LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG
STUDIENGANG COMPUTERLINGUISTIK

Bachelorarbeit

im Studiengang Computerlinguistik

an der Ludwig-Maximilians-Universität München

Fakultät für Sprach- und Literaturwissenschaften

Understanding indirect answers: a cross-lingual transfer learning approach

vorgelegt von
Christin Müller

Betreuer:	Prof. Dr. Barbara Plank
Prüfer:	Prof. Dr. Barbara Plank
Bearbeitungszeitraum:	27. März - 05. Juni 2023

Selbstständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig angefertigt, alle Zitate als solche kenntlich gemacht sowie alle benutzten Quellen und Hilfsmittel angegeben habe.

München, den 05. Juni 2023

. . . .
Christin Müller

Abstract

Yes/no-questions are in fact designed for exactly two different answers – even though it is not uncommon for the actual answer to be hidden in an indirect one that only implies the response. To the question: *"An island? By the sea?"* the answer: *"Islands are generally by the sea"* seems natural in a spoken dialogue, although it might not be the answer the person asking intended to receive. However, conversational systems still struggle to interpret such indirect responses. First successes were made in the last couple of years in question-answering work though, but they all focused on datasets in English and no approach has adopted cross-lingual transfer learning methods so far. Thus, in this thesis, a new corpus is presented for which yes/no-questions and their respective indirect answers are extracted from subtitles. We opted for subtitles since they are a readily available source of conversation and even more so in various languages. The dataset consists of 615 manually annotated pairs for English, and 438 pairs each for French and Spanish. For evaluation, cross-lingual transfer learning approaches based on multilingual BERT will be used by comparing intermediate task training models to zero-shot baseline models. Although the performance results for individual annotation classes are noteworthy, the overall performance leaves room for further improvements, both for the baseline, as well as for the intermediate task training models.

Zusammenfassung

Ja/Nein-Fragen sind im Wesentlichen auf genau zwei verschiedene Antworten ausgelegt – auch wenn die eigentliche Antwort nicht selten in einer indirekten versteckt ist, die die Antwort nur andeutet. Auf die Frage: *"An island? By the sea?"* scheint die Antwort: *"Islands are generally by the sea"* in einem gesprochenen Dialog komplett natürlich, auch wenn es vielleicht nicht der Antwort entspricht, die die Person mit der Frage im Sinn hatte. Konversationssysteme haben allerdings nach wie vor Schwierigkeiten, solche indirekten Aussagen zu interpretieren. In den letzten Jahren wurden erste Erfolge im Bereich des Question-Answering erzielt, jedoch haben sich alle Arbeiten auf Englische Korpora fokussiert und es gibt keine Ansätze, die cross-linguale Transfer-Learning-Methoden nutzen. Aus diesem Grund wird in dieser Arbeit ein neues Korpus vorgestellt, für das Ja/Nein-Fragen und die jeweiligen indirekten Antworten aus Untertiteln extrahiert werden. Wir haben uns für Untertitel entschieden, da sie eine leicht zugängliche Quelle für Konversationen sind, noch dazu in verschiedenen Sprachen verfügbar. Der Datensatz besteht aus 615 manuell annotierten Paaren für Englisch und jeweils 438 Paaren für Französisch und Spanisch. Zur Evaluierung werden auf dem multilingualen BERT-Modell basierende, cross-linguale Transfer-Learning-Ansätze verwendet, indem Intermediate Task Training-Modelle mit Zero-Shot Baseline-Modellen verglichen werden. Obwohl die Performanceergebnisse für einzelne Annotationsklassen erwähnenswert sind, lässt die Gesamtleistung Raum für weitere Verbesserungen, sowohl für die Baseline- als auch für die Intermediate Task Training-Modelle.

Contents

1 Introduction

1.1 The Problem

Even when asked a non-wh question, that is, a yes/no-question, humans do not always state their answer clearly as *yes* or *no*. Because sometimes the answer is more complicated than a simple *yes* or *no*. Or because a *yes* alone doesn't seem affirmative enough. Or simply because human communication is not only binary and consists of so many more levels and options to choose from.

Stenström (1984) defines that in a more scientific way, less based on spontaneous observations. She analyzes various scenarios and reasons why an indirect answer is uttered even when a yes/no-question is asked (Stenström, 1984). One of her observations is that "there is a strong tendency to provide additional information [. . .], which of course does not mean that the speakers are not able and willing to furnish the decision required" (Stenström, 1984, p. 201). Often an indirect answer already implies a *yes*, or a *no*, or even a *maybe*, argues Stenström (1984), "[. . .] or it adds what may seem to be superfluous information, but which is important for social reasons" (Stenström, 1984, p. 202). They are part of a more natural conversation in general (Stenström, 1984), indirect answers can be more convenient than just saying *no* or even act as "a time-saving device, a kind of filler“ (Stenström, 1984, p. 206).

For Raymond (2003), yes/no-questions and their respective answers are "[. . .] one of the most pervasive practices in interaction" (Raymond, 2003, p. 939) and he states that "evidently, speakers have alternative resources for delivering preferred responses and alternative resources for delivering dispreferred responses." (Raymond, 2003, p. 946).

An example for an indirect answer that does imply a *yes* answer, but does not literally say *yes*, is provided next (taken from the *XOpus-QIA* dataset, the corpus that will be introduced in this thesis):

Question: "An island? By the sea?"
Answer: "Islands are generally by the sea."
Question: "Une île ? Près de la mer ?"
Answer: "Les îles sont généralement près de la mer."
Question: "¿Una isla? ¿En el mar?"
Answer: "Las islas generalmente están en el mar."

Even though there is neither a *yes* or a *no* given in the answer, it seems obvious that a *yes* is implied, since one might refer to it as common knowledge. However, no matter how frequent those indirect answers are in everyday communication – or more precisely in "natural language understanding" (Clark et al., 2019, p. 2924) – there is one challenge: "How should a dialog system interpret these indirect answers?" (Louis, Roth, and Radlinski, 2020, p. 7411).

1.2 The Research Goal

In the last couple of years first successes have been achieved in question-indirect answering work in natural language processing. Louis et al. (2020) made a start and introduced – in their own words – "[. . .] the first large scale corpus and models for interpreting such indirect answers" (Louis et al., 2020, p. 7411). This is an important clarification by Louis et al. (2020), since there have been approaches prior to theirs. One of them is Green and Carberry's work in 1999 regarding not only interpreting, but also generating indirect answers to yes/no-questions (Green and Carberry, 1999). Green and Carberry (1999) state that their model's "[...]purpose is to compute the end products of comprehension and generation, and to contribute to a computational theory of conversational implicature" (Green and Carberry, 1999, p. 391). A decade later, de Marneffe, Grimm, and Potts (2009) tried a logic-based approach. For de Marneffe et al. (2009), when an answer does not mean *yes* or *no*, it is uncertain. Which leads them to their "Marcov logic network" (de Marneffe et al., 2009, p. 140): "[...] they allow rich inferential reasoning on relations by combining the power of firstorder logic and probabilities to cope with uncertainty (de Marneffe et al., 2009, p. 140).

Back to more recent approaches, a different corpus than the one by Louis et al. (2020) was collected by Damgaard, Toborek, Eriksen, and Plank (2021). To proceed in this understanding indirect answer task, they opted to choose more natural data extracted from TV scripts (Damgaard et al., 2021), in comparison to prompts and scenarios generated in Louis et al. (2020)'s work and tailored to their exact task. And the most recent work focuses on the importance of context: "[...] as the ground truth changes depending on whether we show annotators context around the question" (Sanagavarapu et al., 2022, p. 4684).

However, all three approaches mentioned above focused on datasets in English (Louis et al., 2020; Damgaard et al., 2021; Sanagavarapu et al., 2022). This thesis tries to overcome the gap Louis et al. (2020) already mentioned at the end of their paper, by proposing a new dataset *XOpus-QIA* consisting of English, French and Spanish subtitle data: "There are exciting avenues for multilingual work to account for language and cultural differences" (Louis et al., 2020, p. 7419).

1.3 The Key Research Questions

One of the key research questions therefore is how and to what extent the model can be adapted to languages other than English – or more specifically – to French and Spanish. Even though French and Spanish are typically not categorized as low resource languages, yet still no datasets are available for this specific task.

To be able to solve this indirect answer understanding problem on a cross-lingual level and to answer the question about what kind of data can be used, a new dataset is needed. The third and also crucial research question of this thesis will be what the best method is for transfer.

1.4 Contributions

To be able to fulfill all three research questions, first we review related and recent work both in the area of understanding indirect answers, but also in the field of approaches suitable for cross-lingual solutions. In this case, this thesis examines transfer learning approaches in more detail in section 2.3.

The second research question leads to a new, cross-lingual dataset – parallel data in English, French and Spanish from the OpenSubtitles Corpus (OPUS) (Lison and Tiedemann, 2016) is collected, with a total of 615 question-indirect answer pairs (QIA pairs) for English and 438 parallel QIA pairs each for evaluation in French and Spanish. This dataset, the XOpus-QIA, is explained in more details in chapter 3 – from preliminary considerations to data collection to annotator agreement.

Chapter 4 explores the methods and explains all conducted experiments. This includes the general setup using the MaChAmp toolkit (van der Goot, Üstün, Ramponi, Sharaf, and Plank, 2021), the baseline models and the models for comparison and evaluation, in this case, intermediate task learning models that use not only the cross-lingual dataset extracted for this thesis for training, but also task-related datasets from previous papers.

To answer the question on the best method for transfer, chapter 5 gives an overview of the respective model performances (accuracy and F1-measures). To conclude this thesis, the results will be summarized, and future work will be discussed in accordance with the take-aways identified.

2 Previous Work

2.1 Indirect Answers

2.1.1 Circa

Louis et al. (2020) introduced *Circa*, a corpus consisting of 32,268 pairs of yes/no-questions, or "polar questions" (Louis et al., 2020, p. 7411) and indirect answers to these questions (Louis et al., 2020). With Circa, they aim to overcome the gap that "previous attempts to interpret indirect yes/no answers have been small scale and without data-driven techniques" (Louis et al., 2020, p. 7411). Their dataset was crowd-sourced and to obtain indirect answers that seem as natural as possible, crowd workers were instructed to stick to realistic conversation scenarios (Louis et al., 2020).

What is special about their annotations and will be relevant for this thesis, is that they not only include *yes* and *no*, but also non-binary answers that, for example, neither mean *yes* nor *no* (Louis et al., 2020). "[...] the variety of examples in our corpus made it certain that just 'yes' and 'no' will not suffice" (Louis et al., 2020, p. 7414). This resulted in their "strict" and "relaxed" label set (Louis et al., 2020, p. 7415). The *strict* label set includes nine labels, next to "yes" and "no" it also contains labels with a certain non-binary uncertainty, such as "probably yes" (Louis et al., 2020, p. 7415) or "probably no" (Louis et al., 2020, p. 7415). In the *relaxed* label set, those uncertain labels are merged with the certain ones, resulting in a reduced label set of six labels, consisting of five class distinctions and one label of annotator disagreement (see table 5.6 for an overview of the label sets).

Strict Label Set	**Relaxed Label Set**
Yes	Yes
No	No
Probably yes / sometimes yes	-
Yes, subject to some conditions	Yes, subject to some conditions
Probably no	-
In the middle, neither yes nor no	In the middle, neither yes nor no
Other	Other
N/A	N/A

Table 2.1: The strict and relaxed label set, as introduced and used by Louis et al. (2020, p. 7415).

As for their models, they experimented on various pre-trained transformer BERT models (Devlin, Chang, Lee, and Toutanova, 2019), but also considered baseline models and models that are trained on either question or answer only (Louis et al., 2020). For both baseline and BERT models, they not only use their own created Circa corpus, but also draw on related corpora, albeit designed for different tasks – such as BoolQ for question-answering (Clark et al., 2019) or MNLI (Williams, Nangia, and Bowman, 2018) for textual entailment (Louis et al., 2020). For the baseline models, they adapt their own test set to the labels given in the BoolQ or MNLI datasets. However, for the BERT models, such adaptations did not take place and as for their BERT-BOOLQ-YN model, they "[...] expect to learn many semantics of yes/no answers from this data" (Louis et al., 2020, p. 7417).

Their results vary, with the adapted BoolQ baseline model being the best one with an accuracy of 63 percent for the relaxed label set (Louis et al., 2020, p. 7418). The question- and answer-only models both outperform the baseline models (Louis et al., 2020, p. 7417) and the overall best scores (88.2 percent accuracy) are reached for the BERT model that uses next to the Circa dataset the textual entailment MNLI dataset (Louis et al., 2020, p. 7417). What seems noteworthy is that their BERT based model without transfer task is almost as good, having an accuracy of 87.8 percent (Louis et al., 2020).

2.1.2 FRIENDS-QIA

Damgaard et al. (2021) proposed a different dataset to solve this problem: their corpus *FRIENDS-QIA* is based upon the famous Friends TV series. Damgaard et al. (2021) collected 5,390 pairs of yes/no-questions and their respective indirect answer. "[...] we attempt to exploit already existing data by simply looking for useful question-answer (QA) pairs in dialogues" (Damgaard et al., 2021, p. 2). They collected, preprocessed and annotated their data manually and in house (Damgaard et al., 2021). Their data annotation scheme is quite similar to the one by Louis et al. (2020), as Damgaard et al. (2021) used the relaxed label set (Louis et al., 2020). Which, as a reminder, consists of six labels in total (Louis et al., 2020; Damgaard et al., 2021).

Their models use BERT or GloVe representations (Pennington, Socher, and Manning, 2014) and the overall structure is based upon 1-dimensional convolutions, "the base CNN" (Damgaard et al., 2021, p. 6). The best result is reached with BERT embeddings (accuracy of 64.08 percent), whereas their baseline models do not perform well; they reach an accuracy of 49.07 percent for the majority baseline, and 52.45 percent for the Naïve Bayes baseline (Damgaard et al., 2021). For Damgaard et al. (2021) those baseline results show that solving this question-indirect answer task is quite a complex one (Damgaard et al., 2021).

Damgaard et al. (2021) noted that models trained on the FRIENDS-QIA corpus differ from when they train models on the Circa corpus by Louis et al. (2020), with higher performance on Circa. They analyze three distinctions that can explain the results: first, they mention the difference in data collection, second the difference in allowing answers that are made up of more than one sentence, "[...] resulting in the CIRCA data being much more concise in meaning and structure than FRIENDS-QIA" (Damgaard et al., 2021, p. 9). And since the Circa corpus is much larger than the Friends-QIA, Damgaard et al. (2021) argue that the CNNs have more data "[...] to learn well from" (Damgaard et al., 2021, p. 9).

2.1.3 SwDA-IA

A third and recent work on yes/no-questions and indirect answers not only considers the isolated answers to a question but the context as well (Sanagavarapu et al., 2022). Sanagavarapu et al. (2022) introduced the *SwDA-IA* dataset with 2,544 yes/no-questions and indirect answers. It contains authentic question-answer pairs from the Switchboard corpus (Jurafsky, Shriberg, and Biasca, 1997). Their dataset, however, is not yet accesible to the public[1]. Sanagavarapu et al. (2022) explain their choice as "[...] we work with real conversations as opposed to artificial, synthetic ones" (Sanagavarapu et al., 2022, p. 4678). Their label set differs from the relaxed label set by Louis et al. (2020), since they only use five labels, but include the uncertainty classes like "probably yes" (Sanagavarapu et al., 2022, p. 4679) and "probably no" (Sanagavarapu et al., 2022, p. 4679).

Their approach focuses mainly on the importance of context, hence they annotate their data in two steps: first, annotators only annotated the question and the immediate answer that follows (Sanagavarapu et al., 2022). Second, annotators were shown more context around the question-answer-pair (three speaker-answerer-turns) (Sanagavarapu et al., 2022). They gained different annotations for both steps which leads to the explanation

[1] Last checked for availability on 23-06-04.

that "[...] context beyond the yes-no question and indirect answer (i.e., the following turn) is needed to determine the ground truth in real conversations" (Sanagavarapu et al., 2022, p. 4680).

Even though they applied three transformer models – namely BERT, the modified transformer model RoBERTa (Liu et al., 2019b), and TOD-BERT, a "task-oriented dialogue BERT" (Wu, Hoi, Socher, and Xiong, 2020, p. 917) – regarding the evaluation they focus on RoBERTa only since it performed best in their setting (Sanagavarapu et al., 2022, p. 4682).

Sanagavarapu et al. (2022) not only used their own corpus, like Louis et al. (2020) they experimented with BoolQ and MNLI, as well as Circa. One interesting result of their work is that when more than the plain question-indirect-answer pairs, that is, the entire context, is used, the results do not improve at all (Sanagavarapu et al., 2022). "This may seem surprising, however, it is known that keeping track of a conversation across several turns is challenging" (Sanagavarapu et al., 2022, p. 4682).

One of their key results is that a model fine-tuned with their Switchboard question-indirect answer pairs from real conversations and combined with different datasets leads to the best performance: "[...] determining indirect answers to yes-no questions in real conversation requires fine-tuning with real conversations [...]" (Sanagavarapu et al., 2022, pp. 4682-4683). Since the MNLI, BoolQ, or Circa dataset do not consist of real conversations, models relying on them do not perform better than models that rely on real conversations (Sanagavarapu et al., 2022, p. 4684).

2.2 Related Tasks

Since yes/no-questions have also been subject of studies in different contexts, this section introduces related tasks that also seem noteworthy and some of them will be used later in this thesis for model setup.

2.2.1 MNLI

MNLI, MultiNLI, or "Multi-Genre Natural Language Inference" (Williams et al., 2018, p. 1112) might not be exactly a dataset related to indirect answers to yes/no-questions. It was introduced – first as a draft – in 2017 and its aim was to make a progress regarding sentence understanding (Williams et al., 2018): "Our chief motivation in creating this corpus is to provide a benchmark for ambitious machine learning research on the core problems of NLU [...]" (Williams et al., 2018, p. 1113). Despite this, this shall not be the only use of the newly created dataset if it is up to the authors, one of their interests is cross-domain transfer learning (Williams et al., 2018, p. 1113).

The MNLI dataset is made up of sentence pairs; one sentence is a premise, extracted from text, and the second sentence of the pair is a hypothesis, formulated by an annotator. Every pair is labeled either as *entailment*, *neutral*, or *contradiction* (Williams et al., 2018, p. 1112) (see example in the appendices section). They trained three different neural network models on MNLI and because it "[...] offers dramatically greater linguistic difficulty and diversity" (Williams et al., 2018, p. 1120) compared to similar, previous datasets, the authors see MNLI as "an effective source task for pre-training and transfer learning [...]" (Williams et al., 2018, p. 1120). This last fact is what makes MNLI not only important for this thesis, but also for related tasks. Even though MNLI does not focus on yes/no-questions, it is used for transfer learning with remarkable results (Clark et al., 2019).

2.2.2 QuAC

QuAC, "Question Answering in Context" (Choi et al., 2018, p. 2174) is a dataset that contains 14,000 naturally formulated questions and answers based on pre-existing texts

(Choi et al., 2018). With this dataset, the authors try to reconstruct natural dialogues that also contain yes/no-questions. In this regard, they make the following observation: "The frequency of yes/no questions increases significantly as the dialogs progress" (Choi et al., 2018, p. 2178).

Even though their specific dialog setup produces more wh-questions, since yes/no-questions are inevitable in natural conversations, their dataset is still made up of 25.8 percent of yes/no-questions (Choi et al., 2018, p. 2175).

2.2.3 BoolQ

One paper that makes use of MNLI and is of crucial importance to this thesis is *BoolQ*, a dataset with yes/no-questions that occur naturally (Clark et al., 2019). The authors state that those type of questions "are unexpectedly challenging" (Clark et al., 2019, p. 2924) and that they "require difficult entailment-like inference to solve" (Clark et al., 2019, p. 2924). BoolQ contains 16,000 yes/no-questions, whereas every question matches an answer, that is, an extracted paragraph from Wikipedia. Every question-answer pair is labeled by an annotator with a Boolean answer (Clark et al., 2019) (see an example in the appendices section).

Clark et al. (2019) receive quite satisfying results using transfer learning methods. They tried various datasets for this task, since they claim that "Yes/No QA is closely related to many other NLP tasks [...]" (Clark et al., 2019, p. 2925). One of those datasets that led to "the best results" for supervised models (Clark et al., 2019, p. 2925) is the dataset for entailment, MNLI – which is in contrast to the results they receive when training on BoolQ only (Clark et al., 2019). They also try paraphrasing, multiple-choice question answer, and extractive question answer datasets for transfer learning.

However, MNLI performed best as a supervised model (Clark et al., 2019). And what the authors call "a surprising result" (Clark et al., 2019, p. 2932) is that datasets similar to BoolQ (that means, containing question and answer) did not perform as expected regarding the transfer learning task (Clark et al., 2019).

As unsupervised model they chose a BERT model (Devlin et al., 2019), which not only leads to the best accuracy scores for unsupervised models, but outperforms all models, even the one that trains on MNLI as transfer task (Clark et al., 2019).

When trained on BERT and fine-tuned first on MNLI, then on BoolQ, the models even reach a test accuracy of 80.43 percent (compared to 76.90 percent for the unsupervised, no-MNLI BERT model) (Clark et al., 2019, pp. 2931-2932). Clark et al. (2019) discuss that yes/no-questions are more often used when the information need is more complex and thus requires more inference (Clark et al., 2019, p. 2928). As well as "[...] a key advantage of MultiNLI is that it contains examples of contradictions" (Clark et al., 2019, p. 2932).

2.3 (Cross-Lingual) Transfer Learning

2.3.1 Why Transfer Learning

There is a challenge when working with machine learning, according to Pan and Yang (2010): "[...] methods work well only under a common assumption: the training and test data are drawn from the same feature space and the same distribution" (Pan and Yang, 2010, p. 1345).

To avoid the problem of rebuilding and/or collecting new data, Pan and Yang (2010) recommend "knowledge transfer or transfer learning" (Pan and Yang, 2010, p. 1345), an approach, that according to them has many names: "learning to learn, life-long learning, knowledge transfer, inductive transfer, multitask learning, knowledge consolidation, context-sensitive learning, knowledge-based inductive bias, metalearning, and incremental/cumulative learning" (Pan and Yang, 2010, p. 1346).

What Pan and Yang (2010) describe as beneficial regarding transfer learning is the fact that it “[...] allows the domains, tasks, and distributions used in training and testing to be different” (Pan and Yang, 2010, p. 1346).

2.3.2 Transfer Learning and BERT

Devlin et al. (2019) argue that pre-training still takes an important part for language understanding models that apply transfer learning. One of those pre-trained models is BERT, as first mentioned in subsection 2.1.1, a pre-trained transformer model (Devlin et al., 2019). What was new about BERT when it was introduced by Devlin et al. (2019) is its bidirectionality. That is, BERT is looking at context not only on the right, but also on the left (Devlin et al., 2019, p. 4171). According to Devlin et al. (2019), this bidirectional approach (hence the name BERT – “Bidirectional Encoder Representations from Transformers”, Devlin et al., 2019, p. 4171) is what makes BERT superior on the sentence level for example in contrast to the back then state-of-the-art approaches with the pre-trained GPT model (Radford and Narasimhan, 2018) or ELMo, a “deep contextualized word representation” (Peters et al., 2018, p. 2227).

Devlin et al. (2019) describe BERT as “conceptually simple and empirically powerful” (Devlin et al., 2019, p. 4171). BERT first trains on data that is unlabeled, then sequentially fine-tunes on labeled data (Devlin et al., 2019).

Pires, Schlinger, and Garrette (2019) acknowledge with their contributions that BERT, or more specifically multilingual BERT (mBERT), “[...] is surprisingly good at zero-shot cross-lingual model transfer [...]” (Pires et al., 2019, p. 4996). Multilingual BERT uses articles from Wikipedia from 104 languages for pre-training, yet it is still a single language model (Devlin et al., 2019; Pires et al., 2019). Wu and Dredze (2019) confirm to BERT an “impressive performance for zero-shot cross-lingual transfer on a natural language inference task” (Wu and Dredze, 2019, p. 833). In their work, Wu and Dredze (2019) experiment on what they call the “potential of mBERT” (Wu and Dredze, 2019, p. 833), by testing BERT for a variety of NLP tasks (Wu and Dredze, 2019), that includes inference tasks and classification tasks. Due to the tokenization algorithm WordPiece (Devlin et al., 2019), there is a large, shared vocabulary (Devlin et al., 2019). As part of their conclusion, Wu and Dredze (2019) recommend mBERT for future work, since it “[...] effectively learns a good multilingual representation with strong cross-lingual zero-shot transfer performance in various tasks” (Wu and Dredze, 2019, p. 841).

2.3.3 Cross-Lingual Transfer Learning Settings

As stated in 2.1, so far there have not been any approaches in understanding indirect answers with non-English datasets. A problem that not only exists for this individual task. “Contemporary natural language processing systems typically rely on annotated data to learn how to perform a task [...]” (Conneau et al., 2018, p. 2475). Annotated data, that in our case, is not available yet for understanding indirect answers to yes/no-questions in Romance languages such as French or Spanish.

The following datasets and models all were created in and applied to the context of cross-lingual transfer learning. What they all have in common is that they are motivated by the need to find solutions to the fact that for many tasks there are often only English datasets available. Two of the datasets that follow in this subsection will also play important roles in this thesis in the corresponding intermediate task training models.

XNLI

Conneau et al. (2018) describe the problem of having labeled data for a variety of languages as an unrealistic scenario and the reason why “[...] there has been a growing interest in cross-lingual understanding and low-resource transfer in multilingual scenarios” (Conneau et al., 2018, p. 2483). They introduced *XNLI*, a dataset based on the

MNLI dataset (Williams et al., 2018). The training data remains in English, nevertheless, for development and test sets the data is translated (human translated) (Conneau et al., 2018). They chose to translate the development and test set instead of creating an entire new corpus for various reasons, one of them is to ensure parallel data (Conneau et al., 2018, p. 2477). Even though they still train on English data and only have development and test sets in different languages available, due to the data alignment, they still receive satisfying results (Conneau et al., 2018, p. 2476).

Considering their data, they encountered some irregularities. That is, once translated, they observed some cases where the gold label changed (Conneau et al., 2018). However, they claim that those cases were quite limited to Chinese and rather seldom (Conneau et al., 2018, p. 2477).

XQA

XQA is a cross-lingual dataset for open-domain question-answering (Liu, Lin, Liu, and Sun, 2019a, p. 2358). The authors rank the cross-lingual open-domain question-answering task as a part of cross-lingual language understanding (Liu et al., 2019a, p. 2358) and for them it is an alternative to generate large-scale datasets in various languages (Liu et al., 2019a). The training set of XQA is in English, while native speakers for each language generate questions for the development and test set, a way to "[...] reflect cultural differences in different languages" (Liu et al., 2019a, p. 2359).

Liu et al. (2019a) implement three different models; two models are based on machine-translation, one is a zero-shot setting with multilingual BERT (mBERT, Devlin et al., 2019) (Liu et al., 2019a, p. 2358). Their result is straightforward: the multilingual BERT model achieves the best performance in almost all target languages, while translation-based methods suffer from the problem of translating name entities (Liu et al., 2019a, p. 2359). Besides the only restriction that English-only datasets and models still perform even better than the mBERT cross-lingual zero-shot models, the authors argue that it still is more practical to train cross-lingual, in this case with XQA (Liu et al., 2019a).

PAWS-X

Instead of generating new data, *PAWS-X*, a sentence-paraphrase dataset for classification, focuses on translations only – machine translation for training data, human translation for development and test data (Yang, Zhang, Tar, and Baldridge, 2019) (see example in the appendices section). The original, monolingual English *PAWS* dataset was set up by Zhang, Baldridge, and He (2019) to "break NLP systems" (Zhang et al., 2019, p. 1299) by creating counterexamples. Zhang et al. (2019) explain in their work, that interpretation is dependent on word order and even small alterations can change meaning. PAWS therefore is according to Zhang et al. (2019) a dataset that consists of pairs that seem identical but differ regarding the word order – which also leads to sentences that seem identical, but do not mean the same thing (Zhang et al., 2019). Zhang et al. (2019) further conclude that "[...] PAWS training data for state-of-the-art models dramatically improves their performance on challenging examples and makes them more robust to real world examples" (Zhang et al., 2019, p. 1306).

Yang et al. (2019) translated the original mono-lingual PAWS (Zhang et al., 2019) dataset to a variety of languages. Yang et al. (2019) explain that it is beneficial to translate datasets, because once the translated sentences are based on the same examples (in this case, monolingual data in English), it favors cross-lingual learning and allows cross-lingual conclusions (Yang et al., 2019, p. 3688).

They evaluate three different models, among others a traditional bag-of-words model, as well as a BERT model (Yang et al., 2019). They also test different approaches on each model: they translated the English training data (machine translation), but they also back-translated the test data into English (also machine translation) (Yang et al., 2019, p.

3689). For further cross-lingual evaluation, they also implement a zero-shot model (Yang et al., 2019). The latter, however, performed worse than the models that are based on machine-translated data (Yang et al., 2019).

Since all BERT models outperform the other models, Yang et al. (2019) conclude: "PAWS-X shows the effectiveness of deep, multilingual pre-training [...]" (Yang et al., 2019, p. 3687).

2.3.4 Intermediate Task Training

Since pretraining and fine-tuning prior to Phang, Févry, and Bowman (2019)'s work led to better results than target task training only, Phang et al. (2019) experimented even further with additional pretraining, that is, intermediate task training (Phang et al., 2019, chap. 1). Their experiments contained three steps: they first set up pre-trained language models (training on unlabeled data), then on labeled data (the intermediate task), and eventually they fine-tuned on the target task and evaluated the model (Phang et al., 2019, chap. 1). For pretraining, they used the transformer models BERT (Devlin et al., 2019) and GPT (Radford and Narasimhan, 2018), as well as the non-transformer, deep contextualized model ELMo (Peters et al., 2018).

For the intermediate task training, they chose four different datasets, one of them MNLI (Phang et al., 2019), which also served as one of the nine target tasks for evaluation (Phang et al., 2019). As a result, even when they used downsampled datasets for training, the performance of the models improved compared to GLUE (Phang et al., 2019). GLUE is the "General Language Understanding Evaluation" by Wang et al. (2018, p. 353) and covers nine tasks in the field of natural language understanding and consists of datasets with a variety of sentence inputs (Wang et al., 2018, p. 353).

To get a better picture of the "when and why intermediate-task training is beneficial" (Pruksachatkun et al., 2020, p. 5231), Pruksachatkun et al. (2020) implemented an empirical approach. To answer the when-question, they train on various intermediate tasks: "it can be answered by a sufficiently exhaustive search over possible intermediate-target task pairs" (Pruksachatkun et al., 2020, p. 5231). Their definition of "good intermediate tasks" (Pruksachatkun et al., 2020, p. 5236) is of such as that they "[...] lead to positive transfer in target task performance" (Pruksachatkun et al., 2020, p. 5236). Pruksachatkun et al. (2020) pair 110 intermediate and target tasks with the modified transformer model RoBERTa (Liu et al., 2019b) and they summarize that reasoning and inference tasks are quite well suited for intermediate training (Pruksachatkun et al., 2020, p. 5236). Pruksachatkun et al. (2020) observe that intermediate task training is most favorable when there is only target task data up to a certain size (Pruksachatkun et al., 2020, p. 5236).

As for the why-question, they focus on the skills a model learns that are helpful (Pruksachatkun et al., 2020, p. 5231) – however, they claim that this question is challenging to answer yet (Pruksachatkun et al., 2020).

2.3.5 Intermediate Task Training for Cross-Lingual Transfer

To evaluate whether intermediate task training also works for non-English environments, Phang et al. (2020) create a zero-shot cross-lingual setup. For evaluation, they use the *XTREME* benchmark (Hu et al., 2020), a benchmark that consists of nine tasks in 40 languages from twelve language families (Hu et al., 2020, chap. 1) and that is particularly designed for cross-lingual transfer learning (Hu et al., 2020, chap. 1). The training data is provided in English, the above mentioned XNLI and PAWS-X are two of the nine different tasks (Hu et al., 2020).

Since intermediate task training performs well in sequential training setups for English, they adapt this setup to the cross-lingual zero-shot setting (Phang et al., 2020). The setup consists of the following steps: a masked language modeling-pre-trained multilingual language encoder is implemented, followed by intermediate task training and fine-tuning on the target task, both in English. Then they evaluate on the target task, but in the desired language (Phang et al., 2020, p. 557).

As a parallel experiment, Phang et al. (2020) also train on translated intermediate task data in German, Russian, and Swahili. The tasks they chose for intermediate training vary from question answering to natural language inference (Phang et al., 2020). As for their pre-trained encoder they opted for XLM-R (or XLM-RoBERTa, Conneau et al., 2020). XLM-R is a transformer model pre-trained on a variety of common crawl data (Conneau et al., 2020), which is also their baseline model (Phang et al., 2020). All other models train in a second phase on intermediate task(s) (Phang et al., 2020).

As for their results, models trained on MNLI for intermediate task performed best (Phang et al., 2020). However, this only applies to the English intermediate task data: "surprisingly, even when evaluating in-language, using target-language intermediate-task data does not consistently outperform using English intermediate-task data in any of the intermediate tasks on average" (Phang et al., 2020, p. 563).

3 XOpus-QIA

To study the understanding of indirect answers in a cross-lingual context, this thesis introduces the *XOPUS-QIA* corpus. This dataset makes use of the benefits of the OPUS OpenSubtitles collection – a large collection of parallel, aligned data in a variety of languages (Lison and Tiedemann, 2016), and to a significant amount human translated. Subtitles in OPUS are from different series and movies and a lot of genres are available, representing a large number of linguistic variation (Lison and Tiedemann, 2016).

The subtitles of the different, available languages are aligned to mostly English, but also to other languages than English, using the available timestamps and applying an algorithm for alignment to it (Lison and Tiedemann, 2016). The result is 1,689 so called „bitexts“ (Lison and Tiedemann, 2016, p. 927), with 62,200 aligned sentence-pairs for English and Spanish, and 43,900 aligned sentence-pairs for English and Spanish (Lison and Tiedemann, 2016). However, one limitation should be mentioned beforehand. Lison and Tiedemann (2016) emphasize that alignment can lead to multilingual corpora (Lison and Tiedemann, 2016, p. 928). Despite this, as a restriction, they indicate that "[...] it is not always possible to find links across more than two languages because different subtitle alternatives may be chosen for different language pairs" (Lison and Tiedemann, 2016, p. 928). Which is the case for the French and Spanish subtitles, since both are aligned to English, but there is no cross-lingual alignment between French-Spanish-English.

3.1 Preliminaries

To extract yes/no-questions and their respective indirect answers, the raw data is downloaded from the OpenSubtitles website[1] in English, French, and Spanish, along with the sentence alignment files for English-French and English-Spanish.

The raw data is stored in files that represent a certain year, with 2015 being the most current year for the English data. To collect question-answer pairs that reflect contemporary language use, subtitles from the most recent year 2015 are chosen. In every year folder, there are various subfolders. Each subfolder represents a specific series or movie. Again, each subfolder contains at least one subtitle file. Both the authors and translators of the subtitles and the quality of the translated data vary within the corpus.

3.2 Data Collection

For the data collection task, various subtitles from two different genres are used:

- comedy
- crime, drama, mystery

On the one hand, choosing a genre possibly allows further conclusions, and at the same time a more varied corpus is expected when using more than one genre for question-answer-pair extraction. For this reason, we choose the genres comedy and crime, drama, mystery, because they are quite contrasting; as for the comedy genre, jokes, irony, and also sometimes quite situational humour was observed and it was quite important to follow the context of dialogues to understand what was supposed to be funny.

[1] `https://opus.nlpl.eu/OpenSubtitles-v2018.php`, availability last checked on 23-06-04.

In total 18 different subtitles files are browsed in English for yes/no-questions and their respective indirect answer (QIA pairs), among them eight files from the genre crime, drama, mystery, and ten from comedy. One of the challenges in the data collection process is to find subtitles that are not only aligned between English-French and English-Spanish, but also between French-Spanish. Eventually, all subtitle files used in this thesis for the QIA pairs are inter-lingual aligned.

Although prior works on understanding indirect answers mentioned the extraction of *polar questions* (Louis et al., 2020; Damgaard et al., 2021), in this thesis only the term *yes/no-question* is used to not limit the work to polar questions only, as they have a stricter definition, and the dataset is supposed to cover a variety of indirect answers.

Before a QIA pair is extracted, the metadata at the end of each file is first checked to ensure that the subtitles are human translated. This is the case, when the field for machine translation is marked with a 0 in the subtitle file. It is also ensured via the metadata provided that the original language of the subtitle is English – or, if the language is not indicated in the metadata, the country of origin must be an English-speaking country such as Australia or the USA – so that French and Spanish translations are compared to an English original version.

As mentioned, all types of questions that can be answered with a *yes* or a *no* are extracted as yes/no-questions. Orthography rules out grammar in this case, that means that if a phrase ends with a question mark and can be answered with a *yes* or a *no*, it is extracted as a question. As for indirect answers, all answers without an explicit *yes* or *no* are collected. Derivatives of *yes* and *no* such as *yeah* or *nope* are also considered as indirect answers. Along with the QIA pairs, some metadata is extracted, in particular order (see table 3.1): the year of the folder, the ID of the series or movie (Movie-ID), the subtitle ID (here called Doc-ID), and the genre. For questions and answers, the sentence IDs (Sentence-ID, Answer-ID) are also extracted. The respective ID belongs to the first line where the question or answer starts.

Language	**Number**	**Year**	**Movie-ID**	**Doc-ID**	**Genre**	**Sentence-ID (Start Question)**	**Question**	**Answer-ID (Start Answer)**	**Answer**
English	0	2015	3591512	6461786	Crime, Drama, Mystery	<s id=”22”>	The coast?	<s id=”23”>	Audrey’s been telling me about it from her magazines - haven’t you, Audrey?
	1	2015	3591512	6461786	Crime, Drama, Mystery	<s id=”23”>	Audrey’s been telling me about it from her magazines - haven’t you, Audrey?	<s id=”24”>	Some Hollywood film star was supposed to have bought it, but no, it’s Mr and Mrs Owen.
	2	2015	3591512	6461786	Crime, Drama, Mystery	<s id=”27”>	An island? By the sea?	<s id=”29”>	Islands are generally by the sea.
French	0	2015	3591512	6442019	Crime, Drama, Mystery	<s id=”19”>	La côte ?	<s id=”20”>	Audrey l’a lu dans ces magazines, n’est-ce pas, Audrey ?
	1	2015	3591512	6442019	Crime, Drama, Mystery	<s id=”20”>	Audrey l’a lu dans ces magazines, n’est-ce pas, Audrey ?	<s id=”21”>	Une star d’Hollywood était censé l’avoir acheté, mais non, M. et Mme Owen l’ont fait.
	2	2015	3591512	6442019	Crime, Drama, Mystery	<s id=”24”>	Une île ? Près de la mer ?	<s id=”26”>	Les îles sont généralement près de la mer.
Spanish	0	2015	3591512	6439216	Crime, Drama, Mystery	<s id=”19”>	¿La costa?	<s id=”20”>	Audrey me ha estado hablando de ella por las revistas, ¿a que sí, Audrey?
	1	2015	3591512	6439216	Crime, Drama, Mystery	<s id=”20”>	Audrey me ha estado hablando de ella por las revistas, ¿a que sí, Audrey?	<s id=”22”>	Se suponía que una estrella de Hollywood iba a comprarla, pero no, el señor y la señora Owen.
	2	2015	3591512	6439216	Crime, Drama, Mystery	<s id=”25”>	¿Una isla? ¿En el mar?	<s id=”27”>	Las islas generalmente están en el mar.

Table 3.1: Example for extracted QIA pairs from the OpenSubtitles dataset (Lison and Tiedemann, 2016) with metadata.

Challenges and Peculiarities During the extraction process of the QIA pairs one difficult challenge is encountered, which is due to the document structure: speaker turns are not necessarily indicated (see an example file in appendices section). This makes it difficult to define where a question ends and an answer begins. To ensure that all QIA pairs are extracted correctly, each QIA pair is matched with the corresponding image material during the collection process. Although this task is time-consuming, it seems necessary to ensure the quality of the QIA pairs.

For English, a total of 615 QIA pairs are collected, 292 of them from the genre crime, drama, mystery, 323 from the genre comedy.

For French and Spanish, the exact same QIA pairs are extracted. As a note to add here: when a yes/no-question is translated, the translation in French or Spanish is not necessarily a question. In the collection process were encountered some imperatives, as well as a *yes* and/or a *no* in the supposed indirect answer, when during the translation process the indirect answer has turned into a direct answer. Nonetheless, the exact same QIA pairs in French and Spanish are extracted as in English, even though translations shifted the characteristics of such. Therefore, this thesis will strictly follow the alignment files for English-French and English-Spanish to extract the exact same QIA pairs in all three languages.

As seen in table 3.1 the Sentence IDs for English, French, and Spanish are not necessarily identical. The metadata for French and Spanish changes regarding Sentence- and Answer-ID as well as for the Doc-ID, since each subtitle file has its own. What remains the same in all three languages is the Movie-ID and the Year.

The French and Spanish QIA pairs are only used for evaluation, therefore we provide fewer QIA pairs than in English, due to limited personal resources. For French and Spanish, 444 QIA pairs each are collected, 205 for the genre crime, drama, mystery, and 239 for the genre comedy (see table 3.2).

Language	**Dataset**	**Extracted QIA pairs**	**Final QIA pairs**
English	Comedy	323	323
	Crime	292	292
	all	615	615
French	Comedy	239	235
	Crime	205	203
	all	444	438
Spanish	Comedy	239	235
	Crime	205	203
	all	444	438

Table 3.2: Total amount of QIA pairs extracted for the genres comedy and crime, drama, mystery (here crime). *All* refers to the total of QIA pairs extracted (both genres combined). Due to data cleaning, the numbers in the *Final QIA pairs* row vary; those are the QIA pairs used for evaluation.

3.3 Data Cleaning

The translated, aligned data in French and Spanish not only shows irregularities when it comes to indirect answers that are translated as direct answers, or yes/no-questions that appear as an imperative sentence in French. There are also further anomalies. Some sentences that appear in English do not appear at all in the aligned files in French or Spanish. Those irregularities are excluded from the French and Spanish QIA pairs. That means, when in English there is a complete QIA pair, but in French or Spanish the question or answer is missing, the QIA pair is deleted from the French and Spanish dataset – not from the raw data, however. The result is that for French and Spanish, even though in each language there are different QIA pairs that are erroneous, the same amount of incomplete QIA pairs was detected and will not be included in the final evaluation data. For French and Spanish, there is a total of 438 QIA pairs that remain for evaluation, 203 each for the genre crime, drama, mystery, and 235 each for the genre comedy (see table 3.2).

3.4 Data Annotation

To label the data, we use a slightly altered version of the *relaxed* label set (see again table 2.1), that was introduced by Louis et al. (2020) and already adapted by Damgaard et al. (2021). We changed the "N/A" label used by Louis et al. (2020, p. 7415) though, since it marks annotator disagreement. In this thesis, the data is mostly annotated by one person only, which makes this label redundant. However, there still will be a label 6 that completes label 5, the "other" label (Louis et al., 2020, p. 7415). "Other" is given when the indirect answer to a yes/no-question does not refer to the question. Label 6 in this thesis will be "lacking context". It is given when the indirect answer to a yes/no-question can not be classified properly, because the conversation situation is confusing or unclear. This was not necessary in Louis et al. (2020)'s work, since they did not extract their question-answer pairs from real, existing conversations but generated them solely for their purpose. As mentioned above, one of the challenges during data collection are the missing turns in the subtitle files of OpenSubtitles. With the "lacking context" label it will be acknowledged that even though the extracted QIA pairs are accurate regarding the speaker turn, they do not necessarily make sense to a person that only reads question and answer without having the corresponding moving images at hand.

The labels used in this thesis are defined as follows:

> *Yes (label 1)* – will be defined here as every answer that can be taken as a *yes,* even if it is not a clear *yes*, but more a *maybe yes* or a *yes* in a weakened form (Damgaard et al., 2021, p. 4).
>
> *No (label 2)* – *no* in comparison to *yes* is every answer that is clearly a *no* or all gradations of *no* (Damgaard et al., 2021, p. 4).
>
> *Yes, subject to some conditions (label 3)* – in this case, the answer means *yes*, but with the restriction that only under certain circumstances (Damgaard et al., 2021, p. 4).
>
> *Neither yes nor no (label 4)* – a label for "in the middle" answers (Louis et al., 2020, p. 7415), when the indirect answer cannot be classified in the binary *yes* or *no* scheme (Damgaard et al., 2021).
>
> *Other (label 5)* – this label marks the situation, when the indirect answer does not match the question (Damgaard et al., 2021, p. 4).
>
> *Lacking context (label 6)* – as mentioned above, this label is used when the answer cannot be clearly categorized as *yes* or *no* simply because the context is missing or unknown to the annotator.

Since the language data in XOpus-QIA is aligned, the annotation process is only done on the English data and the labels are copied to the French and the Spanish QIA pairs.

3.5 Annotation Process and Annotator Agreement

The complete XOpus-QIA with 615 QIA pairs has been annotated by only one annotator (annotator 1, the author of this thesis), hence those annotations also count as the gold standard within the scope of this thesis. However, as a pilot, for approximately 200 QIA pairs in English (from both genres), labeled data from a second annotator (annotator 2, an external annotator) is available. As for the annotators, there is an important difference that must be taken into account when comparing the labeled data: annotator 1 has seen the moving image material that corresponds to the respective subtitle files and QIA pairs, annotator 2 has not.

For the complete XOpus-QIA dataset in English, annotated by annotator 1, the label distribution is quite uneven (see table 3.3).

Language	QIA pairs	Label 1	Label 2	Label 3	Label 4	Label 5	Label 6
English	all (615)	35.61	13.01	1.95	12.36	24.39	12.52
	Crime (292)	35.62	13.36	1.37	15.75	21.92	11.99
	Comedy (323)	35.60	12.69	2.48	9.29	26.93	13.00
French	all (438)	34.93	14.16	1.60	11.42	24.20	13.47
Spanish	all (438)	35.16	14.16	1.60	11.19	24.20	13.70

Table 3.3: Label distribution of the XOpus-QIA dataset in percentages.

The "yes" label is the most frequent one, with 219 QIA pairs it makes up 35.61 percent of the data. The second frequent label is "other" (24.39 percent, 151 examples in total) and the "no" label occurs less than half as often as the "yes" label (80 counts in total, 13.01 percent). The distribution of the most frequent "yes" label is even stable across genres (see again table 3.3). As for the genre crime, drama, mystery, the distribution of "yes" is at 35.62 percent, for comedy at 35.60 percent. For French and Spanish those percentages in the label distribution vary slightly, since the dataset in total is smaller than the English one and a data cleaning process was carried out.

The label distribution looks quite different when the labeled data from annotator 1 is compared to the labeled data from annotator 2 (see figure 3.1). The plotted figure for label distribution is generated using the data analysis library pandas (Wes McKinney, 2010) and the plotting library Matplotlib (Hunter, 2007).

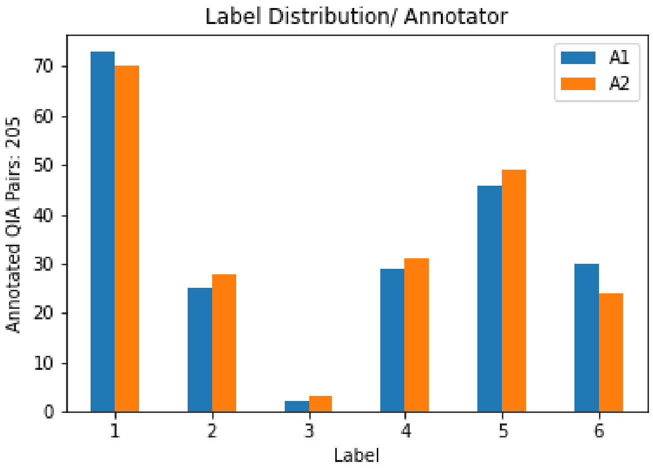

Figure 3.1: Label Distribution for 205 QIA pairs in English, annotated by two annotators, annotator 1 (A1), and annotator 2 (A2).

Annotator 2 labeled the data twice. For the first annotation round, annotator 2 was provided with the label set as well as some theoretical background – including the concept of polar questions and indirect answers. This resulted in only 172 annotated QIA pairs out of 207 QIA pairs provided for annotation, because some QIA pairs did not contain polar questions in a strict sense. This is one of the reasons why the concept has been extended to yes/no-questions. After this clarification, annotator 2 re-labeled the data and only two QIA pairs out of 207 remain unlabeled. Even though the label distribution between annotator 1 and annotator 2 varies, the frequency distribution of the label is at least similar. In both cases "yes" is the most frequent label, followed by "other". For a more detailed confusion matrix see figure 3.2. The confusion_matrix function by scikit-learn (Pedregosa et al., 2011) is used to compute the output.

If inter-annotator agreement scores are calculated for both annotators, it becomes perceptible how challenging the task of identifying the implied response in an indirect answer is: for the total of 205 English QIA pairs that have been labeled by both annotators, the

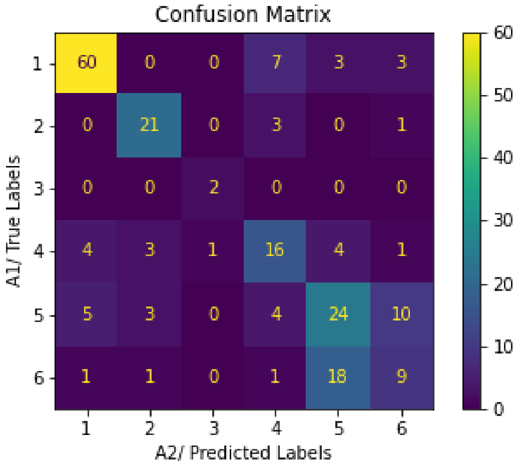

Figure 3.2: Confusion Matrix for 205 QIA pairs and two annotators.

observed agreement score is at 0.64. Whereas if the problem of agreeing not because of agreeing but of chance (Cohen, 1960) is ruled out, the Cohen's Kappa score[2] is at only 0.54, even though it still "[. . .]represent[s] a fair to good level of agreement [. . .]" (Green, 1997, p. 5).

More surprisingly, those scores vary enormously between the two genres provided. For the crime, drama, mystery genre, the observed agreement is at 0.53, the Cohen Kappa at 0.38. A score that even after Green (1997)'s definition is quite low. It is the complete opposite for the comedy genre though; observed agreement is at 0.78, and Cohen Kappa at 0.72. It reaches almost the 0.75 score line for high agreement (Green, 1997). This might be due to the somewhat difficult situational dialogues in the comedy genre, as mentioned in 3.2. Label 6 and label 5 are more frequent in the comedy genre and none of them stands for an answer themselves. As seen in figure 3.3 and compared to 3.4, regarding the comedy genre, both annotators highly agree on label 5, "other". In this case, both of them interpret that there is no clear answer.

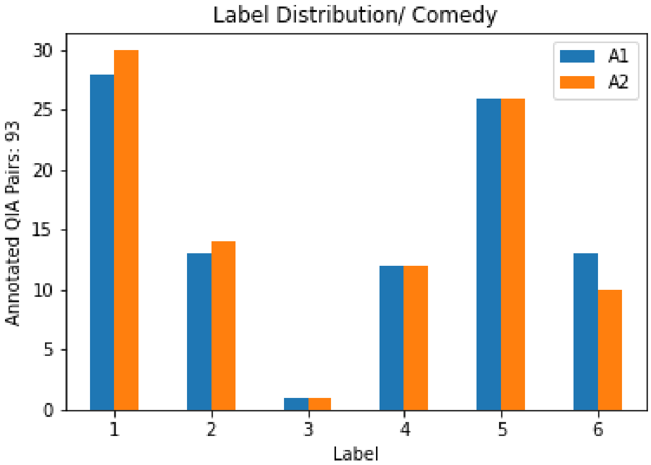

Figure 3.3: Label Distribution for 93 QIA pairs in English, for the genre comedy, annotated by two annotators.

[2] The Cohen Kappa scores are calculated using pandas (Wes McKinney, 2010) and the metrics module of the Natural Language Toolkit (NLTK) library (Bird and Loper, 2004), both assembled and ready to use in a code template by Buntain (2020).

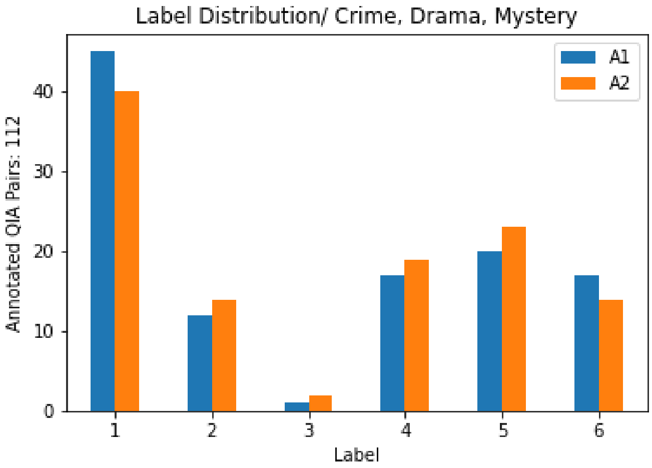

Figure 3.4: Label Distribution for 112 QIA pairs in English, for the genre crime, drama, mystery, annotated by two annotators.

3.6 How XOpus-QIA is used

The XOpus-QIA[3] is used both for training and evaluation, depending on the language. The French and Spanish QIA pairs are only used for evaluation in this thesis. The English QIA pairs are also used for training respectively fine-tuning, it varies from model to model. Therefore, the English dataset is split up into a train, development and test set. For this task, the train_test_split function by scikit-learn (Pedregosa et al., 2011) is applied, with a random state of 42 and shuffle set to True. The data is split up to 80 percent for the training set, and 10 percent each for the development and test set. As for the English XOpus-QIA, this results in a distribution of 492 of the 615 QIA pairs for the training set, 61 QIA pairs for the development set, and 62 QIA pairs for the test set. The files are saved in a tab separated values (tsv) format, since that is a requirement for training with the toolkit that will be used in this work.

[3] The full dataset is available at: `https://github.com/Christin-M/XOpus-QIA`.

4 Models

4.1 MaChAmp Toolkit

To build the baseline and cross-lingual transfer learning intermediate task training models, the MaChAmp[1] (van der Goot et al., 2021) toolkit is used in the version 0.4. MaChAmp is useful for a variety of NLP tasks and can be easily used via the command-line (van der Goot et al., 2021). Training or fine-tuning a model can be done in one line of command, by providing a task-adapted configuration file for one of the desired NLP tasks (van der Goot et al., 2021). A wide range of those tasks is available – including classification, that will be needed in this thesis for making progress in understanding indirect answers (van der Goot et al., 2021). A MaChAmp model consists of both encoder and decoder, whereas the encoder is a pre-trained model, and the decoder is task-specific (van der Goot et al., 2021).

A variety of classification models will be trained using MaChAmp. This includes that for every model (both baseline and intermediate task training), a configuration file and the input data (for training, intermediate task training, and fine-tuning) are provided.

As for the parameters, all models implemented using MaChAmp share the same parameter settings, since the default hyperparameters were found to be robust in multiple tasks (van der Goot et al., 2021). In this case, the multilingual BERT transformer model is enabled and every dataset for training, intermediate task training, and fine-tuning trains for 20 epochs each (see example of a params.json file in the appendices section).

To specify the sentence classification task, the configuration files (an example of a configuration file can also be found in the appendices section) all are set to the classification task and are linked to data paths for training and development. The "sent_idxs" line indicates in which column the sentences – in this case, the question and answer – are to be found, the "column_idx" line tells the model in which column the labels are located.

The pipelines for the baseline models respectively for the cross-lingual transfer learning intermediate task training models are configured as follows: both model types rely on mBERT, which is already pre-trained with MaChAmp. As for the baseline models, we then train (respectively fine-tune) mBERT on the English training and development set of the XOpus-QIA corpus.

As for the intermediate task training models, each model first trains sequentially on an intermediate task, that is, any task that is not the understanding indirect answers task, then fine-tunes on the target task in English, which is the understanding indirect answers task. The third step consists of evaluating the model on the same target task, this time on data in French and Spanish (see figure 4.1 for a schematic overview).

Prediction on a MaChAmp trained model is as simple as training and also done via the command-line. Given an input file in tsv-format and a model to be tested, the evaluation scores along with the labeled output file are easily retrieved. For additional metrics, such as the F1-scores per label, the sklearn.metrics module (Pedregosa et al., 2011) is used.

4.2 Model Overview

For each model considered for evaluation, the pipeline setup will be described in this chapter, including which data is used, and – if not mentioned before in this thesis – how the data is prepared. Every baseline model trains for 20 epochs, every intermediate task

[1] `https://github.com/machamp-nlp/machamp`, availability last checked on 23-06-04.

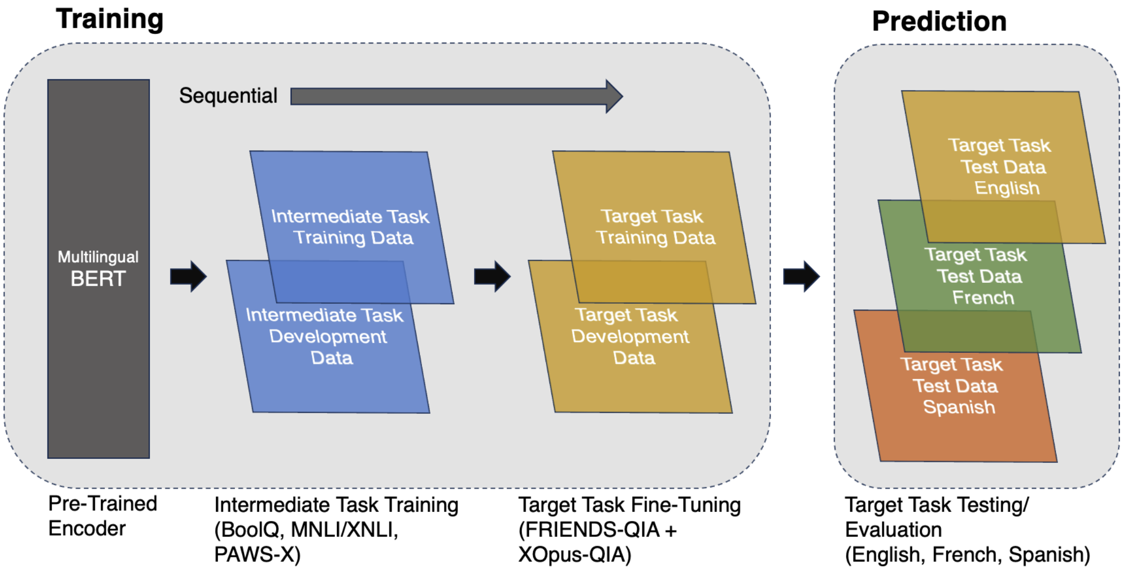
Training
Prediction
Sequential
Multilingual BERT
Intermediate Task Training Data
Intermediate Task Development Data
Target Task Training Data
Target Task Development Data
Target Task Test Data English
Target Task Test Data French
Target Task Test Data Spanish
Pre-Trained Encoder
Intermediate Task Training (BoolQ, MNLI/XNLI, PAWS-X)
Target Task Fine-Tuning (FRIENDS-QIA + XOpus-QIA)
Target Task Testing/ Evaluation (English, French, Spanish)

Figure 4.1: Overview of the sequential intermediate task training setup for the understanding indirect answers classification task with MaChAmp (van der Goot et al., 2021). Figure adapted from Phang et al. (2020).

training model for 20 epochs on each dataset – hence on 40 epochs in total. The best development scores for the best epoch reported here refer to the last dataset the model has been trained on. All models predict on seven different files: first, on all QIA pairs available for English, French, and Spanish, then on the respective genre-specific QIA pairs for French and Spanish.

4.2.1 Baseline

To compare the results of the transfer learning intermediate task training models, three different zero-shot baseline models will be considered. Each model is trained on a different dataset for indirect answers in English, and evaluated on data in English, French, and Spanish. The comparison of the three different baseline models will be discussed in subsection 4.2.2 to choose suitable target task data for the transfer learning intermediate task training models.

XOpus-QIA

The XOpus-QIA baseline model trains on the English XOpus-QIA dataset only, on the 492 QIA pairs retrieved from the train split (see section 3.6) and the 61 QIA pairs retrieved from the development split. The best results are retrieved in epoch 14, with a development score of 0.5738 (see table 4.1). For evaluation, the test set of XOpus-QIA for English is used, as well as the French and Spanish XOpus-QIA pairs for zero-shot evaluation.

Model	Train Loss	Dev Loss	Train Scores	Dev Scores
Baseline				
XOpus-QIA	0.4441	49.6954	1.0000	0.5738
FRIENDS-QIA-E6	1.6629	56.9412	0.9816	0.6448
FRIENDS-QIA + XOpus-QIA	0.7344	72.0100	0.9911	0.6172

Table 4.1: Model predictions for the baseline models.

FRIENDS-QIA-E6

For this baseline, the FRIENDS-QIA dataset[2] by Damgaard et al. (2021) is used. The data is available in csv-format, which is converted to tsv for usage in MaChAmp. Since the label set used in this thesis is based upon the label set by Damgaard et al. (2021), adapted from Louis et al. (2020), the FRIENDS-QIA can be used with only little adjustments. In the FRIENDS-QIA dataset, the label 6 is for the class where no annotator agreement was achieved, in the XOpus-QIA label 6 is for the lack of context. For this baseline model, the label "N/A" (label 6) (Damgaard et al., 2021) will be excluded from FRIENDS-QIA and the same 80-10-10 train-dev-test split is used as for the XOpus-QIA. For the train set, this results in 4,994 QIA pairs, and the development data includes 625 QIA pairs. For prediction, the English, French and Spanish QIA pairs from the XOpus-QIA are used, and again, evaluation is done on the full datasets first, then on the QIA pairs per genre. The best results for this model are also retrieved in epoch 14, the development score for this epoch is 0.6448 (see table 4.1).

FRIENDS-QIA + XOpus-QIA

This baseline model combines the FRIENDS-QIA corpus by Damgaard et al. (2021) with the English data of the XOpus-QIA. For this model, the "N/A" label used by Damgaard et al. (2021) is once again excluded from the FRIENDS-QIA. However, the "lacking context" label of XOpus-QIA remains in the corpus, so that during the training process this label will still be learned. Again, we implement the same 80-10-10 train-dev-test split as in the two previous models. The training set is made up of 5,486 labeled QIA pairs, the development and test set have 687 QIA pairs each. The best development score of 0.6172 (see table 4.1) is retrieved in the 16th epoch.

4.2.2 Notes and Interim Evaluation

At this point, a few of the evaluation results will be discussed beforehand, to determine and justify which dataset will be used for target task fine-tuning for the transfer learning intermediate task training models.

So far, if the three zero-shot baseline models are compared, the XOpus-QIA model, the FRIENDS-QIA-E6 model, and the FRIENDS-QIA + XOpus-QIA model, there are certain results that need clarification (see table 4.2).

Model	**Training Data**	**Test Data**	**Accuracy**	**Macro F1**
Baseline				
XOpus-QIA	XOpus-QIA (en)	en	0.4104	0.3056
	XOpus-QIA (en)	fr	**0.6050**	**0.5497**
	XOpus-QIA (en)	es	**0.6210**	**0.4850**
FRIENDS-QIA-E6	FRIENDS-QIA (en)	en	0.3710	0.2791
	FRIENDS-QIA (en)	fr	0.3014	0.2175
	FRIENDS-QIA (en)	es	0.2785	0.1711
FRIENDS-QIA + XOpus-QIA	FRIENDS-QIA + XOpus-QIA (en)	en	**0.6143**	**0.3768**
	FRIENDS-QIA + XOpus-QIA (en)	fr	0.4406	0.3364
	FRIENDS-QIA + XOpus-QIA (en)	es	0.4406	0.3395

Table 4.2: Performance scores of the baseline models. The best scores are highlighted in bold.

For evaluation data in English, the concatenated model of FRIENDS-QIA + XOPUS-QIA has the best results, with an accuracy score of 0.6143 and a F1-score of 0.3768. However, for the evaluation data in French and Spanish from the XOpus-QIA dataset, adding the FRIENDS-QIA dataset did not help: using the same distribution, the XOpus-QIA data performed best – both for accuracy and macro F1-score. Nevertheless, the

[2] `https://github.com/friendsQIA/Friends_QIA/tree/main`, availability last checked on 23-06-04.

combined FRIENDS-QIA + XOpus-QIA dataset will be used for target task fine-tuning for the following transfer learning intermediate task training models. One of the reasons is a preceding baseline model, excluded in this section from overall evaluation due to a lack of comparability. In a first attempt, the XOpus-QIA model was trained with more data than the one stated above. That was possible because the data had not previously been split into a train-dev-test-split, only an 80-20 train-dev-split was used. The prior XOpus-QIA model was then trained on the train and development data of the XOpus-QIA dataset, however no data for testing has been left in this setting – at least for English. Hence the new setup. What should be mentioned here is that the prior XOpus-QIA model led to lower accuracy and macro F1-scores when evaluated on the French and Spanish XOpus-QIA data (for example for the French QIA pairs, the accuracy prior was at 0.4589, now it is at 0.6050). The current XOpus-QIA model seems to overfit eventually, therefore the concatenated dataset of FRIENDS-QIA and XOpus-QIA (FX) will be used as target task training data for fine-tuning in the transfer learning intermediate task training models.

4.2.3 Intermediate Task Training

For the intermediate task training models, each model uses different intermediate task data for training and trains sequentially on intermediate task data and target task data, for 20 epochs each. The target task fine-tuning data is identical for each model – the concatenated FRIENDS-QIA and XOpus-QIA (FX) dataset is used, respectively the train and the development data. See table 4.1 for the scores of each model.

Model	**Model Name**	**Train Loss**	**Dev Loss**	**Train Scores**	**Dev Scores**
Intermediate Task Training					
BoolQ	BoolQ-FX	2.8447	63.0986	0.9657	0.5852
	BoolQ-FR-FX	30.0699	32.2362	0.5829	0.5837
	BoolQ-ES-FX	10.2512	44.3883	0.8682	0.6012
	BoolQ-PT-FX	20.3495	38.5458	0.7220	0.5866
MNLI/ XNLI	MNLI-FX	0.8732	68.5868	0.9907	0.5910
	XNLI-FR-FX	1.6820	64.3502	0.9776	0.6084
	XNLI-ES-FX	1.8147	67.8732	0.9787	0.5924
PAWS-X	PAWS-X-FX	24.1357	33.5635	0.6619	0.5983
	PAWS-X-FR-FX	16.3376	40.9808	0.7763	0.6012
	PAWS-X-ES-FX	1.0034	70.7338	0.9889	0.5866

Table 4.3: Model scores for the intermediate task training models, extracted in the best epoch of target task fine-tuning.

BoolQ-FX

Four different BoolQ models with the question-answer dataset BoolQ (Clark et al., 2019) and its respective translations are trained. Since the BoolQ dataset is made up of yes/no-questions and answers, it is assumed that this might be a good task for intermediate task training regarding the understanding indirect answers task, since the model can learn from a similar dataset. The original BoolQ dataset is available in English and the train split consists of 9,430 question-answer pairs, the development split of 3,270 question-answer pairs. The data is stored in jsonl-format, which is converted to a tsv-format. Although stated otherwise, the downloaded jsonl-file of the training data only contains 9,427 question-answer pairs[3], which reduces the dataset by three examples.

The BoolQ-FX model trains on the original, English BoolQ dataset and fine-tunes on the FRIENDS-QIA + XOpus-QIA dataset. It performs best in the 14th fine-tuning epoch, with a development score of 0.5852.

[3] The BoolQ dataset has been downloaded under the following link: `https://github.com/google-research-datasets/boolean-questions`, availability last checked on 23-06-04.

BoolQ-FR-FX, BoolQ-ES-FX

The BoolQ dataset in French, along with the BoolQ dataset in Spanish, are available via the Hugging Face Datasets library[4] (Lee, 2023b,a). Unfortunately, the train and validation set are not human translated, but translated using the respective open translation service Opus-MT models (Tiedemann and Thottingal, 2020) for each language.

As a second issue, the currently available, uploaded validation data for French and Spanish are both identical to the training data. Despite asking for the correct validation set, to the present date[5] it has not been uploaded. This problem has hence been solved temporarily by re-splitting up the train sets in French and Spanish into a train and a development set. The remaining question-answer pairs for French and Spanish are for the training data 7,541, 1,886 for the development data. Both models, the BoolQ-FR-FX with the French train and development data, and the BoolQ-ES-FX with the Spanish train and development data, are trained and fine-tuned equivalently to the BoolQ model that uses the original dataset in English.

The BoolQ-FR-FX model performs best in the 4th fine-tuning epoch at a development score of 0.5837. The BoolQ-ES-FX model peaks at 0.6012 in epoch 9 of fine-tuning.

BoolQ-PT-FX

Since the BoolQ-FR-FX and the BoolQ-ES-FX model both encounter certain data issues, a BoolQ model has also been trained on data in Portuguese. The Portuguese version of BoolQ is also available via the Hugging Face Datasets library[6] (Lee, 2023c). It is also translated with Opus-MT (Tiedemann and Thottingal, 2020) in the English-Portuguese version. The difference to the French and Spanish BoolQ models is the availability of the development set. For Portuguese, the development set is not identical to the training set. Since French, Spanish, and Portuguese are all three Romance languages and thus grammatically related and similar, a model trained with Portuguese data will be evaluated for French and Spanish. The number of question-answer pairs is identical to English, with 9,427 pairs for the training set, and 3,270 pairs for the development set.

The BoolQ-PT-FX model thus takes the Portuguese training and development data for intermediate task training, fine-tuning is also done on the FRIENDS-QIA + XOpus-QIA dataset. The model peaks in epoch 6 with a development score of 0.5866.

MNLI-FX

Even though it is an English dataset for inferences, MNLI (Williams et al., 2018) already performed decently in transfer learning approaches for question-answering models, more precisely for BoolQ (Clark et al., 2019). The authors chose BERT (Devlin et al., 2019) as unsupervised model, which not only leads to the best accuracy scores, it outperforms all models, even the one that trains on MNLI as transfer task (Clark et al., 2019). As for this thesis, we hope to gain similar results when we first train on the MNLI dataset, then fine-tune on the FRIENDS-QIA + XOpus-QIA dataset.

The MNLI dataset consists of 433,000 pairs of hypothesis, premise, and label for textual entailment. Therefore, the dataset needs to be downsampled to meet limited hardware capacities for training. First, the data is converted to a tsv-format, since the MNLI data is available in a txt-format[7]. Then the data is downsampled using scikit-learn's resample function (Pedregosa et al., 2011) with a random state of 42 and shuffle set to True. The n_samples parameter is set to 12,696, which results in the same training and development set size as for BoolQ.

[4] `https://huggingface.co/reaganjlee`, availability last checked on 23-06-04.

[5] 23-06-04.

[6] `https://huggingface.co/datasets/reaganjlee/boolq_pt`, availability last checked on 23-06-04.

[7] `https://cims.nyu.edu/~sbowman/multinli/`, availability last checked on 23-06-04.

The MNLI-FX model pipeline then looks as follows: it trains on the downsampled MNLI training and development sets and fine-tunes on FRIENDS-QIA + XOpus-QIA. After 16 epochs of sequential training, the best development score is recorded, being at 0.5910.

XNLI-FR-FX, XNLI-ES-FX

The XNLI dataset[8] (Conneau et al., 2018) is based upon the MNLI dataset and contains human translated data for several languages, including French and Spanish. Translations only are available for development and test data, though. This thesis wants to explore the benefit of XNLI and compare the results on this data to the English-only MNLI-FX model.

For the French XNLI-FR-FX and the Spanish XNLI-ES-FX model, for the intermediate task training, the same training data file is used – the English MNLI downsampled data file. However, the development data files are different to the MNLI-FX model, since the translated XNLI development data for French and Spanish is used. For each language represented in the XNLI dataset, there are 2,490 premise-hypothesis pairs. The models sequentially fine-tune on the FRIENDS-QIA + XOpus-QIA dataset, similar to all intermediate task training models.

For the English/French XNLI-FR-FX model, the best performance is reported in epoch 14 with a development score of 0.6084. The best development score for the English/Spanish XNLI-ES-FX model is measured in epoch 14, but with a development score of 0.5924.

PAWS-X-FX

Even though the PAWS dataset (Zhang et al., 2019) focuses on paraphrases, it will be still assumed in this thesis that the translated PAWS-X (Yang et al., 2019) can help in understanding indirect answers, since the syntactic structure provided in the paraphrases has an impact on interpretation.

Three different models are trained using the available PAWS-X language datasets[9]. One in English, one in French, one in Spanish. As for the training data, the paraphrase datasets for each language will be as well downsampled. Therefore, the same downsampling process is adapted for the PAWS-X training sets as used for the MNLI training data.

This results in 9,427 paraphrase pairs in the training data file, while the human translated development data only contains 2000 sentence-paraphrase pairs. Even though at least for English there would be more pairs available for development from the original PAWS, the 2,000 pairs for development from the PAWS-X dataset will be used, to have the same amount for each language. Each PAWS-X model, the English PAWS-X-FX, the French PAWS-X-FR-FX, and the Spanish PAWS-X-ES-FX, are trained during intermediate task training with 9,427 sentence paraphrase pairs for training and 2,000 pairs for development. The target task training data for fine-tuning is FRIENDS-QIA + XOpus-QIA.

The English PAWS-X-FX performs best in the 5th epoch at a development score of 0.5983, the French PAWS-X-FR-FX got the best development score of 0.6012 in epoch 7, and the Spanish PAWS-X-ES-FX in epoch 17 at a development score of 0.5866.

4.2.4 Final Remarks on the Models

A lot more models have been trained for this thesis than the ones specified above. Models, not mentioned here in detail, because they were canceled during training, because something was odd with the datasets, or because they were not relevant for the evaluation. For example, as a baseline, two different FRIENDS-QIA models were trained. The

[8] `https://github.com/facebookresearch/XNLI`, availability last checked on 23-06-04.

[9] `https://github.com/google-research-datasets/paws/tree/master/pawsx`, availability last checked on 23-06-04.

FRIENDS-QIA dataset is available in two versions: a modified one and a raw version with no modifications (Damgaard et al., 2021). Damgaard et al. (2021) used for their models the modified dataset. However, a model on the raw data was trained for this thesis, just to ensure to use the best performing dataset. The modified dataset – as intended by the authors (Damgaard et al., 2021) – outperformed the original, raw dataset, so the modified one is used as well in this thesis.

There were also various BoolQ models trained. As for the French and Spanish versions of the BoolQ models, the results of both models seemed doubtful. Both peaked at the first epoch of intermediate task training and then the values decreased strongly and continuously, and they both fine-tuned on the FRIENDS-QIA + XOpus-QIA with the same results. The datasets have been checked and it was observed that the development sets were almost identical to the training sets – hence the BoolQ models for French and Spanish were trained twice, the second time with the reduced training data as explained in subsection 4.2.3.

There also has been an attempt with the original-sized MNLI dataset. Since the first two epochs during intermediate task training already trained for more than 48 hours each, the downsampled option was chosen.

5 Results and Discussion

In this chapter the results are discussed. First, the baseline models are compared, followed by the intermediate task training models. We then analyze the results per genre and the F1-scores per label. If not stated otherwise, all metrics mentioned refer to the complete datasets that contain QIA pairs for both genres (that includes 615 QIA pairs for English, and 438 QIA pairs for French and Spanish each).

5.1 Baseline

The main results of the baseline models are already discussed in subsection 4.2.2, however, a more detailed explanation will be given at this point. The FRIENDS-QIA + XOpus-QIA model achieves the best performance for the English test data, both for accuracy and macro F1-score (see table 4.2 for the baseline performance scores). For the French and Spanish test data, the XOpus-QIA baseline model performs best with an accuracy of 0.6050 for French and an accuracy of 0.6210 for Spanish, and macro F1-scores of 0.5496 for French and 0.4850 for Spanish.

If the data for testing is divided by genre, the results are quite similar when it comes to ranking the models (see table 5.1), even though the crime, drama, mystery genre outperforms the comedy genre. The XOpus-QIA is still the model that performs best for both French and Spanish QIA pairs, for both genres.

Model	**Test Data**	**Genre**	**Accuracy**	**Macro F1**	**Genre**	**Accuracy**	**Macro F1**
Baseline							
XOpus-QIA	fr	Comedy	**0.5830**	**0.4920**	Crime	**0.6305**	**0.6656**
	es	Comedy	**0.6298**	**0.4844**	Crime	**0.6108**	**0.4800**
FRIENDS-QIA-E6	fr	Comedy	0.2936	0.1774	Crime	0.3103	0.3499
	es	Comedy	0.2255	0.1456	Crime	0.3399	0.1981
FRIENDS-QIA + XOpus-QIA	fr	Comedy	0.3872	0.2895	Crime	0.5025	0.3898
	es	Comedy	0.3617	0.2771	Crime	0.5320	0.4113

Table 5.1: Baseline F1-scores for the genres comedy, and crime (crime, drama, mystery). Best performance scores are in bold.

If the models are tested with the entire QIA pairs in English, French, or Spanish, except for the French test set predicted with XOpus-QIA, the F1-scores barely surpass values of 0.5000 and as for the FRIENDS-QIA-E6 model, they are generally quite poor.

However, the performance results change drastically if the F1-scores are calculated per class (see table 5.2). For the English test data, the "yes" label reaches an F1-score of 0.7515 with the FRIENDS-QIA + XOpus-QIA model.

Model	**Test Data**	**F1/ Label**	**F1/ Label**	**F1/ Label**	**F1/ Label**	**F1/ Label**	**F1/ Label**
Baseline		1/ Yes	2/ No	3/ Yes (condition)	4/ Neither	5/ Other	6/ Context
XOpus-QIA	en	0.5532	0.2609	0.0000	0.1333	0.5806	0.0000
	fr	**0.7216**	0.5860	0.4444	0.4615	0.5796	0.5047
	es	**0.7000**	0.5676	0.0000	0.3415	0.6637	0.6372
FRIENDS-QIA-E6	en	0.5333	0.4211	0.0000	0.3076	0.1333	0.0000
	fr	0.5031	0.3310	0.2222	0.1946	0.0541	0.0000
	es	0.4820	0.3684	0.0000	0.1577	0.0183	0.0000
FRIENDS-QIA + XOpus-QIA	en	**0.7515**	0.6220	0.2000	0.4370	0.2500	0.0000
	fr	0.6100	0.4459	0.0000	0.2825	0.3239	0.3561
	es	0.6554	0.4260	0.0000	0.2162	0.3546	0.3846

Table 5.2: Baseline F1-scores per label. Best performing scores for "yes" are printed in bold.

For the French and Spanish test data similar values for the "yes" class can be observed. Both peak with the XOpus-QIA model at 0.7216 for French, and 0.7000 for Spanish. But for the other five classes, or labels, the values decrease, for all test sets. Some F1-scores are even 0. For the English test data, none of the three baseline models has a F1-score other than zero for label 6, "lacking context".

For the English test data, the class with the second-best performance is "no" (label 2), with a F1-score of 0.6220 achieved by the FRIENDS-QIA + XOpus-QIA model. For French, it is also the "no" label, with a F1-score of 0.5860 for the XOpus-QIA baseline. And right after, for French, scores label 5, "other", with a F1-score of 0.5796. The situation with Spanish is a little different. Label 5 and label 6 perform better than for the French test data with the XOpus-QIA model, the F1-score for "other" is at 0.6637, closely followed by a F1-score of 0.6372 for "lacking context". The "no" class for the Spanish test data, however, scores lower at only 0.5676.

Since label 1 and label 5 are the most frequent labels, for the French and Spanish data those results are at least somewhat satisfying regarding their respective label distribution. As for the English data, with a F1-score of 0 for the label 6, the results are quite poor, considering that more than 12 percent of the data QIA pairs are assigned to this class. In the predicted output data for English, label 6 does not appear at all.

5.2 Intermediate Task Training

Next, we compare the results of the intermediate task training to the baseline models. In particular, we train 10 models for evaluation. When we compare these models, there is none that clearly stands out regarding the accuracy or macro F1-scores. Although a few models do perform better than others, especially the macro F1-scores are generally quite low. Let's start with the models trained and evaluated on English data first (see table 5.3).

Model	**Model Name**	**Training Data**		**Test Data**	**Accuracy**	**Macro F1**
Baseline						
FRIENDS-QIA + XOpus-QIA	FRIENDS-QIA + XOpus-QIA	FRIENDS-QIA + XOpus-QIA (en)		en	0.6143	0.3768
Model	**Model Name**	**Intermediate Task**	**Target Task**	**Test Data**	**Accuracy**	**Macro F1**
Intermediate Task Training						
BoolQ	BoolQ-FX	BoolQ (en)	FRIENDS-QIA + XOpus-QIA (en)	en	0.6070	0.3913
	BoolQ-FR-FX	BoolQ (fr)	FRIENDS-QIA + XOpus-QIA (en)	en	0.5779	0.2883
	BoolQ-ES-FX	BoolQ (es)	FRIENDS-QIA + XOpus-QIA (en)	en	0.5968	0.4273
	BoolQ-PT-FX	BoolQ (pt)	FRIENDS-QIA + XOpus-QIA (en)	en	0.5881	0.3620
MNLI/ XNLI	MNLI-FX	MNLI (downsampled) (en)	FRIENDS-QIA + XOpus-QIA (en)	en	0.5750	0.3732
	XNLI-FR-FX	MNLI (downsampled) (en) + XNLI (fr)	FRIENDS-QIA + XOpus-QIA (en)	en	0.5822	0.3848
	XNLI-ES-FX	MNLI (downsampled) (en) + XNLI (es)	FRIENDS-QIA + XOpus-QIA (en)	en	0.5953	0.3844
PAWS-X	PAWS-X-FX	PAWS-X (downsampled) (en)	FRIENDS-QIA + XOpus-QIA (en)	en	0.5691	0.3130
	PAWS-X-FR-FX	PAWS-X (downsampled) (fr)	FRIENDS-QIA + XOpus-QIA (en)	en	0.5939	0.3708
	PAWS-X-ES-FX	PAWS-X (downsampled) (es)	FRIENDS-QIA + XOpus-QIA (en)	en	0.5706	0.3591

Table 5.3: Performance scores of the baseline and intermediate task training models evaluated on English data. Results in red determine lower scores as compared to the FRIENDS-QIA + XOpus-QIA baseline, green indicates higher performance scores.

English Test Data As for the English test data, when the intermediate task training models are compared to the FRIENDS-QIA + XOpus-QIA baseline model, they are consistently below the baseline, if accuracy is evaluated (see table 5.3). However, as for the macro F1-scores, some setups are helpful and reach higher scores. The models that stand out are two of the BoolQ models, the Spanish BoolQ-ES-FX model with an F1-score of 0.4273, and the BoolQ-FX model, trained on the original BoolQ dataset in English as intermediate task, with a macro F1-score of 0.3913. And even if the BoolQ-FX model performs lower than the FRIENDS-QIA + XOpus-QIA baseline regarding accuracy, it is the best model when comparing only the intermediate task training models for the English test data output. However, almost all intermediate task training models perform more or less equally, between an accuracy of 0.5691 as the lowest accuracy measured with the PAWS-X-FX model and the highest accuracy score of 0.6070 for the BoolQ-FX model. As an observation, the English in-language setup benefits more from the yes/no-question and answer dataset BoolQ than from the inference dataset MNLI or the paraphrase dataset PAWS-X.

French and Spanish Test Data In the cross-lingual setups, at least for French when evaluated on an intermediate task training model, the accuracy scores surpass the baseline with the MNLI or XNLI dataset. The MNLI-FX model, trained on the downsampled English MNLI data, performs not only best for French, but performs best overall (see table 5.4). In this cross-lingual transfer learning setting, it reaches an accuracy of 0.5000.

Model	**Model Name**	**Training Data**		**Test Data**	**Accuracy**	**Macro F1**
Baseline						
FRIENDS-QIA + XOpus-QIA	FRIENDS-QIA + XOpus-QIA	FRIENDS-QIA + XOpus-QIA (en)		fr	0.4406	0.3364
				es	0.4406	0.3395
Model	**Model Name**	**Intermediate Task**	**Target Task**	**Test Data**	**Accuracy**	**Macro F1**
Intermediate Task Training						
BoolQ	BoolQ-FX	BoolQ (en)	FRIENDS-QIA + XOpus-QIA (en)	fr	0.3904	0.3255
		BoolQ (en)	FRIENDS-QIA + XOpus-QIA (en)	es	0.3813	0.2753
	BoolQ-FR-FX	BoolQ (fr)	FRIENDS-QIA + XOpus-QIA (en)	fr	0.3721	0.2049
		BoolQ (fr)	FRIENDS-QIA + XOpus-QIA (en)	es	0.3493	0.1818
	BoolQ-ES-FX	BoolQ (es)	FRIENDS-QIA + XOpus-QIA (en)	fr	0.4087	0.3419
		BoolQ (es)	FRIENDS-QIA + XOpus-QIA (en)	es	0.3813	0.3320
	BoolQ-PT-FX	BoolQ (pt)	FRIENDS-QIA + XOpus-QIA (en)	fr	0.4018	0.2137
		BoolQ (pt)	FRIENDS-QIA + XOpus-QIA (en)	es	0.3539	0.2083
MNLI /XNLI	MNLI-FX	MNLI (downsampled) (en)	FRIENDS-QIA + XOpus-QIA (en)	fr	0.5000	0.4006
		MNLI (downsampled) (en)	FRIENDS-QIA + XOpus-QIA (en)	es	0.4292	0.3644
	XNLI-FR-FX	MNLI (downsampled) (en) + XNLI (fr)	FRIENDS-QIA + XOpus-QIA (en)	fr	0.4406	0.3597
		MNLI (downsampled) (en) + XNLI (fr)	FRIENDS-QIA + XOpus-QIA (en)	es	0.4201	0.3455
	XNLI-ES-FX	MNLI (downsampled) (en) + XNLI (es)	FRIENDS-QIA + XOpus-QIA (en)	fr	0.4521	0.3843
		MNLI (downsampled) (en) + XNLI (es)	FRIENDS-QIA + XOpus-QIA (en)	es	0.4292	0.3692
PAWS-X	PAWS-X-FX	PAWS-X (downsampled) (en)	FRIENDS-QIA + XOpus-QIA (en)	fr	0.4201	0.2557
		PAWS-X (downsampled) (en)	FRIENDS-QIA + XOpus-QIA (en)	es	0.3950	0.2274
	PAWS-X-FR-FX	PAWS-X (downsampled) (fr)	FRIENDS-QIA + XOpus-QIA (en)	fr	0.3973	0.2525
		PAWS-X (downsampled) (fr)	FRIENDS-QIA + XOpus-QIA (en)	es	0.3402	0.2143
	PAWS-X-ES-FX	PAWS-X (downsampled) (es)	FRIENDS-QIA + XOpus-QIA (en)	fr	0.4292	0.3527
		PAWS-X (downsampled) (es)	FRIENDS-QIA + XOpus-QIA (en)	es	0.4384	0.3650

Table 5.4: Performance scores of the baseline and intermediate task training models evaluated on French and Spanish data. Green cells indicate improvement over the respective baseline and the values in red drop below the baseline scores. The accuracy score with no color is identical to the baseline score.

What is striking is the fact that the XNLI-FR-FX model, the one that trained on the English MNLI train set and the development set of the French XNLI, performed worse than the MNLI-FX model for the French test set. With an accuracy score of 0.4406 it outperformed for example a BoolQ-based or PAWS-X-based model though, however, even the Spanish XNLI-ES-FX performed better for the French data prediction with an accuracy of 0.4521, which makes it the only model – along with the MNLI-FX model – that surpasses the baseline accuracy scores.

When evaluated on Spanish, the model results for accuracy all stay below the baseline accuracy score of 0.4406. However, the accuracy score of the PAWS-X-ES-FX model, that used the Spanish translation of the PAWS-X dataset for development, is quite close to the baseline result at 0.4384 and performs best among the models that predicted on the Spanish QIA pairs.

The results are unambiguous for the macro F1-scores, though. Several F1-scores of the intermediate task training models surpass the baseline scores, including all three models that trained on the downsampled inference dataset MNLI – for both the French and Spanish test data. The overall best macro F1-score for the French test data is reached with the MNLI-FX model – hence the same setup as for the accuracy score. For Spanish, it is also a MNLI-based model, with the XNLI-ES-FX model having the highest macro F1-score of 0.3693. The macro F1-score with the PAWS-X-ES-FX model is at 0.3650 though, so there are hardly any differences and thus this score shows the consistency of the PAWS-X-ES-FX model for the Spanish QIA pairs, considering the accuracy scores as well.

What is noteworthy, is that for the English QIA pairs for testing, the models based on the yes/no-question answering dataset BoolQ work best. For French and Spanish, it is even the complete opposite – BoolQ tested on French and Spanish QIA pairs delivers the worst results among all evaluated models and only the F1-score of the BoolQ-ES-FX model outperforms the baseline, if evaluated with the French QIA pairs.

Most surprisingly, the French BoolQ-FR-FX model is even the least effective model setup for French – the predicted output data reached a low point at 0.2049 for the macro F1-score. However, as for the BoolQ-FR-FX model and the BoolQ-ES-FX model, they both trained on a reduced dataset, due to the erroneous development data.

Results per Genre When performance results are broken down by genre, the dominant models for French and Spanish that exceed the respective baseline scores mostly remain the same (see table 5.5).

Model	Model Name	Test Data	Genre	Accuracy	Macro F1	Genre	Accuracy	Macro F1
Baseline								
FRIENDS-QIA + XOpus-QIA	FRIENDS-QIA + XOpus-QIA	fr	Comedy	0.3872	0.2895	Crime	0.5025	0.3898
		es	Comedy	0.3617	0.2771	Crime	0.5320	0.4113
Intermediate Task Training								
BoolQ	BoolQ-FX	fr	Comedy	0.3702	0.2900	Crime	0.4138	0.3759
		es	Comedy	0.3362	0.2643	Crime	0.4335	0.2883
	BoolQ-FR-FX	fr	Comedy	0.3574	0.1944	Crime	0.3892	0.2105
		es	Comedy	0.3106	0.1620	Crime	0.3941	0.2034
	BoolQ-ES-FX	fr	Comedy	0.3745	0.2573	Crime	0.4483	0.4716
		es	Comedy	0.3319	0.2892	Crime	0.4384	0.3598
	BoolQ-PT-FX	fr	Comedy	0.3532	0.1908	Crime	0.4581	0.2386
		es	Comedy	0.2894	0.1684	Crime	0.4286	0.2586
MNLI/ XNLI	MNLI-FX	fr	Comedy	0.4596	0.3312	Crime	0.5468	0.5039
		es	Comedy	0.3617	0.3345	Crime	0.5074	0.3652
	XNLI-FR-FX	fr	Comedy	0.4085	0.2921	Crime	0.4778	0.5224
		es	Comedy	0.3617	0.3116	Crime	0.4877	0.3603
	XNLI-ES-FX	fr	Comedy	0.3830	0.2822	Crime	0.5320	0.5812
		es	Comedy	0.3617	0.3299	Crime	0.5074	0.4013
PAWS-X	PAWS-X-FX	fr	Comedy	0.3872	0.2252	Crime	0.4581	0.2834
		es	Comedy	0.3447	0.2027	Crime	0.4532	0.2569
	PAWS-X-FR-FX	fr	Comedy	0.3489	0.2005	Crime	0.4532	0.3133
		es	Comedy	0.2809	0.1712	Crime	0.4089	0.2625
	PAWS-X-ES-FX	fr	Comedy	0.3702	0.2676	Crime	0.4975	0.4871
		es	Comedy	0.3915	0.3413	Crime	0.4926	0.3711

Table 5.5: Accuracy and F1-scores per genre of the intermediate task training models. Crime is used as short form for crime, drama, mystery. Scores highlighted in green exceed the baseline scores.

For the French QIA pairs for the comedy, and crime, drama, mystery genre, the MNLI-based models perform best and better than baseline scores. For example, for the genre comedy, the baseline accuracy score for French is 0.3872, for the MNLI-FX model it is 0.4596. For Spanish, for comedy, the PAWS-X-ES-FX model outperforms the FRIENDS-QIA + XOpus-QIA, for crime, drama, mystery, the baseline beats the intermediate task training models, even though among them, the XNLI-ES-FX and the MNLI-FX model achieve the best scores for Spanish with an accuracy of 0.5074 each.

An interesting fact that should be mentioned here is that the two genres perform quite differently. As for the crime, drama, mystery genre both accuracy and macro F1-scores exceed the scores of the comedy genre, for all models. A presumable reason for this could be the genre of comedy itself, since humor has many nuances, is situation-dependent and may sometimes include irony, sarcasm, or the like, which even in human interaction might lead to misunderstandings. However, this explanation is not data-based, it is just an attempt to classify the obvious differences regarding the accuracy and F1-scores. And as seen in the label distribution in section 3.5, the comedy genre produces more ambiguous output, where context is needed. Hence if the models are evaluated on the crime, drama, mystery genre, they seem to predict more responses that can be answered unequivocally.

Even though there does not exist an explanation yet, another observation must be pointed out. When evaluated on Spanish, the comedy genre obtains better results compared to the baseline, but for French, the crime, drama, mystery genre performs better in the baseline comparison (see again table 5.5). Although both languages are Romance languages and hence related, models seem to struggle more with humor in French. This could result from the subtitle translation, as humor is very culture-specific and generally difficult to translate. However, this is just a presumption that needs a more detailed study of the translated utterances for ratification.

F1-scores per Label We have seen that regarding overall performances, the accuracy and F1-score do not particularly stand out and the FRIENDS-QIA + XOpus-QIA baseline scores are hard to beat. This discrepancy tells us to look at single labels, as for the overall performance, the intermediate task training step does not consistently help.

In table 5.6, the F1-scores per label can be observed and there are once again no surprises regarding the best dominant intermediate task training models per language. For the English test data, almost all six labels score best with a BoolQ model, even though not necessarily the English one. As for the "yes" class, the highest overall score is reached with the XNLI-ES-FX model at 0.7446, followed by the BoolQ-FX model at 0.7428 – so there is hardly any difference between the two. However, both are still outperformed by the FRIENDS-QIA + XOpus-QIA baseline.

Model	**Model Name**	**Test Data**	**F1/ Label**	**F1/ Label**	**F1/ Label**	**F1/ Label**	**F1/ Label**	**F1/ Label**
Baseline			1/ Yes	2/ No	3/ Yes (condition)	4/ Neither	5/ Other	6/ Context
FRIENDS-QIA + XOpus-QIA	FRIENDS-QIA + XOpus-QIA	en	0.7515	0.6220	0.2000	0.4370	0.2500	0.0000
		fr	0.6100	0.4459	0.0000	0.2825	0.3239	0.3561
		es	0.6554	0.4260	0.0000	0.2162	0.3546	0.3846
Intermediate Task Training								
BoolQ	BoolQ-FX	en	0.7428	0.5802	0.3077	0.5028	0.2253	0.0000
		fr	0.5449	0.4516	0.1667	0.2559	0.2560	0.2778
		es	0.6084	0.4268	0.0000	0.2464	0.1951	0.1750
	BoolQ-FR-FX	en	0.7154	0.5546	0.0000	0.3518	0.1081	0.0000
		fr	0.5283	0.3109	0.0000	0.1404	0.2500	0.0000
		es	0.5261	0.2843	0.0000	0.0308	0.2500	0.0000
	BoolQ-ES-FX	en	0.7321	0.5350	0.3333	0.4720	0.3093	0.1818
		fr	0.5345	0.4615	0.2000	0.2394	0.3014	0.3146
		es	0.4915	0.4615	0.2000	0.1628	0.4114	0.2651
	BoolQ-PT-FX	en	0.7257	0.5559	0.2400	0.4153	0.2353	0.0000
		fr	0.5738	0.3382	0.0000	0.1867	0.1500	0.0333
		es	0.5150	0.3474	0.0000	0.1798	0.1440	0.0635
MNLI/ XNLI	MNLI-FX	en	0.7229	0.5277	0.3200	0.4728	0.1957	0.0000
		fr	0.6647	0.4493	0.1429	0.3699	0.3871	0.3896
		es	0.5904	0.4586	0.3333	0.2282	0.3439	0.2319
	XNLI-FR-FX	en	0.7055	0.5572	0.2609	0.4844	0.3011	0.0000
		fr	0.5869	0.4172	0.2222	0.2479	0.3506	0.3333
		es	0.5472	0.4746	0.2000	0.2302	0.3613	0.2597
	XNLI-ES-FX	en	0.7446	0.5798	0.3077	0.4767	0.1975	0.0000
		fr	0.6098	0.4493	0.2222	0.2857	0.3490	0.3896
		es	0.5704	0.4483	0.1667	0.2584	0.3867	0.3846
PAWS-X	PAWS-X-FX	en	0.6920	0.5697	0.0000	0.3636	0.2524	0.0000
		fr	0.5847	0.3567	0.0000	0.209	0.3182	0.0656
		es	0.5685	0.2945	0.0000	0.1205	0.3810	0.0000
	PAWS-X-FR-FX	en	0.7159	0.5740	0.3158	0.4402	0.1791	0.0000
		fr	0.5941	0.4279	0.0000	0.16	0.2080	0.125
		es	0.5531	0.3653	0.0000	0.1549	0.1481	0.0645
	PAWS-X-ES-FX	en	0.7069	0.5556	0.1739	0.4680	0.2500	0.0000
		fr	0.6276	0.4684	0.2000	0.2871	0.2979	0.2353
		es	0.6471	0.4459	0.2222	0.2538	0.3239	0.2973

Table 5.6: F1-scores per label for the intermediate task training models. Green cells indicate higher performance than the baseline.

For Spanish and French, it is once again either an MNLI-based model or a PAWS-X-based model with the best performance and a BoolQ-based model with the worst performance.

In terms of "yes", it is overall the class with the best performance scores, the best one peaks at 0.7515 with the FRIENDS-QIA + XOpus-QIA baseline. The lowest score of 0.4915 for this label is tested with the Spanish QIA pairs and the BoolQ-ES-FX model. However, although this is the lowest value for label 1, it is still better than those of labels 3, 5, or 6. Label 5, "other", is the second most frequent label in the XOpus-QIA dataset, and it seems difficult to predict, not even the baseline retrieves a good score for this label (0.2500). The best F1-score for the "other" class evaluated on English is at 0.3039, the one for French at 0.3871 and the one for Spanish is the highest at 0.4114 with a BoolQ-based model, surprisingly. The "no" class, which is a less frequent label than "other", yet still produces better results with F1-scores of 0.5802 for English, 0.4746 for Spanish, and 0.4648 for French. However, the English results for label 2 are below the baseline results, whereas the cross-lingual F1-score results for French and Spanish are both above the baseline.

When we look at the less frequent labels according to the label distribution (see again table 3.3), the results are more surprising regarding the type of class that performs better with the intermediate task training models than the baseline models.

As for "yes, subject to some conditions" and "lacking context", the scores are in general at a low point of even zero – for a lot of models. For English, the only model that predicted "lacking context" at all, was the Spanish BoolQ-ES-FX model. And the overall scores for label 6 are in general somewhat poor, since the class is not at all an uncommon label with a frequency of 13.47 percent for French, and 13.70 for Spanish, and 12.52 percent for the English QIA pairs. Maybe it is important here to also consider that the model only sees this label from the XOpus-QIA part of the training/ fine-tuning data, not from the FRIENDS-QIA part (since the label 6 is excluded from the FRIENDS-QIA dataset, as mentioned earlier).

Whereas for label 3, "yes, subject to some conditions", which accounts for only 1.60 percent of the data sets for French and Spanish, the zero-performance in a lot of models is not that unexpected for this class. However, all MNLI-based models, evaluated on all three languages, outperform the baseline for this label. Label 3 reaches a solid 0.3333 macro F1-score with the English MNLI-FX model. While with the baseline model, for the Spanish QIA pairs, the macro F1-score is zero.

The MNLI-based intermediate task training models also reach similar results for "neither yes nor no". All scores are higher than the FRIENDS-QIA + XOpus-QIA baseline scores, except for the XNLI-FR-FX model evaluated on French, astonishingly. The Spanish PAWS-X-ES-FX model also mostly outperforms the baseline for those two labels. For label 5, "other", which is more frequent than label 3 or 4, the MNLI-based models still stand out and especially the XNLI-FR-FX model achieves the best score for all three languages.

Comparison to the XOpus-QIA and FRIENDS-QIA-E6 baselines We shortly want to include the other two baseline models, the XOpus-QIA and the FRIENDS-QIA-E6, in the discussion.

As described above, if we only compare baseline models, the XOpus-QIA model performs best when tested with French and Spanish QIA pairs. For the English QIA pairs, the FRIENDS-QIA + XOpus QIA model performs best.

If sequential intermediate task training is added to the models, the BoolQ-based models achieve the best scores with English, the MNLI-based models perform best with French, and the PAWS-X-based models best with Spanish.

Comparing now both baseline and intermediate task training models, the baseline models stand out. When tested on English, French, and Spanish QIA pairs from XOpus-QIA, the intermediate task training models are all outperformed regarding accuracy. As for the macro F1-scores in the cross-lingual setting, the intermediate task training models mostly are outperformed by the XOpus-QIA model, which only trains on the XOpus-QIA corpus. As for the FRIENDS-QIA-E6 model, the intermediate task training model stand out regarding the accuracy scores, and even for the macro F1-scores the FRIENDS-QIA-E6 model did not produce convincing output (see table 4.2).

And also, when we both compare the F1-scores per label for the baseline and intermediate task training models, the XOpus-QIA baseline model clearly wins in the cross-lingual context. For French, all labels receive the best result with this baseline model, for the Spanish test data, for five labels it is also the XOpus-QIA baseline model that performs best, except for label 3, the "yes, subject to some conditions" class. For the English test data, there is a draw; for three labels each it is either the baseline or the intermediate task training model that performs best. However, as we have seen and discussed, for the in-language setup with English, the FRIENDS-QIA + XOpus-QIA model is the best baseline model.

5.3 Final Considerations

As an overall result, the task of understanding indirect answers to yes/no-questions still seems quite challenging. For the "yes" class, the results are quite presentable. However, once the answer is not a clear *yes* or *no*, or we look at the overall results, the task remains demanding. Nevertheless, we have seen that for the less frequent labels the intermediate task training improved the results partially clearly and the inference datasets MNLI and XNLI as well as the paraphrase dataset PAWS-X helped within the scope of cross-lingual transfer learning.

One explanation for low overall performance results, as given by Damgaard et al. (2021), might be the dataset structure. Both the XOpus-QIA as well as the FRIENDS-QIA have no limitations regarding the number of answer sentences. Damgaard et al. (2021) argue that in this case, "[. . .] the speaker might change their mind in-between the sentences, further complicating the task of interpreting the questions and answers" (Damgaard et al., 2021, p. 8), which influences performance scores.

There might also be several more factors. Already in the process of data collection it was observed that using subtitles with no cues of who is speaking is difficult to understand even for humans. Furthermore, the label distribution shows that labels 5 and 6 are quite common, and both do not stand for an answer themselves, but annotate the fact that no answer can be given.

And as an explanation why the XOpus-QIA baseline model outperforms the intermediate task training models in the cross-lingual setup, the XOpus-QIA with the collected QIA pairs matches the evaluation data best, since they are both extracted and translated from the same series. Additionally, it is quite a small dataset compared to the FRIENDS-QIA by Damgaard et al. (2021) or the Circa dataset by Louis et al. (2020). However, it could also be an explanation that using BERT alone is already quite powerful, as seen in chapter 2. To conclude, experimenting with transfer learning for this cross-lingual task might not have led overall to the desired results, since the zero-shot approaches outperformed the intermediate task training models mostly. Nevertheless, it is a finding that can be followed up to further refine and optimize the models.

6 Conclusion and Future Work

This thesis presented a cross-lingual transfer learning approach for the task of understanding indirect answers not only in English, but also in French and Spanish.

As for the related work, the most recent contributions within the scope of this work are by Louis et al. (2020), and Damgaard et al. (2021), since this thesis adapts the label set by Louis et al. (2020) for annotation and the FRIENDS-QIA dataset by Damgaard et al. (2021) for the zero-shot baseline models and transfer learning intermediate task training models.

To address the gap in resources beyond English, we introduced XOpus-QIA, a novel dataset. It consists of pairs of yes/no-questions and indirect answers, extracted from the OpenSubtitles corpus, provided by Lison and Tiedemann (2016). A total of 615 QIA pairs for English, and 438 QIA pairs for French and Spanish each, all translations of the English QIA pairs, are available for training and/or testing. During the data collection process, several challenges were encountered. A first challenge is that the subtitle data lacks speaker information, it is not always clear who is speaking.

The second challenge relates to the task of finding triggers of indirect answers in subtitle data. This is distinct from the setup by Louis et al. (2020), who generated question answer pairs. Related to this challenge is the fact that answers may be ambiguous if analyzed at an utterance level. This impacts annotation significantly; hence a new label “lacking context” is introduced in this work that substitutes the “N/A” label for annotator disagreement of the existing relaxed label set by Louis et al. (2020).

The difficulty of the task of understanding indirect answers is also visible in the inter-annotator agreement. For the QIA pairs annotated by two annotators, the Cohen’s Kappa score is at only 0.54. And furthermore, it is also reflected in the model results, especially when the metrics are calculated per label and per genre. Human annotation was surprisingly easier for comedy, and led to higher agreement scores, while models struggled more on that genre and produced more erroneous output than for the crime, drama, mystery genre.

Several classification models were set up using the MaChAmp toolkit by van der Goot et al. (2021). The zero-shot baseline models are based on the multilingual BERT (Devlin et al., 2019) encoders and train either on the XOpus-QIA or FRIENDS-QIA dataset only, or a concatenation of both. The intermediate task training models for cross-lingual transfer learning are also based on multilingual BERT (Devlin et al., 2019), then train sequentially on an intermediate task dataset, then fine-tune on the concatenated FRIENDS-QIA + XOpus-QIA dataset. As for the intermediate tasks, we investigated three sources, namely: the BoolQ yes/no-question-answer set (Clark et al., 2019) in the original English version and different translations, as well as the MNLI inference dataset (Williams et al., 2018) and XNLI, the corresponding translated dataset (Conneau et al., 2018), plus PAWS-X, the cross-lingual paraphrase dataset (Yang et al., 2019). All model performances were tested on the English, French, and Spanish versions of the XOpus-QIA dataset.

The results are somewhat unexpected. For the overall performance, the zero-shot baseline models, especially the one only trained on XOpus-QIA, tend to work best if tested for the French and Spanish XOpus-QIA. Using the MAChAmp toolkit with the pre-trained multilingual BERT model, without any intermediate task training, is already quite efficient and mainly speaks for BERT models. However, for the intermediate task training models for cross-lingual transfer learning, datasets based on MNLI are generally a good option to get at least an acceptable result for overall performance.

When the performance is measured separately per label, for the three less frequent labels the cross-lingual transfer learning models that use intermediate task training perform better than the baseline models. As an overall analysis, the F1-scores per label vary fundamentally. As for the “yes” class, performance results peak for every model, ranging from 0.6554 for the Spanish QIA pairs to 0.7515 for the English QIA pairs, both tested on the FRIENDS-QIA + XOpus-QIA baseline model.

Especially the “lacking context” label, which makes up more than twelve percent of the testing datasets, produces low performance results. Yet again, this seems like a confirmation that either the task, or the dataset, or both are challenging and thus there is great potential for model adaptations and improvements. Since the cross-lingual transfer learning approach has not yielded a significant gain compared to the zero-shot baseline models, for future work there are several options.

First, the multilingual BERT was used as pre-trained model with MaChAmp. Sanagavarapu et al. (2022) reported satisfying results with the adapted BERT model RoBERTa (Liu et al., 2019b). One of the options is hence to train all models on RoBERTa. A second option is multitask learning (Phang et al., 2019). The transfer learning models in this thesis were trained sequentially; thus, the joint learning option is still to be tested. And even though the MNLI or XNLI datasets for inferences do not perform too badly for French and Spanish, there are way more datasets, even cross-lingual datasets, to choose from and to train on for this task. The BoolQ dataset, that seems to be a good fit for this task since it consists of yes/no-questions and answers, only performed well for the English QIA pairs. Hence it can be interesting to test a dataset for intermediate task training that at first does not seem appropriate for this task because it is lacking similarity with the target task.

Limitations And ultimately, as a limiting factor, the XOpus-QIA dataset is a rather small dataset to draw definitive conclusions on this task, since only 615 QIA pairs for English are available for training, development, and testing. Additionally, the disadvantage of subtitles is that dialogues are optimized to fit to images and no specific details are given. Perhaps the key to better performance results for further work could be to involve the context of question-indirect-answer pairs, as done by Sanagavarapu et al. (2022). This would lead to a cross-lingual approach not only based on more, but on richer data. Or maybe it can be even beneficial for this task to switch to more natural conversations instead of subtitle scripts. Overall, understanding indirect answers to yes/no-questions is a field that offers much scope for future work.

References

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Cody Buntain. 2020. agreement.py. `https://gist.github.com/cbuntain/9dd7e42d5d8ab34609162410e06f3270`. Retrieved 23-05-30.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

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

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Cathrine Damgaard, Paulina Toborek, Trine Eriksen, and Barbara Plank. 2021. "I'll be there for you": The one with understanding indirect answers. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 1–11, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Scott Grimm, and Christopher Potts. 2009. Not a simple yes or no: Uncertainty in indirect answers. In *Proceedings of the SIGDIAL 2009 Conference*, pages 136–143, London, UK. Association for Computational Linguistics.

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

Annette M. Green. 1997. Kappa statistics for multiple raters using categorical classifications. In *Proceedings of the Twenty-Second Annual Conference of SAS Users Group*, San Diego, USA.

Nancy Green and Sandra Carberry. 1999. Interpreting and generating indirect answers. *Computational Linguistics*, 25(3):389–435.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization.

John D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. Technical Report Draft 13, University of Colorado, Institute of Cognitive Science.

Reagan Lee. 2023a. boolq_es. `https://huggingface.co/datasets/reaganjlee/boolq\_es`. Retrieved 23-05-30.

Reagan Lee. 2023b. boolq_fr. `https://huggingface.co/datasets/reaganjlee/boolq\_fr`. Retrieved 23-05-30.

Reagan Lee. 2023c. boolq_pt. `https://huggingface.co/datasets/reaganjlee/boolq\_pt`. Retrieved 23-05-30.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2019a. XQA: A cross-lingual open-domain question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach.

Annie Louis, Dan Roth, and Filip Radlinski. 2020. "I'd rather just go to bed": Understanding indirect answers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7411–7425, Online. Association for Computational Linguistics.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng.*, 22(10):1345–1359.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, 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)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 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)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. 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*, pages 557–575, Suzhou, China.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2019. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks.

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

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Geoffrey Raymond. 2003. Grammar and social organization: yes/no interrogatives and the structure of responding. *American Sociological Review*, 68(6):939–967.

Krishna Sanagavarapu, Jathin Singaraju, Anusha Kakileti, Anirudh Kaza, Aaron Mathews, Helen Li, Nathan Brito, and Eduardo Blanco. 2022. Disentangling indirect answers to yes-no questions in real conversations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4677–4695, Seattle, United States. Association for Computational Linguistics.

Anna-Brita Stenström. 1984. *Questions and responses in English conversation.* CWK Gleerup Malmo, Sweden.

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

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

List of Figures

List of Tables

Appendices

Example of BoolQ

Example of a question-answer-pair of the BoolQ dataset (Clark et al., 2019).

```
{"question": "is windows movie maker part of windows essentials",
"title": "Windows Movie Maker",
"answer": true,
"passage": "Windows Movie Maker (formerly known as Windows Live
Movie Maker in Windows 7) is a discontinued video editing software
by Microsoft. It is a part of Windows Essentials software suite
and offers the ability to create and edit videos as well as to
publish them on OneDrive, Facebook, Vimeo, YouTube, and Flickr."}
```

Example of MNLI

Example sentence of the MNLI dataset (Williams et al., 2018).

```
neutral ( ( Conceptually ( cream skimming ) ) ( ( has ( ( ( two
( basic dimensions ) ) - ) ( ( product and ) geography ) ) ) . ) )
( ( ( Product and ) geography ) ( ( are ( what ( make ( cream
( skimming work ) ) ) ) ) ). ) )
(ROOT (S (NP (JJ Conceptually) (NN cream) (NN skimming))
(VP (VBZ has) (NP (NP (CD two) (JJ basic) (NNS dimensions))
(: -) (NP (NN product) (CC and) (NN geography)))) (. .)))
(ROOT (S (NP (NN Product) (CC and) (NN geography)) (VP (VBP are)
(SBAR (WHNP (WP what)) (S (VP (VBP make) (NP (NP (NN cream))
(VP (VBG skimming) (NP (NN work)))))))) (. .)))
Conceptually cream skimming has two basic dimensions - product and
geography. Product and geography are what make cream skimming work.
31193 31193n government neutral
```

Example of PAWS-X

Example of a sentence-paraphrase pair of the PAWS-X dataset (Yang et al., 2019).

```
Sentence 1:
In Paris , in October 1560 , he secretly met the English ambassador ,
Nicolas Throckmorton , asking him for a passport to return to England
through Scotland .
Sentence 2:
In October 1560 , he secretly met with the English ambassador , Nicolas
Throckmorton , in Paris , and asked him for a passport to return to
Scotland through England .
Label:
0
```

Example of an OpenSubtitles file

File from 2015, folder ID 3591512, document ID 6461786, sentence ID 1-21 (Lison and Tiedemann, 2016).

```
 <s id="1">
    <time id="T1S" value="00:01:22,320" />
WHISPERS ECHO:
  </s>
  <s id="2">
I love you, I love you...
    <time id="T1E" value="00:01:25,080" />
  </s>
  <s id="3">
    <time id="T2S" value="00:01:47,520" />
EXPLOSIONS ECHO Cyril!
  </s>
  <s id="4">
Cyril!
    <time id="T2E" value="00:01:50,840" />
  </s>
  <s id="5">
    <time id="T3S" value="00:02:06,960" />
RAINFALL
    <time id="T3E" value="00:02:09,440" />
  </s>
  <s id="6">
    <time id="T4S" value="00:02:14,960" />
The, er...position is for a secretary.
    <time id="T4E" value="00:02:18,440" />
  </s>
  <s id="7">
    <time id="T5S" value="00:02:18,440" />
Then the agency shouldn’t have given you my name.
    <time id="T5E" value="00:02:20,560" />
  </s>
  <s id="8">
    <time id="T6S" value="00:02:20,560" />
My typing and shorthand isn’t good enough.
    <time id="T6E" value="00:02:22,440" />
  </s>
  <s id="9">
    <time id="T7S" value="00:02:22,440" />
- Assistant, then.
  </s>
  <s id="10">
- TYPEWRITER CLACKS
    <time id="T7E" value="00:02:24,360" />
  </s>
  <s id="11">
    <time id="T8S" value="00:02:24,360" />
With some minor secretarial duties.
    <time id="T8E" value="00:02:26,520" />
  </s>
  <s id="12">
```

```
    <time id="T9S" value="00:02:26,520" />
I sent her all the details.
  </s>
  <s id="13">
She chose you.
    <time id="T9E" value="00:02:30,120" />
  </s>
  <s id="14">
    <time id="T10S" value="00:02:30,120" />
Really?
    <time id="T10E" value="00:02:31,600" />
  </s>
  <s id="15">
    <time id="T11S" value="00:02:31,600" />
You’re a teacher?
    <time id="T11E" value="00:02:33,080" />
  </s>
  <s id="16">
    <time id="T12S" value="00:02:33,080" />
Games mistress, yes.
    <time id="T12E" value="00:02:34,680" />
  </s>
  <s id="17">
    <time id="T13S" value="00:02:34,680" />
Teachers are good at organising.
    <time id="T13E" value="00:02:37,120" />
  </s>
  <s id="18">
    <time id="T14S" value="00:02:37,120" />
Mrs Owen is expecting a lot of guests.
    <time id="T14E" value="00:02:39,320" />
  </s>
  <s id="19">
    <time id="T15S" value="00:02:39,320" />
Whereabouts in the country?
    <time id="T15E" value="00:02:41,040" />
  </s>
  <s id="20">
    <time id="T16S" value="00:02:41,040" />
The Devon coast.
  </s>
  <s id="21">
Soldier Island.
    <time id="T16E" value="00:02:43,440" />
  </s>
```

Example of a metadata block at the end of a subtitle file

File from 2015, folder ID 3591512, document ID 6461786 (Lison and Tiedemann, 2016).

```
<meta>
  <conversion>
    <encoding>utf-8</encoding>
    <sentences>725</sentences>
    <corrected_words>0</corrected_words>
    <ignored_blocks>0</ignored_blocks>
    <truecased_words>118</truecased_words>
    <unknown_words>14</unknown_words>
    <tokens>5204</tokens>
  </conversion>
  <source>
    <year>2015</year>
    <HD>0</HD>
    <genre>Crime,Drama,Mystery</genre>
    <cds>1/1</cds>
    <duration>56</duration>
  </source>
  <subtitle>
    <rating>1.0</rating>
    <version>1</version>
    <date>2016-01-14</date>
    <confidence>1.0</confidence>
    <blocks>648</blocks>
    <duration>00:55:29,120</duration>
    <language>English</language>
    <machine_translated>0</machine_translated>
  </subtitle>
</meta>
```

Example for params.json

Example of a params.json file as used in this thesis for model setup (van der Goot et al., 2021).

```
{
  "transformer_model": "bert-base-multilingual-cased",
  //"transformer_model": "xlm-roberta-large",
  "random_seed": 8446,
  "default_dec_dataset_embeds_dim": 12,
  "encoder": {
    "dropout": 0.2,
    "max_input_length": 128,
    "update_weights_encoder": true
  },
  "decoders": {
    "default_decoder": {
      "loss_weight": 1.0,
      "metric": "accuracy",
      "additional_metrics": ["accuracy", "f1_micro", "f1_macro"],
      "topn": 1,
      "layers_to_use": [-1]
    },
    "classification": {
    },
    "dependency": {
      "arc_representation_dim": 768,
      "tag_representation_dim": 256,
      "metric": "las"
    },
    "mlm": {
      "metric": "perplexity"
    },
    "multiclas": {
      "metric": "multi_acc",
      "threshold": 0.7
    },
    "multiseq": {
      "metric": "multi_acc",
      "threshold": 0.7
    },
    "regression": {
      "metric": "avg_dist"
    },
    "seq": {
    },
    "seq_bio": {
      "metric": "span_f1"
    },
    "string2string": {
    },
    "tok": {
      "pre_split": true
    }
  },
  "batching": {
    "max_tokens": 1024,
    "batch_size": 32,
    "sort_by_size": true,
    "sampling_smoothing": 1.0 // 1.0 == original size, 0.0==all equal
  },
  "training": {
    "keep_top_n": 1,
    "learning_rate_scheduler": {
      //"type": "slanted_triangular",
      "cut_frac": 0.3,
      "decay_factor": 0.38,
      "discriminative_fine_tuning": true,
      "gradual_unfreezing": true
    },
    "num_epochs": 20,
    "optimizer": {
      //"type": "adamw",
      "betas": [
        0.9,
        0.99
      ],
      "lr": 0.0001,
      "correct_bias": false,
      //"patience": 5, // disabled, because slanted_triangular \
      changes the lr dynamically
      "weight_decay": 0.01
    }
  }
}
```

Example of a configuration file in json-format (van der Goot et al., 2021)

Classification task with the inference dataset MNLI (Williams et al., 2018).

```
{
    "RTE": {
        "train_data_path": "machamp/data/qia/intermediate_mnli/mnli_train.tsv",
        "dev_data_path": "machamp/data/qia/intermediate_mnli/mnli_dev.tsv",
        "sent_idxs": [1,2],
        "tasks": {
            "rte": {
                "task_type": "classification",
                "column_idx": 0,
                "metric": "accuracy",
                "additional_metrics": ["accuracy", "f1_micro", "f1_macro"]
            }
        }
    }
}
```

Inhalt des verfügbaren Cloud-Ordners

Please note: The following link leads to the file storage service Google Drive. The link provided will be accessed in editor mode, hence all files and folders should be downloadable.

`https://drive.google.com/drive/folders/1H7rko2zwbIeOFJjzSATa_4aUDX3aQgBp?usp=sharing`

The link contains a folder with the following content:

- PDF version of this thesis
- OpenSubtitles files and alignment files
- XOpus-QIA files
- Folder that contains all models trained with MaChAmp and the data used for training and evaluation
- Excel-file with all model results