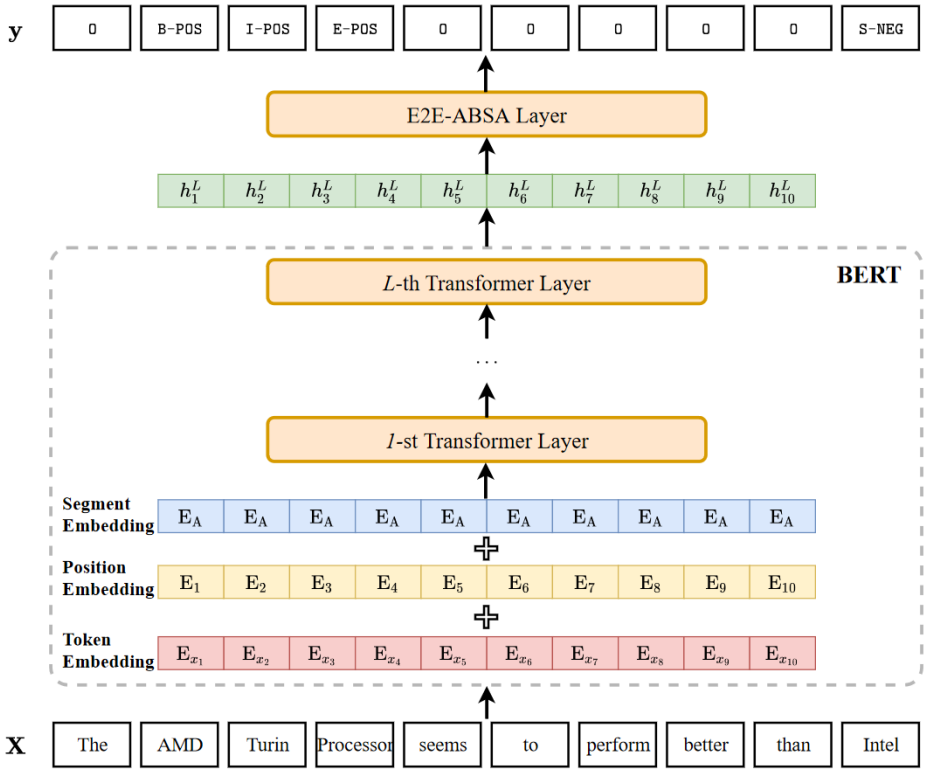


FIGURE 4.1: Model architecture (Li et al., 2019b)

We have also seen in Chapter 2 how E2E-ABSA can be formulated as a sequence tagging task. Given an input sequence of tokens $X = (x_1, ..., x_N)$ of length N, we use a BERT encoder model to obtain contextual representations of all tokens $H^L = (h_1^L, ..., h_N^L) \in \mathbb{R}^{N \times dim_h}$ where dim_h corresponds to the dimension of the representation vector and L to the final Transformer layer. These contextual representations are then fed into a linear classification layer to predict the sequence of labels $(y_1, ..., y_N)$. A softmax activation function is used to compute the predictions for

each token as follows:

$$P(y_i|x_i) = \text{softmax}(W_o h_i^L + b_o) \tag{4.1}$$

where $W_o \in \mathbb{R}^{dim_h \times |\mathcal{Y}|}$ is a learnable parameter. Using the `BIOES` encoding as our tagging scheme, the possible values for tag y_i are `B-{POS, NEG, NEU}`, `I-{POS, NEG, NEU}`, `E-{POS, NEG, NEU}`, `S-{POS, NEG, NEU}`, or `O`, which respectively denote beginning of aspect, inside of aspect, end of aspect, and single-word aspect with possible sentiment values positive, negative, and neutral, as well as outside of aspect. As BERT relies on the WordPiece tokenization algorithm which may split words into several subwords, we use the first subword of a split word for predicting its tag.

The E2E-ABSA layer is trained using the cross-entropy loss function.

$$\begin{aligned}\mathcal{L}_{CE}(\hat{y}_i, y_i) &= -\log p(y_i|x_i) \\ &= -y \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \\ &= -[y \log \text{softmax}(W_o h_i^L + b_o) + (1 - y_i) \log(1 - \text{softmax}(W_o h_i^L + b_o))]\end{aligned} \tag{4.2}$$

where $\hat{y}_i$ refers to the classifier output and y_i to the correct label.

4.2 Multi-task Learning

As discussed in Chapter 2, ABSA subtasks are correlated to each other. Accordingly, there is a tendency to learn several tasks jointly since pipeline approaches tend to propagate errors and these tasks can benefit each other. Throughout this work, we consider E2E-ABSA for two reasons: i) for the sake of clarity and consistency when evaluating various data configurations, and ii) because E2E-ABSA is more relevant to the end user in the context of VisualCRM. E2E-ABSA can be considered as a form of multi-task learning in which two different tasks are learned simultaneously. To further investigate the adequacy of multi-task learning and combining related subtasks, we consider two other ABSA tasks: global polarity classification and joint aspect category detection (ACD) and ASC. The three tasks are learned simultaneously using a weighted loss function.

$$\mathcal{L} = \frac{1}{3} \cdot \mathcal{L}_{E2E-ABSA} + \frac{1}{3} \cdot \mathcal{L}_{polarity} + \frac{1}{3} \cdot \mathcal{L}_{ACD+ASC} \tag{4.3}$$

where each individual loss is the cross-entropy loss.

Global polarity classification seeks to assign a label y_c from a set of classes $C = \{\texttt{positive}, \texttt{negative}, \texttt{neutral}, \texttt{mixed}\}$ which corresponds to the overall polarity of an utterance. Formally, we reduce the contextual representations of all tokens $H^L = (h_1^L, ..., h_N^L) \in \mathbb{R}^{N \times dim_h}$ to a global representation $h_{glob} \in \mathbb{R}^{dim_h}$ via max-pooling.[1] We then pass this representation through a linear layer and compute the predicted label using a softmax activation function.

$$P(y_c|X) = \text{softmax}(W_o h_{glob} + b_o) \tag{4.4}$$

ACD can be treated either as a global task by identifying all aspect categories evoked in an utterance, or as a local task by detecting the specific aspect category

[1] We compared this method with predicting from a dedicated classification token [CLS] and found no difference in performance.

associated with an opinion target. We focus on the latter task. Accordingly, we treat joint ACD and ASC in a similar way to E2E-ABSA by formulating it as a sequence tagging task and combining boundary labels with task-specific labels. An example is shown in Table 4.1.

The	app	is	slow	but	I	like	the	new	interface	.
O	S-efficiency-neg	O	O	O	O	O	O	B-usability-pos	E-usability-pos	O

TABLE 4.1: Example sequence of labels for joint ACD and ASC.

Given an input sequence of tokens $X = (x_1, ..., x_N)$, we predict a sequence of labels $(y_1, ...y_N)$, as with E2E-ABSA. The difference is that the size of the tagset is much larger due to the number of aspect categories: $4mn + 1$, where m is the number of aspect categories[2] and n the number of polarity classes. Since this results in a complex task to learn, we simplify it by systematically assigning an aspect category label with polarity to each token in the predicted span boundaries from E2E-ABSA. In that way, we can reduce the size of the tagset to $mn + 1$ by using labels with no boundary indication, e.g. `efficiency-neg`.

In a production setting, multi-task learning is obviously advantageous since a single model can tackle many useful tasks at once. But we are also interested in finding out whether the benefits of multi-task learning can be extended beyond just two related tasks as with E2E-ABSA, and if we can improve the model's performance across all these tasks. Global polarity classification could encourage the model to learn a global representation of an utterance which would be beneficial to other tasks. Similarly, joint ACD and ASC, being a complex task, could force the model to learn how different components of an utterance interact together, and thus introduce more refined information. On the flip side, the combination of all these tasks could become too complex for the model to learn efficiently, and some trade-offs would have to be made to account for all this information.

[2] 30 for the restaurant domain in SemEval and 11 for the in-house datasets.

Chapter 5

Translation-based Adaptation

As discussed in Chapter 2, NMT can enable cross-lingual transfer by translating a reference corpus into a target language. In this chapter, we describe the method we use to generate synthetic data and project annotations for sequence tagging tasks.

5.1 Annotation Projection for Sequence Tagging

We use the Marian NMT toolkit (Junczys-Dowmunt et al., 2018)[1] to translate the source language corpora into the other languages, and vice versa. Marian NMT is open source and allows us to use 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]

We use English and French as source languages for SemEval and in-house datasets respectively. The choice of which language to translate from is based on the following considerations: 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 expect annotations for the English and French corpora to be of higher quality as they went through a more extensive review process. In the case of the in-house datasets, French data is also more plentiful.

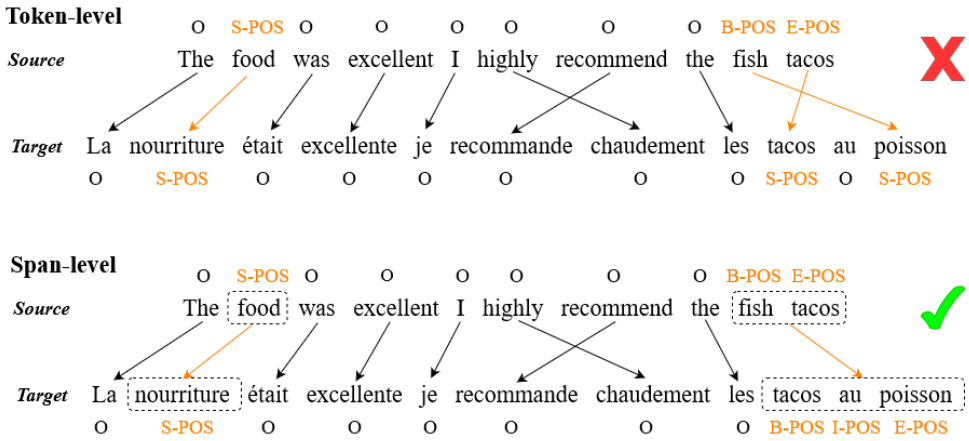


FIGURE 5.1: Different strategies for annotation projection using word alignments.

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

For sentence classification tasks, label projection is not a concern and the translated data can be used right away using the labels from the reference corpus. In contrast, sequence tagging datasets are annotated at the token level and require an additional pre-processing step since no word-to-word correspondence between source and target utterances is available. Such correspondence can be obtained using word alignment tools. One simple approach to annotation projection would then be 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 5.1. Consequently, we introduce a span-based alignment approach to project gold annotations at the span-level.

5.1.1 Span-based Alignment

Let $x^{src}_{1:N} = (x_1, ..., x_N)$ be the source utterance and $x^{trg}_{1:M} = (x_1, ..., x_M)$ the translated utterance. We obtain word alignments $a^{src}_{1:N} = (a_1, ..., a_N)$ 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^{src}_i. For each annotated span $x^{src}_{i:j}$ in the source utterance from the i-th word to the j-th word, we identify the projected span in the target utterance as $x^{trg}_{p:q}$, where

$$\begin{aligned} p &= min(a^{src}_{ij}) \\ q &= max(a^{src}_{ij}) \end{aligned} \tag{5.1}$$

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}`. This span-level alignment method prevents misprojections that could occur with a word-level alignment. Note that both methods are equivalent when source and target spans have the same number of words and the same order.

The tool of choice for obtaining word alignments is not guaranteed to be entirely accurate. First, the translations that the tool relies on are machine-generated and can therefore be inadequate. Second, as we have seen in Chapter 3, user-generated content 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.

Since our projection method uses the minimum and maximum indices of the aligned tokens as the offsets of the projected span, some gaps can appear in the projected spans where some of the translated tokens have not been aligned to an opinion target token in the source utterance. An example is shown in Figure 5.2. We call these tokens 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 as ill-formed and are filtered out of the translated dataset. The choice of this hyperparameter will be discussed in Chapter 6, where we provide an extrinsic evaluation of allowing one or more insertions per gap in the projected spans.

To obtain word alignments, we use SimAlign[3] (Jalili Sabet et al., 2020), a tool that leverages multilingual word embeddings for high quality word alignments. We refer the reader to the original publication for a comparison with other word alignment

[3]Using the Itermax algorithm.

Algorithm 1: Span-based alignment

```
function align (src_utterance, trg_utterance, src_spans, α)
a^src_{1:N} = get_word_alignment(src_utterance, trg_utterance)
trg_spans = []
for x^src ∈ src_spans do
    x^trg = []
    for i ∈ x^src do
        if i ∈ 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 > α then
        return False
    else
        x^trg = range(min(x^trg), max(x^trg))
        trg_spans.append(x^trg)
    end
end
return trg_spans
```

methods. Unlike statistical word aligners and attention-based alignments obtained from NMT models, this method has the advantage of not requiring parallel data as it is fully unsupervised. Note that with this method, a source token can be aligned with multiple target tokens and vice versa.

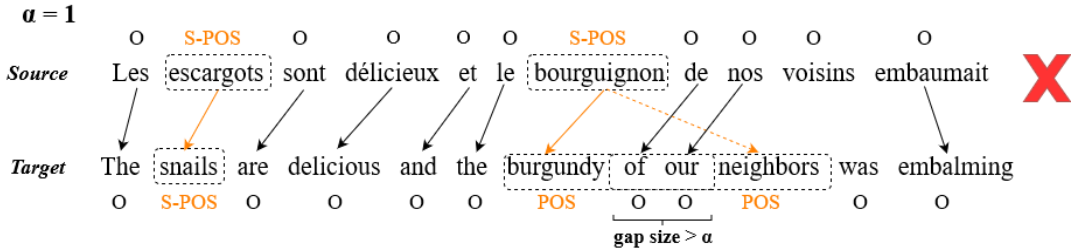


FIGURE 5.2: The word alignment tool can introduce fallacies. We use a heuristic based on gap size to filter out utterances that potentially contain an ill-formed projected span. The dashed line indicates an erroneous alignment.

5.2 Data Configurations

In this section, we describe the various data configurations that we compared in our experiments with a common multilingual PTM as the backbone architecture. We conducted experiments for two scenarios: cross-lingual adaptation and data augmentation.

5.2.1 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. This scenario can be considered as a low-resource setting. The question here is whether the original annotated data is sufficient to fine-tune a multilingual PTM and process unseen languages, or if translating this data into the other languages facilitates cross-lingual transfer, a process illustrated in Figure 5.3. In other words, we would like to find out whether multilingual word representations are truly language agnostic or not. A summary of the different data configurations is shown in Table 6.5. **O** and **Tr** respectively denote original and translated data. **S** and **T** refer to source and target languages.

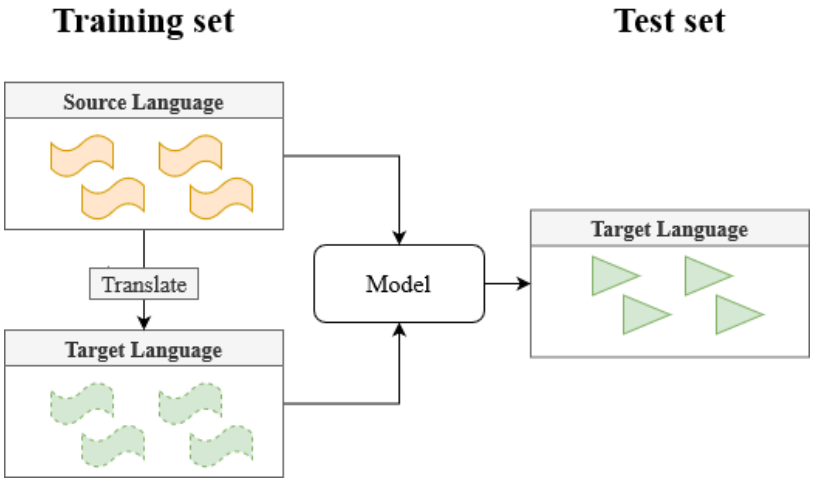


FIGURE 5.3: An illustration of cross-lingual adaptation ($\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow T}$). The dotted lines denote the synthetic dataset obtained with our translation-based adaptation method.

$\mathbf{O}_S$ (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 an unseen target language. This configuration serves as the baseline method for cross-lingual adaptation

$\mathbf{Tr}_{S\rightarrow T}$ 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.

$\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow T}$ 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. This configuration should tell us if the original source language data is important.

$\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow others}$ To assess the relevance of specifically fine-tuning on the target language, we use a combination of the source data and its translations into all the other languages except the target language to be evaluated.

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 (en/fr)**	**Target**	**Others**
O_S (Zero-shot)	✓		
$Tr_{S \to T}$		✓	
O_S + $Tr_{S \to T}$	✓	✓	
O_S + $Tr_{S \to others}$	✓		✓
O_S + $Tr_{S \to all}$	✓	✓	✓

TABLE 5.1: Data configurations for cross-lingual adaptation.

5.2.2 Data Augmentation

Data augmentation refers to a high-resource setting where annotated data is available in all the languages we wish to process. Multilingual PTMs can be fine-tuned on a multilingual corpus to make use of as much supervision as possible. We are interested in studying how our translation-based adaption approach will affect the model's performance if we augment the original corpora with the translated data, as shown in Figure 5.4. Table 5.2 summarizes the different data configurations.

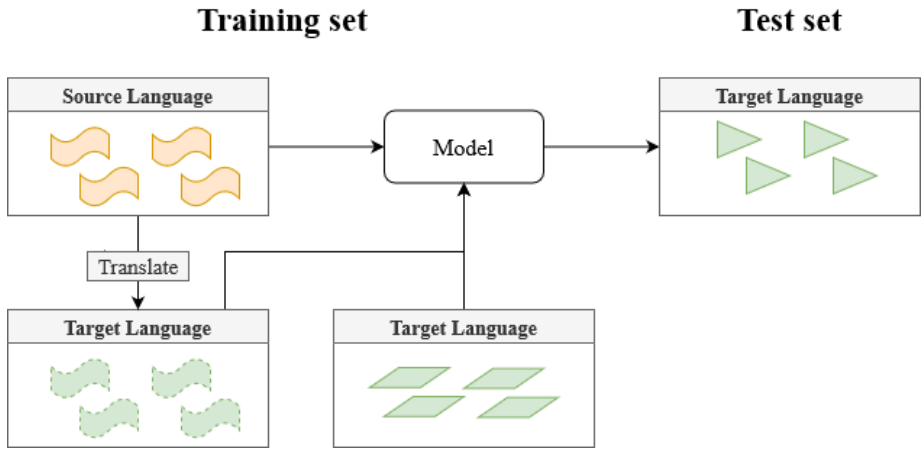


FIGURE 5.4: An illustration of data augmentation (EN/FR → OTHERS). The dotted lines denote the synthetic dataset obtained with our translation-based adaptation method.

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. A multilingual PTM should benefit from this additional data.

O$_{all}$ **+ Tr**$_{S \to all}$ We fine-tune a model on the concatenation of the original data in all languages with the translated data from either English for SemEval or French for the in-house datasets into the other languages. With this configuration, we evaluate the relevance of our approach for data augmentation in a high-resource setting.

O$_{all}$ **+ Tr**$_{all \to S}$ We also experiment with translating the corpora from all the other languages into English for SemEval or French for the in-house datasets. With this configuration, we would like to understand how different translated corpora affect 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 $\to$ all**	**all $\to$ 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 5.2: Data configurations for data augmentation.

Chapter 6

Experiments

This chapter describes the experiments that we conducted. In the first set of experiment, we only take into account the original corpora and explore generic models through multi-task learning and combining heterogeneous datasets. The second set of experiment is dedicated to our translation-based adaptation method. We study the relevance of using the synthetic data obtained through this method in two scenarios: Cross-lingual adaptation, where annotated data is available in only one language, and data augmentation, where corpora are available in multiple languages and we seek to improve performance by augmenting the datasets.

6.1 Experimental Setup

If not stated otherwise, the same setup is used throughout our experiments. We use mBERT as our multilingual PTM initialized with pre-trained weights[1] available from HuggingFace Transformers (Wolf et al., 2020). For each configuration, a model is fine-tuned with a linear classification layer added on top of mBERT's architecture to predict a label for each token. We train each model for a maximum of 50 epochs with a batch size of 128. We use the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate set to $5e^{-5}$. We use the `BIOES` encoding as our tagging scheme. [2] In the case of E2E-ABSA, 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. We ran our experiments on a GeForce RTX 3090 GPU with 24 GB memory. It takes around one hour to fine-tune a model with this configuration.

6.2 Original Data

In this section, we experiment with the original corpora only and consider ways of combining different tasks and domains. As original data, we take the concatenation of the datasets in all the languages to learn a single multilingual model.

6.2.1 Multi-task Learning

As we discussed in Chapter 4, E2E-ABSA is a form of multi-task learning which performs OTE and ASC simultaneously. In previous works, BERT-based models have been successfully applied to E2E-ABSA (Li et al., 2019b). In this experiment, we are interested in extending this multi-task approach with two other related tasks: global sentiment classification, and joint ACD and ASC. Learning all these correlated tasks

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

[2] We compared various tagging schemes and found that `BIOES` provided the best results.

Task	Setting	ar	en	es	fr	nl	ru
Polarity	**Monotask**	79.5	72.9	73.8	67.8	70.4	71.6
	Multi-task	**80.4**	**73.3**	**74.7**	**67.8**	**71.6**	**71.7**
E2E-ABSA	**Monotask**	61.6	64.3	67.8	52.0	**60.3**	60.8
	Multi-task	**61.9**	**64.5**	**67.8**	**53.1**	60.1	**61.8**
ACD + ASC	**Monotask**	8.7	12.4	14.3	6.8	10.1	12.3
	Multi-task	**10.7**	**30.6**	**29.2**	**12.2**	**26.4**	**27.1**

TABLE 6.1: Results of multi-task learning on SemEval datasets.

simultaneously has the potential of increasing the model's sensitivity to various elements related to ABSA. As this results in a more complex task to estimate, we increase the maximum number of epochs to 100 to allow the model to converge to a local minimum. In the case of global sentiment classification, we report the F_1 score on correct polarity labels. Joint ACD and ASC is evaluated in a similar way to E2E-ABSA with exact matches of span boundaries and labels, and results are reported using the micro F_1 score.

Table 6.1 and Table 6.2 show the results for SemEval and in-house datasets respectively, comparing the performance of each task learned individually (Monotask) vs. learned simultaneously with the other tasks (Multi-task). For both sets of corpora, multi-task learning provides a consistent improvement across all tasks. The in-house datasets which are smaller especially benefit from this transfer learning approach. Note that joint ACD and ASC in the monotask setting includes boundary labels increasing its tagset size, which explains the low results.

Task	Setting	ar	en	es	fr	nl	pl	ro
Polarity	**Monotask**	78.0	84.8	94.8	87.7	90.3	88.0	80.1
	Multi-task	**79.3**	**84.5**	**95.0**	**89.0**	**91.9**	**89.4**	**80.2**
E2E-ABSA	**Monotask**	48.0	54.0	54.8	56.8	64.5	56.4	41.3
	Multi-task	**51.2**	**60.4**	**59.4**	**57.1**	**66.5**	**58.2**	**42.8**
ACD + ASC	**Monotask**	7.1	6.1	4.1	3.5	5.0	6.3	2.9
	Multi-task	**24.4**	**20.0**	**18.1**	**25.7**	**28.5**	**20.5**	**12.8**

TABLE 6.2: Results of multi-task learning on the in-house datasets.

These results show the relevance of multi-task learning for ABSA. As an added bonus, the multi-task model has the advantage of tackling many tasks at once, which is desirable in a production setting. In the following experiments, we turn back to E2E-ABSA only, looking at how various data configurations affect performances on this task.

6.2.2 Combining SemEval and In-house Datasets

The data analysis from Chapter 3 revealed that SemEval and in-house datasets show marked dissimilarities. The main difference is that they pertain to two distinct domains, i.e. restaurant/hotel reviews and mobile application reviews. Nonetheless, both sets of corpora were annotated in a similar way. To find out whether a single

	SemEval						In-house						
Data	**ar**	**en**	**es**	**fr**	**nl**	**ru**	**ar**	**en**	**es**	**fr**	**nl**	**pl**	**ro**
Separate	61.6	64.3	**67.8**	52.0	60.3	60.8	48.0	54.0	54.8	56.8	64.5	**56.4**	41.3
Combined	**62.5**	**67.4**	67.2	**53.1**	**61.2**	**61.6**	**49.2**	**58.3**	**57.7**	**57.8**	**66.2**	55.4	**41.6**

TABLE 6.3: Results of combining SemEval and in-house datasets for E2E-ABSA.

model is capable of handling different domains, we fine-tune a model on the concatenation of SemEval and in-house datasets and compare it to the single-domain baseline.

The results can be found in Table 6.3. The combined model brings a considerable improvement across most languages. This shows not only that two different domains can be tackled by one model, but also that out-of-domain data is beneficial for E2E-ABSA. We expect opinion targets to follow similar syntactic structures. One possible explanation for these results is that information learned from similar syntactic structures is abstracted away from the domain and transferred adequately to both SemEval and in-house datasets.

6.3 Translation-based Adaptation

This section concerns the experiments where we tested our translation-based adaptation method in both cross-lingual adaptation and data augmentation scenarios.

6.3.1 Setting the Number of Allowed Insertions

In our span-based projection method, the number of allowed insertions per gap in a projected span is considered as a hyperparameter α. To establish what is the optimal value for this hyperparameter on our task, we conduct a preliminary experiment with extrinsic evaluation. Using the translations of the French in-house dataset in the other languages, we create different sets of synthetic corpora by projecting the annotations with different values for α. We also evaluate our projection method when allowing any number of insertions to assess the relevance of this filtering approach. We fine-tune these models in the cross-lingual adaptation scenario with the $\mathbf{O}_S$ + $\mathbf{Tr}_{S \to all}$ configuration. We also compare SimAlign with `fast_align`, a popular word alignment tool. The results are shown in Table 6.4.

Word alignments	α	**ar**	**en**	**es**	**fr**	**nl**	**pl**	**ro**	**avg.**	**% out**
fast_align	**>2**	24.9	38.7	31.7	50.6	53.8	22.0	19.8	34.5	0
	2	28.0	41.8	37.8	54.1	55.7	31.0	28.6	39.6	12.9
	1	27.9	41.2	35.1	52.5	**56.1**	30.1	26.1	38.4	13.3
	0	28.9	40.9	34.3	52.7	54.9	29.3	25.6	38.1	15.0
SimAlign	**>2**	22.8	31.4	29.5	39.1	37.9	26.2	23.3	30.0	0
	2	32.3	43.1	**38.8**	52.4	51.1	36.9	**32.1**	41.0	4.8
	1	31.9	42.4	38.7	53.2	54.5	**37.7**	31.5	**41.4**	5.3
	0	**33.3**	**43.4**	38.7	**54.5**	49.8	35.8	30.9	40.9	11.4

TABLE 6.4: Study of the impact of constraining the number of insertions within projected opinion targets during label alignment. Results for the $\mathbf{O}_S$ + $\mathbf{Tr}_{S \to all}$ configuration on the in-house datasets.

By constraining the number of allowed insertions per gap, we observe a significant improvement in performance compared to not filtering any of the translated utterances. This validates our filtering approach and shows that not all translated utterances are useful. We also obtain better results with SimAlign. Consequently, when projecting annotations in the following experiments, we use SimAlign to obtain word alignments and we set α to 1, i.e. the value with the highest average score.

6.3.2 Cross-lingual Adaptation

Config.	ar	es	fr	nl	ru	en
$\mathbf{O}_S$	13.2	51.2	35.1	38.2	34.9	62.7
$\mathbf{Tr}_{S\rightarrow T}$	7.6	25.8	37.6	24.0	29.1	_
$\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow T}$	31.4	54.3	40.7	49.4	47.7	_
$\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow others}$	18.2	52.6	40.9	43.9	38.2	_
$\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow all}$	**34.7**	54.8	41.5	**51.0**	**47.9**	**64.1**
(Li et al., 2020)	_	**58.2**	**46.9**	49.9	44.9	_

TABLE 6.5: Results of cross-lingual adaptation on the SemEval datasets.

The results for cross-lingual adaptation are shown in Table 6.5 and Table 6.6 for the SemEval and in-house datasets respectively. 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 $\mathbf{O}_S$. We observe that zero-shot cross-lingual transfer is least efficient for distant languages (e.g. Russian and Arabic). For more closely related languages (e.g. Spanish and Dutch), the results are closer to those obtained with translation-based adaptation models. For SemEval, the gains observed for Arabic despite the domain shift (hotel reviews) demonstrate the usefulness of fine-tuning a model on synthetic data in the target language and the model's capacity to handle different domains.

Config.	ar	en	es	nl	pl	ro	fr
$\mathbf{O}_S$	13.5	23.9	24.8	33.8	19.3	15.4	48.6
$\mathbf{Tr}_{S\rightarrow T}$	24.5	30.3	31.5	38.8	31.9	27.2	_
$\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow T}$	10.3	42.6	38.7	**55.6**	34.2	29.9	_
$\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow others}$	30.3	38.4	37.8	49.8	**37.2**	28.2	_
$\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow all}$	**31.3**	**43.3**	**39.4**	55.5	35.6	**31.2**	**51.4**

TABLE 6.6: Results of cross-lingual adaptation on the in-house datasets.

$\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow others}$ provides an improvement over $\mathbf{O}_S$, but is limited compared to $\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow T}$ and $\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow all}$, confirming that multilingual PTMs benefit from being fine-tuned on the target data specifically. In most cases, $\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow T}$ and $\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow all}$ configurations come relatively close. However, $\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow all}$ can be considered as superior as a unique model can be applied to all languages. $\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow T}$ provides an improvement over $\mathbf{Tr}_{S\rightarrow T}$, confirming the importance of the original source language data. We also observe that with $\mathbf{O}_S$ + $\mathbf{Tr}_{S\rightarrow all}$, the translated data contributes to an improvement in performance for the source language, i.e. English

or French, highlighting the relevance of the translation-based adaptation method for data augmentation.

For comparison, we also include the results obtained by Li et al., 2020 on SemEval. Their approach consists in warming up the model's parameters using the translated data and then applying $\mathbf{O}_S$. While the results between their approach and ours are similar, our approach has the advantage of being conceptually simpler and easier to implement.

6.3.3 Data Augmentation

The results for data augmentation are shown in Table 6.7 and Table 6.8 for the SemEval and in-house datasets respectively.

Config.	ar	en	es	fr	nl	ru
$\mathbf{O}_T$	61.1	62.7	58.9	34.4	52.3	35.5
$\mathbf{O}_{all}$	61.1	64.3	67.1	51.9	59.6	60.4
$\mathbf{O}_{all}$ + $\mathbf{Tr}_{S\rightarrow all}$	61.6	**65.1**	67.7	51.8	61.8	60.1
$\mathbf{O}_{all}$ + $\mathbf{Tr}_{all\rightarrow S}$	**61.9**	55.9	**68.0**	52.2	60.0	61.0
$\mathbf{O}_{all}$ + $\mathbf{Tr}_{S\leftrightarrow all}$	61.3	57.5	66.9	**52.4**	**62.4**	**61.6**

TABLE 6.7: Results of data augmentation on the SemEval datasets.

When comparing $\mathbf{O}_T$ to $\mathbf{O}_{all}$, we observe a significant improvement when fine-tuning a model on the combination of all languages, highlighting the efficacy of multilingual PTMs.

For SemEval, the augmented data provides a consistent improvement over $\mathbf{O}_{all}$. Using the translations of the non-English corpora into English is detrimental to the performance on the English test set, while other languages tend to benefit from this translation direction. Overall, $\mathbf{O}_{all}$ + $\mathbf{Tr}_{S\rightarrow all}$ is the configuration that brings the most consistent improvements across 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.

Config.	ar	en	es	fr	nl	pl	ro
$\mathbf{O}_T$	10.3	42.6	38.7	48.6	55.6	34.2	29.9
$\mathbf{O}_{all}$	48.0	54.0	54.8	**56.8**	64.5	**56.4**	41.3
$\mathbf{O}_{all}$ + $\mathbf{Tr}_{S\rightarrow all}$	**48.5**	53.2	56.0	55.9	61.6	54.5	41.0
$\mathbf{O}_{all}$ + $\mathbf{Tr}_{all\rightarrow S}$	46.9	54.1	**56.2**	54.1	64.1	55.3	40.8
$\mathbf{O}_{all}$ + $\mathbf{Tr}_{S\leftrightarrow all}$	48.4	**54.9**	54.2	55.2	**64.6**	54.6	**42.0**

TABLE 6.8: Results of data augmentation on the in-house datasets.

6.4 Putting It All Together

The underlying goal of this work was to research generic approaches to ABSA that can address different domains and languages at once. While we mostly focused on E2E-ABSA for clarity, multi-task learning is also desirable to tackle other ABSA subtasks. In this last experiment, we combine the insights from our exploration into

a final model. Specifically, we would like to uncover if the benefits of the translation-based adaptation method also hold in multi-task, multi-domain settings. As a kind of ablation study, we use a baseline model and gradually add components that were beneficial individually.

The results for cross-lingual adaptation on E2E-ABSA are shown in Table 6.9. We use $\mathbf{O}_S$ (Zero-shot) as the baseline configuration. Combining the source language data with its translations into the other languages provides a significant increase in performance as we have seen before. Applying multi-task learning with two additional ABSA subtasks yields a consistent improvement across all languages. It is especially beneficial for the in-house datasets. Finally, concatenating the original source language data in both English and French as well as all the translations is useful to almost all languages, showing how larger and more diverse datasets can benefit the model.

	SemEval						In-house						
Config.	**ar**	**es**	**fr**	**nl**	**ru**	**en**	**ar**	**en**	**es**	**nl**	**pl**	**ro**	**fr**
$\mathbf{O}_S$	13.2	51.2	35.1	38.2	34.9	62.7	13.5	23.9	24.8	33.8	19.3	15.4	48.6
+ Translations	34.7	54.8	41.5	**51.0**	47.9	64.1	31.3	43.3	39.4	55.5	35.6	31.9	51.4
+ Multi-task	34.2	55.2	42.9	**51.0**	49.8	63.8	35.5	47.2	40.9	58.3	**40.0**	**33.8**	53.6
+ Combining	**39.1**	**57.0**	**44.2**	50.7	**51.0**	**64.6**	**38.2**	**50.4**	**41.5**	**58.9**	39.8	33.5	**57.1**

TABLE 6.9: Combining different tasks, domains, and languages into one model in the cross-lingual adaptation setting. F_1 score for E2E-ABSA.

In the case of data augmentation, we use $\mathbf{O}_{all}$ as the baseline model. The results can be seen in Table 6.10. We first increment this model with multi-task learning, which provides a steady improvement. Again, the in-house datasets especially benefit from multi-task learning. One possible explanation is that this set of corpora is smaller and thus benefits the most from transfer learning. The next configuration adds the combined SemEval and in-house datasets. The gains are modest and in the case of the in-house datasets, this approach is sometimes detrimental. The final model includes all original and translated data ($\mathbf{O}_{all}$ **+** $\mathbf{Tr}_{S \rightarrow all}$). As for the data augmentation experiment, the SemEval datasets benefit from the augmented data, while the results are mixed for the in-house datasets.

	SemEval						In-house						
Config.	**ar**	**en**	**es**	**fr**	**nl**	**ru**	**ar**	**en**	**es**	**fr**	**nl**	**pl**	**ro**
$\mathbf{O}_{all}$	61.1	64.3	67.1	51.9	59.6	60.4	48.0	54.0	54.8	56.8	64.5	56.4	41.3
+ Multi-task	61.9	64.5	67.8	53.1	60.1	61.8	51.2	60.4	**59.4**	57.1	**66.5**	**58.2**	42.8
+ Combining	61.8	65.6	68.7	52.3	60.0	62.0	50.8	58.0	58.1	**59.4**	**66.5**	56.4	42.8
+ Translations	**62.1**	**65.8**	**69.3**	**53.7**	**60.3**	**64.6**	**51.2**	**61.1**	56.0	58.50	64.5	56.1	**44.2**

TABLE 6.10: Combining different tasks, domains, and languages into one model in the data augmentation setting. F_1 score for E2E-ABSA.

Impressively, these results show that we can successfully combine different tasks, domains, and languages into one model, an ideal approach in a production setting.

Chapter 7

Analysis

In line with the results from the previous chapter, we conduct a refined analysis to shed some light on several aspects of this work such as the amount of annotated data needed and dealing with noisy user-generated text.

7.1 To Annotate, or not to Annotate

The process of annotating resources in various languages is costly and can be error-prone. The variety of languages means that this process has to be outsourced. The quality of the resulting annotations is difficult to evaluate with no knowledge of the target language. In light of the results from Chapter 6, one question arises: is it sufficient to use an annotated corpus in a reference source language that we expect to be more reliable, or is it necessary to annotate resources in multiple languages? To further highlight the relevance of our translation-based adaptation method, we compare the best results obtained with the final models from Section 6.4 for both cross-lingual adaptation and data augmentation scenarios, i.e. the low-resource vs. high-resource settings.

	ar	**es**	**fr**	**nl**	**ru**	**en**
Low-resource	39.06	57.00	44.18	50.75	50.96	64.60
High-resource	62.07	69.30	53.68	60.34	64.58	65.77

TABLE 7.1: Comparing different resource settings on the SemEval datasets. F_1 score for E2E-ABSA.

The comparisons for SemEval and in-house datasets can be seen in Table 7.1 and Table 7.2 respectively. The results for the source languages, i.e. English and French, are shown separately since original annotated data is available for these languages in both settings. In both sets of corpora, we observe that the resource-poor setting provides reasonable performances despite the absence of gold annotated data in the target language. However, the resource-rich setting yields a considerable improvement in performance, increasing the F_1 score by around 10 to 20 points depending on the language. This gap is especially marked for languages that are more distant from the source language, e.g. Arabic, Russian, and Polish. Regarding the source languages, having access to more annotated data in other languages seems to be beneficial, bringing a modest increase in performance.

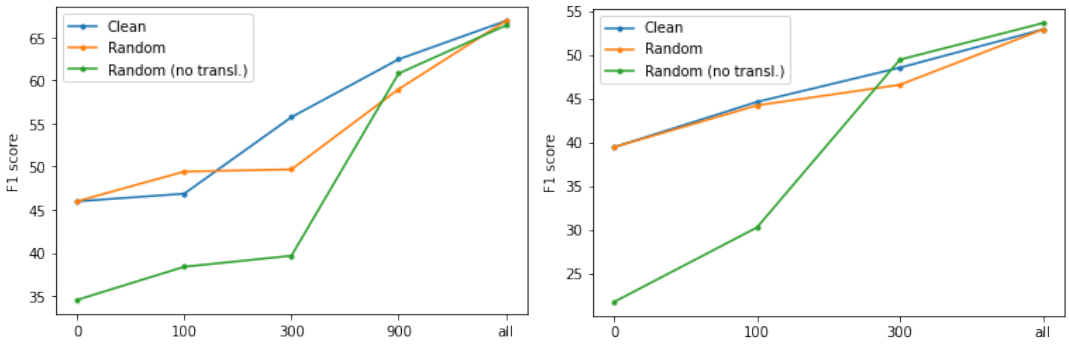
In light of this comparison, one could then wonder how much annotation in the target languages is necessary to bridge the remaining gap between low- and high-resource settings. To answer this question, we conducted a follow-up comparison of different resource settings with increasing amounts of annotated examples in the

	ar	en	es	nl	pl	ro	fr
Low-resource	38.2	50.4	41.5	58.9	39.8	33.5	57.1
High-resource	51.2	61.1	56.0	64.5	56.1	44.2	58.5

TABLE 7.2: Comparing different resource settings on the in-house datasets. F_1 score for E2E-ABSA.

target languages. To select these examples, we compared random sampling with a clean sampling method, i.e. for each language, we compute the SER for all examples and select the n examples with the lowest SER. We also compare fine-tuning on original annotated data only vs. adding the translated data in the target languages from the full source language corpus. The results for E2E-ABSA are shown in Figure 7.1. For simplicity, we averaged the results across all languages.

We find that in low-resource settings, i.e. 0 to 300 annotated examples in the target languages, the translated data is largely beneficial, showing how translation-based adaptation can be combined effectively with a small amount of annotated examples. With more annotated data, this benefit is less marked. At least up to the full annotated corpora, the results improve linearly with increasing amounts of annotated data, highlighting the importance of the annotation process. When comparing the two sampling strategies, we observe that clean sampling provides an improvement over random sampling. This reflects how noisy text with its deviations from standard form tends to lean towards a more random distribution which is difficult to model. Accordingly, this sampling method can help guide the selection of examples to annotate during the initial phase of the annotation process.



(A) SemEval datasets

(B) In-house datasets

FIGURE 7.1: Comparing different resource settings with different sampling strategies. Averaged F_1 score across languages for E2E-ABSA.

To sum up, we have explored ways of fully exploiting the available annotated data in different resource settings. When no annotated data is available in a target language, zero-shot cross-lingual transfer can be used to tackle this language. However, as we have seen in Chapter 6, this approach can be limited. Thanks to our translation-based adaptation method, the gap between resource-poor and resource-rich settings can be reduced, and small amounts of annotated data in the target languages can boost the results. All in all, whether to annotate or not depends on the needs. If high performance is crucial, then annotating datasets for other languages would be preferred, and it seems the more annotated examples the better. Otherwise, adapting available resources to a target language is a satisfactory approach.

7.2 Processing Noisy Text

When applied to the SemEval datasets, data augmentation in the resource-rich setting yields a consistent improvement in performance. For the in-house datasets however, the translated data can be detrimental for some languages. In Chapter 3, an analysis of the amount of noise in the data compared both sets of corpora and revealed that the in-house datasets are comparatively more noisy, i.e. they contain more spelling errors. Hence, the hypothesis we make is that this relatively higher noise in the data could be perturbing the translations, and thus negatively impacting the rest of the pipeline. Similarly, the in-house datasets contain a more specialized vocabulary, e.g. brand names and product names, which is also challenging for the NMT model. To confirm our intuition, we conduct two additional experiments that aim at reducing this perturbation from noisy text. We focus on the case where our translation-based adaptation method failed to improve the baseline results, i.e. data augmentation on the in-house datasets with the $\mathbf{O}_{all}$ + $\mathbf{Tr}_{S\rightarrow all}$ configuration.

Follow-up experiment 1: Filtering noisy utterances Considering the SER as an indicator of noisy text, we can filter out the most noisy utterances from the French reference corpus before translating it into the other languages. We compute the SER 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 7.3.

	ar	**en**	**es**	**fr**	**nl**	**pl**	**ro**	**avg.**
No filter	**48.3**	54.9	54.5	54.6	63.2	**55.4**	41.1	53.1
Filter 5%	47.4	**55.3**	**55.6**	55.5	**64.0**	54.7	**41.7**	**53.4**
Filter 10%	46.8	53.2	54.5	**59.2**	63.3	54.8	40.9	53.2
Filter 20%	47.4	54.5	52.6	55.6	63.8	53.1	38.3	52.2

TABLE 7.3: Study of filtering out the n percent utterances with the largest SER from the French corpus before translating it into the other languages. Results for E2E-ABSA with data augmentation on the in-house datasets with the $\mathbf{O}_{all}$ + $\mathbf{Tr}_{S\rightarrow all}$ configuration.

We observe a slight increase in performance when filtering out the 5% and 10% 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. This first follow-up experiment seems to support our hypothesis. As a consequence, we conducted another experiment to explore this issue further.

Follow-up experiment 2: Text normalization If noisy user-generated text is indeed what is perturbing the NMT model and consequently the rest of the pipeline, could we automatically reduce noise from the data before translating it? Would results on the downstream task be affected by such an approach? To answer these questions, we conduct another follow-up experiment focusing on the automatic correction of the French in-house corpus.

A basic approach we tried first was to apply a spellchecker (Hunspell) on each utterance for a word-by-word correction. This approach has the advantage of preserving token-level annotations. However, a qualitative analysis of the corrected utterances revealed that such an approach was not adequate as it introduced more noise in the data than what was already present.

Modern GEC models have proven to be quite accurate for correcting erroneous text. Several GEC models for English have been released publicly (Choe et al., 2019; Omelianchuk et al., 2020). However, no GEC model for French is available. This is mostly due to the lack of large scale annotated GEC resources for French. Katsumata and Komachi, 2020 showed that large pre-trained seq2seq models such as BART (Lewis et al., 2020) could be fine-tuned effectively for GEC, thus leveraging transfer learning and reducing the amount of parallel data needed. Following this work, we fine-tuned our own GEC model for French based on a French adaption of BART called BARThez (Eddine, Tixier, and Vazirgiannis, 2020).

While GEC resources for French are scarce, some can be found. The WiCoPaCo corpus (Max and Wisniewski, 2010) is a publicly available dataset of corrections and paraphrases in French retrieved from the Wikipedia revision history. The authors released a subset of this corpus containing corrections of misspelled words only using two heuristics to filter out unwanted modifications (Wisniewski, Max, and Yvon, 2010). This subset contains 138,875 entries. In addition, another dataset of 8443 gold annotated parallel erroneous-corrected French examples from conversations between customers and assistants is available in-house.

Following other GEC works which first train a model on synthetic data and then fine-tune on gold data (Kiyono et al., 2019), we conducted a first stage of fine-tuning of BARThez on the subset of the WiCoPaCo corpus, followed by a second stage on the in-house dataset. For details on fine-tuning parameters, the interested reader can refer back to the original BARThez publication on which we based our parameters. We evaluated this model on a held-out test set using ERRANT (Bryant, Felice, and Briscoe, 2017). Results for the different stages can be seen in Table 7.4.

	P	**R**	**$F_{0.5}$**
Stage 0	0.4	0.9	0.4
Stage I	34.3	24.5	31.8
Stage II	71.5	57.5	68.2

TABLE 7.4: Performance of BARThez on the in-house GEC test set after each fine-tuning stage. Stage 0 denotes the results obtained with the off-the-shelf model.

Using the model from stage II, we are able to automatically correct the French in-house ABSA dataset. Since this is a seq2seq model, the number of tokens in the predicted correction can be modified. An additional pre-processing step is required to preserve token-level annotations. To this end, we apply our adaptation method described in Chapter 5 to project span-level annotations from the original utterance to the corrected one. This results in around 1% of filtered utterances for $\alpha = 1$, indicating that the projection is reliable. Finally, we apply our usual pipeline of translating from the corrected French corpus into the other languages, and then leveraging the augmented data to fine-tune a model for ABSA. The results are shown in Table 7.5.

We observe no performance improvement compared to the uncorrrected baseline approach. These results suggest two possible explanations:

- Our GEC model does not perform well enough on the ABSA dataset. While the results on the held-out test set are quite good (conversational assistance), the ABSA dataset pertains to another domain, i.e. written reviews.

- The NMT model struggles with the in-house datasets due to its specific lexicon and colloquial language, regardless of spelling errors. This results in subpar translations which are not helpful in the data augmentation scenario.

	ar	**en**	**es**	**fr**	**nl**	**pl**	**ro**
Not corrected	**48.3**	**54.9**	**54.5**	54.6	63.2	55.4	41.1
Corrected	47.1	54.6	52.3	**54.7**	**63.4**	**55.8**	**41.3**

TABLE 7.5: Study of automatically correcting the French corpus before translating it into the other languages. Results are for data augmentation on the in-house datasets with the $\mathbf{O}_{all}$ + $\mathbf{Tr}_{S \rightarrow all}$ configuration.

In the next section, we turn to a qualitative analysis of filtered utterances when translating and projecting from the French in-house corpus to shed some light on the translation of noisy text.

7.3 Analysis of the Filtering Approach

In Chapter 5, we introduced a filtering heuristic to prevent misalignments when projecting annotations. The results of our experiments showed the relevance of this filtering approach, with around 5% of the translated utterances filtered. This section shows how our translation-based adaptation method performs on noisy user-generated text. We sampled some of the filtered utterances when projecting annotations from the French in-house corpus to its translation in English. Table 7.6 shows four of these examples, with comments on what went wrong in the process and why these utterances were filtered.

Source text	**Translated text**	**Comment**
cette appli m' enerve	this app pisses me off .	A wrong word alignment leads to a misprojection.
Appli très performante et agréable à naviguer simple et efficace	Very efficient and pleasant to navigate simple and efficient	"Appli" ignored by the NMT model.
Application vraiment bidon :/ à quoi sert t elle mis à part nous dire quel est notre forfait !!? Aucun suivi conso !! Je ne recommande absolument pas !!!! Dev : votre application est à revoir dans sont intégralité !!!	Application really fake :/ What is it for besides telling us what our package is !!? No follow - up conso !! I absolutely do not recommend !!!! Dev : your application is to be reviewed in are full !!!	Limitation of the filtering heuristic: "follow - up conso" is a correct opinion target but is filtered because of a gap size > 1.
Pas de probleme ! Ca marche bien service au rendez vous	No problem , it works well on the date .	Colloquial language difficult to translate leads to omitted opinion target.

TABLE 7.6: Analysis of filtered examples when adapting the French in-house corpus to English. Correct and incorrect opinion targets are underlined with straight dotted lines respectively. Wrong word alignments are shown in red.

Through these examples, we can see that the NMT model handles noisy text somewhat satisfactorily, e.g. text with missing diacritics is translated correctly. However, colloquial language and specific lexicon such as "Appli" and "conso" pose a problem and can lead to subpar translations. The opinion target may be omited,

or the meaning of the source utterance may not be rendered properly. We also observe that colloquial language can produce wrong word alignments, e.g. "appli" gets aligned to both "app" and "off".

All in all, this analysis shows the difficulty of processing noisy user-generated text. Though our filtering approach can prevent fallacies to some extent, errors are bound to appear and produce undesired results through the multiple processing steps of translating, aligning, and encoding. This explains why translation-based adaptation in the data augmentation scenario might not have been as successful for the in-house datasets as for the relatively cleaner SemEval datasets.

7.4 Syntactic Analysis

In Chapter 6, we have seen how SemEval and in-house datasets could be combined into a single model. Not only is this desirable as a generic approach in a production setting, but performances were also improved despite the domain shift. Our intuition is that opinion targets tend to appear in similar syntactic structures regardless of the domain. In this section, we conduct a syntactic analysis to test this hypothesis and help explain how these two sets of corpora with distinct domains can be combined and can benefit each other.

Using SpaCy (Honnibal et al., 2020), we analyzed the syntactic categories of opinion targets based on i) their part-of-speech (POS) tag, and ii) their POS tag combined with their dependency label. The first analysis serves as a broad overview of opinion targets, while the second will provide a more refined analysis of what roles opinion targets typically play in a sentence. We carried out this analysis across English, French, Spanish, and Dutch, languages that are structurally similar and that are common to both SemEval and in-house datasets.

POS Tag				**POS Tag + Dependency Label**			
SemEval		**In-house**		**SemEval**		**In-house**	
NOUN	62.5	NOUN	38.7	NOUN nsubj	20.3	NOUN nsubj	9.9
PROPN	5.0	PROPN	10.2	NOUN ROOT	10.0	NOUN ROOT	7.2
ADJ	3.1	VERB	5.8	NOUN conj	7.8	NOUN obj	6.9
VERB	2.7	N_ADP_N	3.2	NOUN nmod	5.4	NOUN nmod	3.6
N_ADP_N	2.2	ADJ	3.1	NOUN obj	4.1	PROPN nsubj	3.6

TABLE 7.7: Five most common syntactic categories for opinion targets with coverage in percent. N_ADP_N stands for NOUN_ADP_NOUN, e.g. "brownie with ginger".

The five most common syntactic categories are shown in Table 7.7. Looking at POS tags, we can see that a majority of opinion targets are nouns, regardless of the domain. It is interesting to note the higher proportion of proper nouns in the in-house datasets, where more brand and product names tend to appear. The most common POS tags combined with dependency labels are also similar between SemEval and in-house datasets. Opinion targets typically appear in noun phrases as e.g. subjects, objects, or nominal modifiers. This analysis suggest that based on similar syntactic structures common to both domains, the E2E-ABSA model is able to extract high-level properties that apply to any ABSA domain. This is another example of the usefulness of transfer learning, extracting information from one domain and applying it to another.

7.5 Probing mBERT Layers

To investigate how multilingual word representations are affected by translation-based adaptation, we conduct a probing analysis of mBERT's representations across its twelve layers. More specifically, we compare $\mathbf{O}_S$ to $\mathbf{O}_S$ + $\mathbf{Tr}_{S \to T}$ fine-tuned for E2E-ABSA on the English SemEval dataset. We then use these models for inference on the English test set and its translation in the target language, and retrieve the encoded representations of each utterance at each layer. Each of these representations $H^L = (h_1^L, ..., h_N^L) \in \mathbb{R}^{N \times dim_h}$ is reduced to $\mathbb{R}^{dim_h}$ using max pooling. For each model, we then compute the Pearson correlation coefficient between the pooled representations of each English utterance and its translation into the target language. We compare French and Dutch, two languages closely related to English, with Russian and Arabic, two more distant languages. Figure 7.2 shows this comparison, where we averaged the Pearson correlation coefficients for visualization.

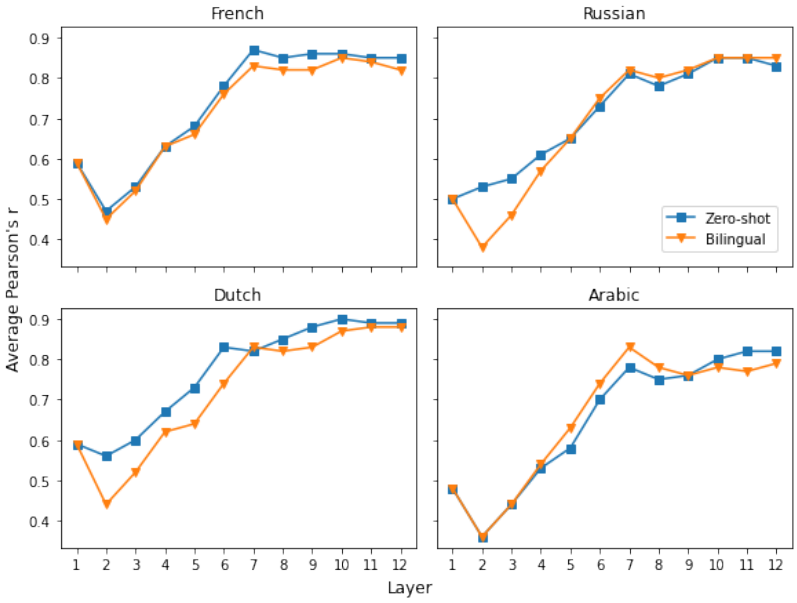


FIGURE 7.2: Comparing $\mathbf{O}_S$ (Zero-shot) with $\mathbf{O}_S$ + $\mathbf{Tr}_{S \to T}$ (Bilingual) using the average Pearson correlation coefficient between mBERT encodings of English utterances and their translations into a target language.

We observe that representations for source and translated utterances tend to be more correlated for French and Dutch compared with Russian and Arabic. This corresponds to our observation that cross-lingual transfer is most effective between related languages. Another aspect that supports this observation, though not very marked, is the higher correlation for $\mathbf{O}_S$ for French and Dutch compared with $\mathbf{O}_S$ + $\mathbf{Tr}_{S \to T}$, especially in the final layers. This probing analysis also reveals how multilingual representations are learned through the different layers. Early layers with lower correlations seem to be less concerned with specific languages and are perhaps dedicated to low-level properties of language in general. The final layers (from the seventh layer onward) are then capable of producing multilingual word representations.

Chapter 8

Conclusion

8.1 Discussion

In this work, we set out to explore generic approaches to ABSA. Thanks to transfer learning, a single model is able to leverage data from different domains and languages to learn generalizable representations. We have also shown how similar ABSA subtasks can be learned simultaneously and benefit each other using multitask learning.

Another thread was the efficient use of annotated resources. We have introduced a translation-based adaptation method to generate synthetic data and facilitate cross-lingual transfer for sequence tagging tasks. We extend word alignments for annotation projection by directly projecting spans to safely adapt resources to a target language. Our span-based projection method preserves token-level annotations and prevents misalignments thanks to a filtering heuristic. Starting from a reference annotated corpus, this approach can be used to tackle other languages for which no annotation is available.

We tested this method in two scenarios: cross-lingual adaptation, where annotated data is available for a source language only, and data augmentation, where annotated data is available in multiple languages. While zero-shot cross-lingual transfer can be applied somewhat successfully in the first scenario, our translation-based adaptation method greatly improves cross-lingual transfer, increasing the F_1 score by up to 20 points. These results suggest that multilingual PTMs are not fully language agnostic and that they benefit from being fine-tuned on the target language specifically, even if the unseen language is similar to the source language.

For data augmentation, the benefits of fine-tuning on additional translated data are less marked. This approach did not improve performance over the multilingual baseline for the in-house datasets. We argued that this was due to noisy text which is difficult to translate, and proposed several methods to alleviate this issue with modest success. But translation-based adaptation is still potentially helpful for data augmentation: the results improved for the cleaner SemEval corpora. In a final experiment, we successfully combined all the transfer learning methods that were introduced into one generic model. This suggests that this model is able to extract information learned from various domains, tasks, and languages and abstract it into generalizable high-level properties.

8.2 Future Work

We settled on one approach to translation-based adaptation, i.e. span-level projection using word alignments obtained from SimAlign which performed well. Other word alignment methods could be investigated, e.g. using attention weights from an

NMT model. Another possibility is to use an end-to-end approach to jointly project annotations and predict labels, as done by Xu, Haider, and Mansour, 2020. Similarly, it would be interesting to study how the use of different MT toolkits impacts performances on the downstream task.

The overall success of translation-based adaptation shows that it could be fruitful to apply it to other sequence tagging tasks. Projection trials showed that annotation projection is most effective for tasks with relatively short spans (e.g. noun phrases) such as named entity recognition and question answering. The projection of longer spans tends to be more error-prone. Some experiments on FrameNet semantic parsing showed that this method can also bring performance gains to other tasks.

Our attempt to improve adaptation by correcting noisy textual data before translating it was not successful. Processing noisy user-generated text is often much more difficult than dealing with clean text and remains an open problem. One possible avenue for improvement is to address this problem. Throughout the experiments, we used mBERT as our PTM for simplicity. However, mBERT is pre-trained on a large multilingual corpus of clean text from Wikipedia. In contrast, XLM-R was pre-trained on a larger corpus from clean CommonCrawl data. This corpus contains text scraped from the entire web. As such, it is more likely to contain noisy text. Using XLM-R as the multilingual PTM could therefore be useful to deal with noisy user-generated text.

Another promising approach is domain adaptation (Ramponi and Plank, 2020). The copious amounts of unlabeled user-generated text could be leveraged to prime a PTM to this type of data using self-supervision. This second phase of pre-training in-domain has shown a lot of potential to tailor a PTM to the domain of a target task (Gururangan et al., 2020). This approach has been succesfully applied to single ABSA tasks (Xu et al., 2019; Xu et al., 2020a; Rietzler et al., 2020). Going even further, some have devised dedicated self-supervised objectives to introduce linguistic knowledge related to sentiment analysis into PTMs (Ke et al., 2020). Domain adaptation is another transfer learning approach that could be integrated into our final model.

8.3 Closing Remarks

Deep neural networks require large amounts of annotated data to be efficient. Nowadays, NLP approaches mainly rely on deep learning, and many of the recent improvements can be attributed to sequential transfer learning and the efficient use of unlabeled text. While several new model architectures have brought leaps in performance, the quality of annotated data for a downstream task is crucial and remains a bottleneck. At the same time, the annotation process is costly and is a limiting factor. As a result, most works focus on high-resource languages.

This work showed how to adapt available annotated resources in a reference language to process unseen languages. With the increasing language coverage and quality of MT, this approach can help address the issue of scarce annotated data and thereby promote work on low-resource languages. Nonetheless, we showed how small amounts of original annotated data in the target languages could improve cross-lingual adaptation. These results highlight the importance of the annotation process to obtain quality annotated corpora.

Bibliography

Akbik, Alan, Duncan Blythe, and Roland Vollgraf (2018). "Contextual String Embeddings for Sequence Labeling". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 1638–1649.

Banea, Carmen et al. (2008). "Multilingual Subjectivity Analysis Using Machine Translation". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 127–135.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent (2013). "Representation Learning: A Review and New Perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8, pp. 1798–1828.

Bojanowski, Piotr et al. (2017). "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146.

Boyd, Adriane (2018). "Using Wikipedia Edits in Low Resource Grammatical Error Correction". In: *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*. Association for Computational Linguistics, pp. 79–84.

Brown, Tom et al. (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 1877–1901.

Bryant, Christopher, Mariano Felice, and Ted Briscoe (2017). "Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 793–805.

Chen, Yun et al. (2020). "Accurate Word Alignment Induction from Neural Machine Translation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 566–576.

Choe, Yo Joong et al. (2019). "A Neural Grammatical Error Correction System Built On Better Pre-training and Sequential Transfer Learning". In: *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, pp. 213–227.

Choenni, Rochelle and Ekaterina Shutova (2020). "What does it mean to be language-agnostic? Probing multilingual sentence encoders for typological properties". In: *ArXiv*.

Chollampatt, Shamil and Hwee Tou Ng (2018). "A Multilayer Convolutional Encoder-Decoder Neural Network for Grammatical Error Correction". In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

Conneau, Alexis et al. (2018). "XNLI: Evaluating Cross-lingual Sentence Representations". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Conneau, Alexis et al. (2020). "Unsupervised Cross-lingual Representation Learning at Scale". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 8440–8451.

Dai, Hongliang and Yangqiu Song (2019). "Neural Aspect and Opinion Term Extraction with Mined Rules as Weak Supervision". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Desai, Neelmay and Meera Narvekar (2015). "Normalization of Noisy Text Data". In: *Procedia Computer Science* 45. International Conference on Advanced Computing Technologies and Applications (ICACTA), pp. 127–132.

Devlin, Jacob et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, pp. 4171–4186.

Do, Hai Ha et al. (2019). "Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review". In: *Expert Systems with Applications* 118, pp. 272–299.

Duh, Kevin, 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*, pp. 429–433.

Dyer, Chris, 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*. Association for Computational Linguistics, pp. 644–648.

Eddine, Moussa Kamal, Antoine J-P Tixier, and Michalis Vazirgiannis (2020). "BARThez: a Skilled Pretrained French Sequence-to-Sequence Model". In: *ArXiv*.

Eger, Steffen et al. (2018). "Cross-lingual Argumentation Mining: Machine Translation (and a bit of Projection) is All You Need!" In: *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, pp. 831–844.

Fan, Zhifang et al. (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. URL: https://aclanthology.org/N19-1259.

Firth, J. R. (1957). "A synopsis of linguistic theory 1930-55." In: 1952-59, pp. 1–32.

Grundkiewicz, Roman, Marcin Junczys-Dowmunt, and Kenneth Heafield (2019). "Neural Grammatical Error Correction Systems with Unsupervised Pre-training on Synthetic Data". In: *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, pp. 252–263.

Gururangan, Suchin et al. (2020). "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 8342–8360.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780.

Honnibal, Matthew et al. (2020). *spaCy: Industrial-strength Natural Language Processing in Python*.

Hu, Minqing and Bing Liu (2004). "Mining and Summarizing Customer Reviews". In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge*

Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery, 168–177.

Jain, Alankar, 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*. Association for Computational Linguistics, pp. 1083–1092.

Jalili Sabet, Masoud et al. (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*. Association for Computational Linguistics, pp. 1627–1643.

Jebbara, Soufian 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)*. Association for Computational Linguistics, pp. 2486–2495.

Joty, Shafiq et al. (2017). "Cross-language Learning with Adversarial Neural Networks". In: *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Association for Computational Linguistics, pp. 226–237.

Junczys-Dowmunt, Marcin et al. (2018). "Marian: Fast Neural Machine Translation in C++". In: *Proceedings of ACL 2018, System Demonstrations*. Association for Computational Linguistics, pp. 116–121.

K, Karthikeyan et al. (2020). "Cross-Lingual Ability of Multilingual BERT: An Empirical Study". In: *International Conference on Learning Representations*.

Katsumata, Satoru and Mamoru Komachi (2020). "Stronger Baselines for Grammatical Error Correction Using a Pretrained Encoder-Decoder Model". In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, pp. 827–832.

Ke, Pei et al. (2020). "SentiLARE: Sentiment-Aware Language Representation Learning with Linguistic Knowledge". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 6975–6988.

Keung, Phillip, Yichao Lu, and Vikas Bhardwaj (2019). "Adversarial Learning with Contextual Embeddings for Zero-resource Cross-lingual Classification and 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)*. Association for Computational Linguistics, pp. 1355–1360.

Kiyono, Shun et al. (2019). "An Empirical Study of Incorporating Pseudo Data into Grammatical Error Correction". 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)*. Association for Computational Linguistics, pp. 1236–1242.

Kumar, Ankit, Piyush Makhija, and Anuj Gupta (2020). "Noisy Text Data: Achilles' Heel of BERT". In: *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*. Association for Computational Linguistics, pp. 16–21.

Lewis, Mike et al. (2020). "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Li, X. et al. (2020). "Unsupervised Cross-lingual Adaptation for Sequence Tagging and Beyond". In: *ArXiv*.

Li, Xin 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*. Association for Computational Linguistics, pp. 2886–2892.

Li, Xin et al. (2019a). "A unified model for opinion target extraction and target sentiment prediction". In: *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6714–6721.

Li, Xin et al. (2019b). "Exploiting BERT for End-to-End Aspect-based Sentiment Analysis". In: *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*. Association for Computational Linguistics, pp. 34–41.

Liu, Bing and Lei Zhang (2012). "A Survey of Opinion Mining and Sentiment Analysis". In: *Mining Text Data*. Ed. by Charu C. Aggarwal and ChengXiang Zhai. Springer US, pp. 415–463.

Loshchilov, Ilya and Frank Hutter (2019). "Decoupled Weight Decay Regularization". In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Luo, Huaishao et al. (2019). "DOER: Dual Cross-Shared RNN for Aspect Term-Polarity Co-Extraction". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 591–601.

Malykh, Valentin, Varvara Logacheva, and Taras Khakhulin (2018). "Robust Word Vectors: Context-Informed Embeddings for Noisy Texts". In: *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*. Association for Computational Linguistics, pp. 54–63.

Max, Aurélien and Guillaume Wisniewski (2010). "Mining Naturally-occurring Corrections and Paraphrases from Wikipedia's Revision History". In: *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Ed. by Nicoletta Calzolari et al. European Language Resources Association (ELRA).

Mikolov, Tomas et al. (2013). "Efficient Estimation of Word Representations in Vector Space". In: *CoRR* abs/1301.3781.

Muller, Benjamin, Benoit Sagot, and Djamé Seddah (2019). "Enhancing BERT for Lexical Normalization". In: *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*. Association for Computational Linguistics, pp. 297–306.

Omelianchuk, Kostiantyn et al. (2020). "GECToR – Grammatical Error Correction: Tag, Not Rewrite". In: *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Seattle, WA, USA â†' Online: Association for Computational Linguistics, pp. 163–170.

Östling, Robert and Jörg Tiedemann (2016). "Efficient word alignment with Markov Chain Monte Carlo". In: *Prague Bulletin of Mathematical Linguistics* 106, pp. 125–146.

Pang, Bo and Lillian Lee (2008). "Opinion Mining and Sentiment Analysis". In: *Found. Trends Inf. Retr.* 2.1–2, 1–135.

Peng, Haiyun et al. (2020). "Knowing What, How and Why: A Near Complete Solution for Aspect-Based Sentiment Analysis". In: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, pp. 8600–8607.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 1532–1543.

Peters, Matthew E. et al. (2018). "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237.

Pires, Telmo, Eva Schlinger, and Dan Garrette (2019). "How Multilingual is Multilingual BERT?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 4996–5001.

Pontiki, Maria et al. (2014). "SemEval-2014 Task 4: Aspect Based Sentiment Analysis". In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Association for Computational Linguistics, pp. 27–35.

Pontiki, Maria et al. (2015). "SemEval-2015 Task 12: Aspect Based Sentiment Analysis". In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics, pp. 486–495.

Pontiki, Maria et al. (2016). "SemEval-2016 Task 5: Aspect Based Sentiment Analysis". In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, pp. 19–30.

Popescu, Ana-Maria 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*. Association for Computational Linguistics, 339–346.

Pouran Ben Veyseh, Amir et al. (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)*. Online: Association for Computational Linguistics, pp. 8947–8956.

Radford, Alec and Karthik Narasimhan (2018). "Improving Language Understanding by Generative Pre-Training". In:

Radford, Alec et al. (2019). "Language Models are Unsupervised Multitask Learners". In:

Raffel, Colin et al. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* 21.140, pp. 1–67.

Ramponi, Alan and Barbara Plank (2020). "Neural Unsupervised Domain Adaptation in NLP—A Survey". In: *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, pp. 6838–6855.

Rietzler, Alexander et al. (2020). "Adapt or Get Left Behind: Domain Adaptation through BERT Language Model Finetuning for Aspect-Target Sentiment Classification". In: *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, pp. 4933–4941.

Ruder, Sebastian (2019). "Neural Transfer Learning for Natural Language Processing". PhD thesis. National University of Ireland, Galway.

Schwenk, Holger 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).

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016). "Neural Machine Translation of Rare Words with Subword Units". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 1715–1725.

Sumbler, P et al. (2018). "Handling Noise in Distributional Semantic Models for Large Scale Text Analytics and Media Monitoring". In: *Proceedings of the Abstract in the Fourth Workshop on Noisy User—Generated Text (W-NUT 2018), Brussels, Belgium*. Vol. 1.

Sun, Yifu and Haoming Jiang (2019). "Contextual Text Denoising with Masked Language Model". In: *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*. Association for Computational Linguistics, pp. 286–290.

Tian, Leimin, Catherine Lai, and Johanna Moore (2018). "Polarity and Intensity: the Two Aspects of Sentiment Analysis". In: *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*. Association for Computational Linguistics, pp. 40–47.

Tiedemann, Jörg 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)*.

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc.

Wang, Wenya et al. (2017). "Coupled Multi-Layer Attentions for Co-Extraction of Aspect and Opinion Terms". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI Press, 3316–3322.

Wisniewski, Guillaume, Aurélien Max, and François Yvon (2010). "Recueil et Analyse d'un corpus écologique de corrections orthographiques extrait des révisions de Wikipédia". In: *Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2010)*.

Wolf, Thomas et al. (2020). "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, pp. 38–45.

Wu, Shijie 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)*. Association for Computational Linguistics, pp. 833–844.

Wu, Zhen et al. (2020a). "Grid Tagging Scheme for Aspect-oriented Fine-grained Opinion Extraction". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 2576–2585.

Wu, Zhen et al. (2020b). "Latent Opinions Transfer Network for Target-Oriented Opinion Words Extraction". In: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, pp. 9298–9305.

Xu, Hu et al. (2018). "Double Embeddings and CNN-based Sequence Labeling for Aspect Extraction". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Xu, Hu et al. (2019). "BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis". 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, pp. 2324–2335.

Xu, Hu et al. (2020a). "DomBERT: Domain-oriented Language Model for Aspect-based Sentiment Analysis". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, pp. 1725–1731.

Xu, Lu et al. (Nov. 2020b). "Position-Aware Tagging for Aspect Sentiment Triplet Extraction". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 2339–2349.

Xu, Weijia, 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)*. Association for Computational Linguistics, pp. 5052–5063.

Yang, Yinfei et al. (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)*. Association for Computational Linguistics, pp. 3687–3692.

Zhou, Jie et al. (2019). "Deep learning for aspect-level sentiment classification: Survey, vision, and challenges". In: *IEEE Access* 7, pp. 78454–78483.

Zhou, Xinjie, Xiaojun Wan, and Jianguo Xiao (2016). "Cross-Lingual Sentiment Classification with Bilingual Document Representation Learning". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 1403–1412.

Zouhar, Vilém and Daria Pylypenko (2021). "Leveraging Neural Machine Translation for Word Alignment". In: *ArXiv*.