IndirectQA: Understanding Indirect Answers to Implicit Polar Questions in French and Spanish

Christin Müller **Barbara Plank**
MaiNLP lab, Center for Information and Language Processing, LMU Munich, Germany
Munich Center for Machine Learning (MCML), Munich, Germany
Department of Computer Science, IT University of Copenhagen, Denmark
christin.muller@googlemail.com, b.plank@lmu.de

Abstract

Polar questions are common in dialogue and expect exactly one of two answers (yes/no). It is however not uncommon for speakers to bypass these expected choices and answer, for example, *"Islands are generally by the sea"* to the question: *"An island? By the sea?"*. While such answers are natural in spoken dialogues, conversational systems still struggle to interpret them. Seminal work to interpret indirect answers were made in recent years—but only for English and with strict question formulations. In this work, we present a new corpus for French and Spanish—IndirectQA —where we mine subtitle data for indirect answers to study the labeling task with six different labels, while broadening polar questions to include also implicit polar questions (statements that trigger a yes/no-answer which are not necessarily formulated as a question). We opted for subtitles since they are a readily available source of conversation in various languages, but also come with peculiarities and challenges which we will discuss. Overall, we provide the first results on French and Spanish. They show that the task is challenging: the baseline accuracy scores drop from 61.43 on English to 44.06 for French and Spanish.

Keywords: Indirect Answers, Natural Language Understanding, Corpora and Annotation

1. Introduction

Even when asked a simple polar question, humans often do not directly answer with a clear *yes* or *no*. This is characteristic of human conversations, however, at the same time it is a major challenge for Natural Language Understanding (NLU).

Linguists have analyzed why humans do not provide direct answers. Stenström (1984) found that humans have a tendency to provide further information, and that adding superfluous information has important social reasons. Overall, indirect answers make conversations more natural (Stenström, 1984).

The following example extracted from our corpus illustrates an indirect answer that does imply a *yes*, but does not literally say *yes*.

Question: "An island? By the sea?"
Answer: "Islands are generally by the sea."

Recently, work on understanding indirect answers is increasing. Seminal work by Louis, Roth, and Radlinski (2020) collected Circa, a large corpus of question and answer pairs by crowdsourcing. Following up on their work, Damgaard, Toborek, Eriksen, and Plank (2021) opted to choose more natural data extracted from TV scripts (Damgaard et al., 2021), in comparison to the prompts and scenarios generated in Louis et al. (2020). A recent approach focuses on the importance of context and analyzed whether labels change when more context is provided (Sanagavarapu et al., 2022).[1]

All three of the above contributions focus on English (Louis et al., 2020; Damgaard et al., 2021; Sanagavarapu et al., 2022). We propose a new dataset that includes French and Spanish for classification of indirect answers into six classes, following prior proposed label schemes. As our data source we propose to mine the OpenSubtitles Corpus (OPUS) (Lison and Tiedemann, 2016), as it consists of aligned data in different languages. As part of our work, we outline our considerations on how to collect questions and answers from a natural dataset as OPUS, to incentivize future work on extending IndirectQA's language coverage.

Contributions In this paper we will present the new, cross-lingual evaluation dataset over two genres: IndirectQA[2] – with parallel data in English, French and Spanish, collected from the OpenSubtitles Corpus (OPUS) (Lison and Tiedemann, 2016). It contains a total of 1,053 question-indirect answer pairs (QIA pairs), which cover 615 English and 438 parallel, annotated QIA pairs in French and Spanish for six different labels. IndirectQA spans two genres, namely comedy and crime, drama, mystery. With this new corpus, we want

[1] We were unable to get their data which unfortunately is not yet publicly available at: https://github.com/krishna-chaitanya-sanagavarapu/SwDA-IA

[2] Data is available at: https://github.com/mainlp/indirectQA

LREC-COLING 2024, pages 9025–9035
20-25 May, 2024.

to fill the gap that for French and Spanish – even though they are actual high-resource languages – no such dataset is available. We provide results for zero-shot transfer. Additionally, we investigate intermediate task training to see if English, French or Spanish NLI intermediate task data helps to improve zero-shot results.

2. Related Work

Seminal work on modeling the meaning of indirect answers to polar questions has been introduced by Louis et al. (2020), discussed next.

Circa Louis et al. (2020) created *Circa*, a corpus consisting of 32,268 pairs of polar questions and indirect answers to these questions. Their dataset was crowd-sourced both for collecting questions and answers, where questions were collected for ten prompts. To obtain indirect answers that seem as natural as possible, crowd workers were instructed to imagine realistic conversation scenarios (Louis et al., 2020).

What is special about their work is that they not only include binary labels (*yes* and *no*), but they introduced multiple labels, as they found that a variety of examples in their corpus does not fit into just a binary annotation scheme (Louis et al., 2020). This resulted in their "strict" and "relaxed" label set, given in Table 1. The *strict* label set includes, next to the yes/no labels, also labels with a certain uncertainty, such as "probably yes" and "probably no". 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. We use an adapted version of the latter to annotate our own corpus.

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 1: The strict and relaxed label set, as introduced and used by Louis et al. (2020).

FRIENDS-QIA Follow-up work by Damgaard et al. (2021) proposed *FRIENDS-QIA*, a corpus mined from transcripts of the famous Friends TV series. Damgaard et al. (2021) collected 5,390 pairs of polar questions and their respective indirect answer. 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 they used the relaxed label set (Louis et al., 2020).

Damgaard et al. (2021) noted that models trained on the FRIENDS-QIA corpus differ from models trained on the Circa corpus by Louis et al. (2020), with higher performance on Circa. They analyze three reasons 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). Since the Circa corpus is much larger than the FRIENDS-QIA, Damgaard et al. (2021) argue that their CNNs have more data to learn well from (Damgaard et al., 2021). All three reasons mentioned also apply to our own contribution, so we will return to them in Section 3.

SwDA-IA A third and most recent corpus on polar questions with indirect answers is by Sanagavarapu et al. (2022). They introduced the *SwDA-IA* dataset with 2,544 instances. The data comes from the Switchboard corpus (Jurafsky, Shriberg, and Biasca, 1997).

Their approach focuses also on gathering more natural data, similar to Damgaard et al. (2021), and additionally on the importance of context. Hence they annotate their data in two ways: in one annotators only annotated the question and the immediate answer that follows (Sanagavarapu et al., 2022). In the second, annotators were shown more context around the question-answer-pair (three speaker-answer-turns) (Sanagavarapu et al., 2022). They gained different annotations for the two annotation setups, showing that context is important and influences annotator decisions. Unfortunately, the data is not publicly released yet.

3. IndirectQA

To study indirect answers in a cross-lingual context, we introduce the IndirectQA 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), which is 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. However, one difficulty 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. Both the authors and translators of the subtitles and the quality of the translated data vary within the corpus.

3.1. Data Collection

For the data collection task, we collected subtitles from two genres:

1. comedy
2. crime, drama, mystery[3]

We were interested in comparing genres to cover two different text types and challenges during collection and annotation, and at the same time to have a more varied corpus. Future work could include additional genres.

We choose the two genres because they are quite contrasting; in comedy, jokes and irony sometimes lead to quite situational humor. It is quite important to follow the context of dialogues to understand what was supposed to be funny.

In total, we browsed subtitle files in English for yes/no-questions and their respective indirect answer (QIA pairs), among them half from the genre crime, drama, mystery, and slightly more than half 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, in our final corpus we included data for which all subtitle files are inter-lingually aligned.

On Polar Questions Although prior works on understanding indirect answers mentioned the extraction of *polar questions* (Louis et al., 2020; Damgaard et al., 2021), we do not want to limit ourselves to the extraction of polar questions only (in the strict sense as being formulated as a proper question). Instead, we aim to collect all yes/no-questions and statements we see in the corpus – using a less strict definition, so that the dataset is supposed to cover a variety of indirect answers.

We only extracted QIA pairs from human translated subtitles, where the original language of the subtitle is English, so that French and Spanish translations are compared to an English original version.

As yes/no-questions, we extracted all type of questions that could be answered with a *yes* or a *no*. To illustrate this, here are examples of what is included.

In the first example, we have a clear case of a polar question along with an indirect answer.
Question: "Did Mrs Owen leave any instructions for me? - I'm the secretary." Answer: "Only to ensure that you were comfortable and had everything you wished for, Miss Claythorne."

The second example is more indistinct in a sense that the question can be answered with a *yes* or a *no*, however, the status of being a polar question becomes more ambigious in this case, especially if you look at the French translation – it can be interpreted as a question, but also as a demand. *Question: "Perhaps you could call me John?"*

Answer: "Thank you, John."
Question: "Vous pourriez peut-être m'appeler John."

Answer: "Merci, John."

Along with the QIA pairs, some metadata is extracted, in particular: 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.

Challenges and Peculiarities During the extraction process of the QIA pairs one difficult challenge is encountered: speaker turns are not indicated. 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 in the video during the collection process. Although this task is time-consuming, it seems necessary to ensure high quality of the QIA pairs.

For English, we collected a total of 615 QIA pairs, 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 we 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.

[3]The term *crime, drama, mystery* stands for one type of subtitle genre in the OPUS corpus.

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 comedy (see Table 2).

The translated and 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, the same amount of incomplete QIA pairs was detected and will not be included in the final evaluation data. For French and Spanish, this yields a total of 438 QIA pairs, 203 for the genre crime, drama, mystery, and 235 for the genre comedy (see Table 2).

3.2. Data Annotation

We use a slightly altered version of the *relaxed* label set (see again Table 1) used by prior work (Louis et al., 2020; Damgaard et al., 2021). We changed the "N/A" label used by Louis et al. (2020), since it marks annotator disagreement. In this work, the data is mostly annotated by one person, which makes this label redundant. However, we add a label to indicate indirect answers with "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 in their case they did not extract their question-answer pairs from real, existing conversations but collected solicited data solely for their purpose. As mentioned above, one of the challenges during data collection are the missing turns in the subtitle data. With the "lacking context" label we acknowledge 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 video material at hand. Therefore, the final set of labels used in this work is defined as follows, where 1-5 are from Damgaard et al. (2021):

1. *Yes* – every answer that can be interpreted as a *yes,* even if it is not a clear *yes*, but more a *maybe yes* or a *yes* in a weakened form. Ex.: *Q: Are you Mr Narracott? A: Ain't no-one else holding the sign.*

2. *No* – clearly a *no* or all gradients of *no*. Ex.: *Q: Have they not telephoned? A: There's no telephone on the island, Madam.*

3. *Yes, subject to some conditions* – in this case, the answer means *yes*, but with the restriction that it holds only under certain circumstances. Ex.: *Q: Are you a betting man, Lombard? A: It depends.*

4. *Neither yes nor no* – a label for "in the middle" answers (Louis et al., 2020), when the indirect answer cannot be classified in the binary *yes* or *no* scheme. Ex.: *Q: You ran me off the road and then you have the temerity to tell me it's my fault? A: Careful, old boy. Getting a little red in the face there.*

5. *Other*– this label marks the situation, when the indirect answer does not match the question. Ex.: *Q: 21 men? A: I always thought someone would blab.*

6. *Lacking context* – 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. Ex.: *Q: May I show you to the drawing room, sir? Perhaps an aperitif, whilst you await the other guests? A: Ah... Mr Davis. You look like a man who could use a drink.*

Since the languages in IndirectQA are aligned, the annotation process is only done on the English data and the labels are then transferred to the French and the Spanish QIA pairs. The complete IndirectQA with 615 QIA pairs in English has been annotated by one annotator (annotator 1). These annotations are treated as gold standard within the scope of this work.

Quality estimation procedure To estimate annotation quality, given approximately 200 QIA pairs in English (from both genres), we asked a second annotator to independently label the data. 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 video material that corresponds to the respective subtitle files and QIA pairs, annotator 2 has not. This knowledge advantage means that annotator 1 might better classify indirect responses, since visual cues (gestures and facial expressions) provided important clues.

The annotation was performed in two rounds. In both rounds, annotators were instructed to *both*

Language	Extracted #	Final # QIA pairs	Label 1	Label 2	Label 3	Label 4	Label 5	Label 6
English	615	all (615)	35.61	13.01	1.95	12.36	24.39	12.52
	292	Crime (292)	35.62	13.36	1.37	15.75	21.92	11.99
	323	Comedy (323)	35.60	12.69	2.48	9.29	26.93	13.00
French	239/205	all (438)	34.93	14.16	1.60	11.42	24.20	13.47
Spanish	239/205	all (438)	35.16	14.16	1.60	11.19	24.20	13.70

Table 2: Final label distribution of the IndirectQA dataset (in percentages). Total amount (#) of QIA pairs extracted for the genres comedy and crime, drama, mystery (abbrev. crime), indicated as x/x, respectively. 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. Label numbers correspond to the descriptions provided in Section 3.2.

identify question-indirect answer pairs and label them with one of the six labels. After the first round it emerged that annotators had a different interpretation of what triggers an indirect answer. Round 1 resulted in only 172 annotated QIA pairs out of 207 QIA pairs provided for annotation by one annotator, while the other annotator identified 207. This was because some QIA pairs did not contain polar questions in a strict sense. After a discussion round, we decided to extend the concept to include implicit polar questions, as understanding these felt important and more natural for conversations. Therefore, in a second round, after refining the concept, the data was re-labeled and only two QIA pairs out of 207 remain unlabeled. Therefore, on 205 out of 207 pairs both annotators agree that they constitute valid QIA pairs.

Figure 1 shows the distribution over the labels provided by both annotators. The distribution is skewed, as shown in prior work as well. We note consistent trends in the label distribution. In both cases "yes" (label 1) is the most frequent label, followed by "other" (label 5). The distribution of the most frequent "yes" label is even stable across genres (see Table 2 for details). As for the genre crime, drama, mystery, the distribution of "yes" is at 35.62%, for comedy at 35.60%. 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 most infrequent label is 3 (Yes, subject to some conditions), constituting around or less than 2% of the data.

Inter-annotator agreement is good yet varies per genre For the total of 205 doubly-annotated English QIA pairs, the observed agreement is 64.39. Cohen's Kappa (Cohen, 1960) is 0.54, which constitutes a fair to good level of agreement (Green, 1997). What is surprising is that agreement varies enormously between the genres, and crime turns out more difficult. For the crime, drama, mystery genre, the observed agreement is at 52.68, the Cohen Kappa at 0.38. A score that even after Green (1997)'s definition is

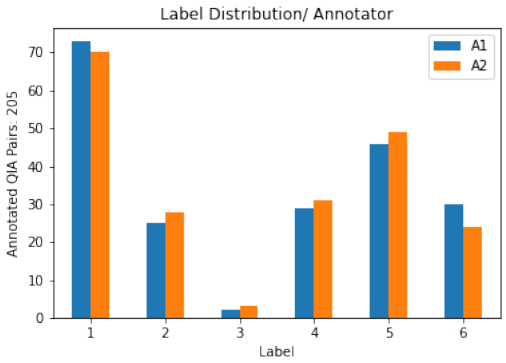

Figure 1: Label Distribution for 205 doubly-annotated QIA pairs in English.

quite low. It is different for comedy though; observed agreement is at 78.49, 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 Section 3.2. Label 6 and label 5 are more frequent in the comedy genre. Both annotators highly agree on label 5, "other", in comedy. In this case, both of them interpret that there is no clear answer.

Data split The French and Spanish QIA pairs in IndirectQA are only used for evaluation. The English QIA pairs are also used for training. Therefore, the English IndirectQA data is split up into a train, development and test set, in 80:10:10 proportions. As for the English IndirectQA part, 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.

4. Models

To evaluate the IndirectQA dataset, we implement baseline and cross-lingual transfer learning setups where we apply intermediate task training (Phang et al., 2020), inspired by experiments in Louis et al.

(2020). For classification, we use the MaChAmp toolkit[4] by van der Goot et al. (2021). We rely on the default hyperparameters, since they were found to be robust in multiple tasks (van der Goot et al., 2021). In this case, the multilingual BERT (mBERT) transformer model is used and every dataset for training, intermediate task training, and fine-tuning trains for 20 epochs each. As for the baseline models, we train (respectively fine-tune) mBERT on the English training and development set of the IndirectQA corpus.

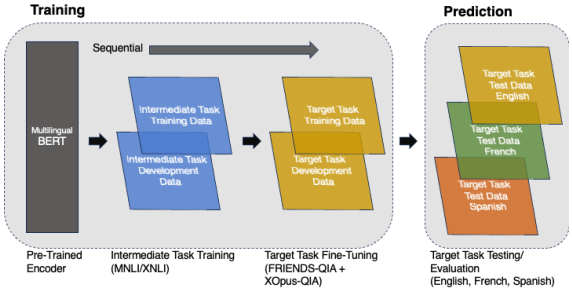

Figure 2: Overview of the sequential intermediate task training setup for the understanding indirect answers classification task. Figure adapted from Phang et al. (2020).

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. See Figure 2 for a schematic overview. We use MNLI data as intermediate task data, as that performed best in preliminary experiments, further detailed below.

4.1. Baseline

For the baseline model, we combine the FRIENDS-QIA corpus by Damgaard et al. (2021) with the English data of the IndirectQA to give the model more data to learn from. With the adaption that the "N/A" label used by Damgaard et al. (2021) is excluded from the FRIENDS-QIA. However, the "lacking context" label of IndirectQA remains in the corpus, so that during the training process this label will still be learned. The training set is made up of 5,486 labeled QIA pairs, the dev and test set have 687 QIA pairs each.

4.2. Intermediate Task Training Models

For the intermediate task training models, we train sequentially on intermediate task data and target task data, for 20 epochs each. The target task fine-tuning data is the concatenated FRIENDS-QIA and IndirectQA (FX) dataset, respectively the train and the development data. As intermediate task data we opt for the MNLI dataset for inferences (Williams et al., 2018), since it already performed decently in transfer learning approaches for question-answering models (Clark et al., 2019).

The MNLI dataset consists of 433,000 pairs of hypothesis, premise, and label for textual entailment. The dataset needs to be downsampled to meet limited hardware capacities for training. We downsampled it to the smallest intermediate task dataset used in preliminary results, which was 9k samples (size from BoolQ). The final MNLI-FX model pipeline then is as follows: it trains on the downsampled MNLI training (9,427 samples) and development sets (3,270 samples) and fine-tunes on FRIENDS-QIA + IndirectQA.

The XNLI dataset (Conneau et al., 2018) is based upon the MNLI dataset and contains human translated development and test data for several languages, including French and Spanish. We want to explore the results on XNLI and compare them to the English-only MNLI-FX model. We therefore train two models using the XNLI dataset: the French XNLI-FR-FX and the Spanish XNLI-ES-FX model. Both models use the downsampled English MNLI dataset for training, but the respective XNLI dataset for each language for development. For each language represented in the XNLI dataset, there are 2,490 premise-hypothesis pairs. The models then sequentially fine-tune on the FRIENDS-QIA + IndirectQA dataset.

5. Results and Discussion

In this section we discuss the results. We then analyze the results per genre and the F1-scores per label. All metrics refer, unless otherwise mentioned, to the evaluation dataset that contains QIA pairs for both genres.

5.1. Baseline

Overall performance In Table 3, first we observe that on French and Spanish, the zero-shot baseline model – trained on English data – performs low with 44.06 accuracy on both French and Spanish, and a macro F1-score of 33.64 for French, and 33.95 for Spanish. This is considerably lower than in-language English results (provided for comparison).

Results per Genre: Comedy is hard If we look at each genre, the results differ quite substantially from the overall performance scores (see Table 4). Interestingly the crime, drama, mystery genre is easier for the model and comedy turns out to be

[4] https://github.com/machamp-nlp/machamp, version 0.4.

Model	Model Name	Training Data		Test	Acc.	F1
English		FRIENDS-QIA + IndirectQA (en)		en	61.43	37.68
Baseline	FRIENDS-QIA + IndirectQA	FRIENDS-QIA + IndirectQA (en)		fr	44.06	33.64
				es	44.06	33.95
Model	**Model Name**	**Intermediate Task**	**Target Task**	**Test**	**Acc.**	**F1**
Intermediate	MNLI-FX	MNLI (en)	FRIENDS-QIA + IndirectQA (en)	fr	50.00	40.06
		MNLI (en)	FRIENDS-QIA + IndirectQA (en)	es	42.92	36.44
	XNLI-FR-FX	MNLI (en) + XNLI (fr)	FRIENDS-QIA + IndirectQA (en)	fr	44.06	35.97
		MNLI (en) + XNLI (fr)	FRIENDS-QIA + IndirectQA (en)	es	42.01	34.55
	XNLI-ES-FX	MNLI (en) + XNLI (es)	FRIENDS-QIA + IndirectQA (en)	fr	45.21	38.43
		MNLI (en) + XNLI (es)	FRIENDS-QIA + IndirectQA (en)	es	42.92	36.92

Table 3: Performance scores (Acc.=Accuracy, F1=Macro F1) of the baseline and intermediate task training models evaluated on French and Spanish data (EN given as reference). Green cells indicate improvement over the respective baseline, red drop below the baseline scores, no color is identical to the baseline score.

the more difficult genre. Models struggle more with humor in French, in contrast to the observed annotation difficulty, where humans struggled more with crime.

Lang	Genre	Acc.	F1	Genre	Acc.	F1
fr	Comedy	38.72	28.95	Crime	50.25	38.98
es	Comedy	36.17	27.71	Crime	53.20	41.13

Table 4: Results per genre, comedy vs crime (crime, drama, mystery).

F1-Scores per Label Unsurprisingly due to the skewed label distribution, performance varies considerably per class (Table 5). While "yes" reaches highest performance, some F1-scores remain even 00.00 – for example, for the infrequent "yes" (condition) for both French and Spanish. We next examine whether intermediate task training helps.

Lang	F1/ Label 1/ Yes	F1/ Label 2/ No	F1/ Label 3/ Yes (cond.)	F1/ Label 4/ Neither	F1/ Label 5/ Other	F1/ Label 6/ Context
en	75.15	62.20	20.00	43.70	25.00	00.00
fr	61.00	44.59	00.00	28.25	32.39	35.61
es	65.54	42.60	00.00	21.62	35.46	38.46

Table 5: Baseline F1-scores per label. English results provided for comparison.

5.2. Intermediate Task Training

As shown in Table 3, for French the intermediate task training with MNLI or XNLI surpasses the baseline . The MNLI-FX model performs not only best for French, but performs best overall (see Figure 3 in Section 8 for a detailed confusion matrix).

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. Even the Spanish XNLI-ES-FX performed better for the French data prediction with an accuracy of 45.21, which makes it the only model – along with the MNLI-FX model – that surpasses the baseline accuracy scores. This shows that cross-dataset effects do impact intermediate task training substantially, already visible when examined in terms of accuracy.

When evaluated on Spanish, the results drop below the baseline accuracy score of 44.06, thus intermediate task training hurts in turns of accuracy.

Instead, all F1-scores of the MNLI-trained intermediate task training models surpass the baseline scores – for both the French and Spanish test data. The overall best macro F1-score for the French test data is reached again with the MNLI-FX model. For Spanish, the XNLI-ES-FX model reaches the highest macro F1-score of 36.92 (see Figure 4 in Section 8 for a confusion matrix). As differences between accuracy and macro F1 clearly pertain, we analyse per-class F1 scores next.

F1-scores per Label We have seen that regarding overall performances, the accuracy and F1-score do not particularly differ and the FRIENDS-QIA + IndirectQA 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 6, the F1-scores per label are given for intermediate task training. In terms of "yes", it is overall the class with the best performance. The lowest score of 54.72 for this label is obtained with the Spanish QIA pairs and the XNLI-FR-FX model. Label 5, "other", is the second most frequent label in the IndirectQA dataset, and it seems difficult to predict, not even the baseline reaches a good score for this label. The best F1-score for the "other" class evaluated is for French at 38.71 and the one for Spanish is the highest at 38.67. The "no" class, which is a less frequent label than "other", still reaches better F1-scores of 47.46 for Spanish, and 44.93 for French.

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 + IndirectQA	FRIENDS-QIA + IndirectQA	en	75.15	62.20	20.00	43.70	25.00	00.00
		fr	61.00	44.59	00.00	28.25	32.39	35.61
		es	65.54	42.60	00.00	21.62	35.46	38.46
Intermediate Task Training								
MNLI/ XNLI	MNLI-FX	en	72.29	52.77	32.00	47.28	19.57	00.00
		fr	66.47	44.93	14.29	36.99	38.71	38.96
		es	59.04	45.86	33.33	22.82	34.39	23.19
	XNLI-FR-FX	en	70.55	55.72	26.09	48.44	30.11	00.00
		fr	58.69	41.72	22.22	24.79	35.06	33.33
		es	54.72	47.46	20.00	23.02	36.13	25.97
	XNLI-ES-FX	en	74.46	57.98	30.77	47.67	19.75	00.00
		fr	60.98	44.93	22.22	28.57	34.90	38.96
		es	57.04	44.83	16.67	25.84	38.67	38.46

Table 6: F1-scores per label for the intermediate task training models. Green cells indicate higher performance than the baseline.

When we look at the less frequent labels according to the label distribution (see again Table 2), the results bear promise regarding intermediate task training. As for "3/yes, subject to some conditions", we see consistent improvements over the baseline. All MNLI-based models, evaluated on all three languages, outperform the baseline for this label. Instead, the performance drops on "1/Yes".

The MNLI-based intermediate task training models also reach similar results for "4/neither yes nor no". All scores are higher than the FRIENDS-QIA + IndirectQA baseline scores, except for the XNLI-FR-FX model evaluated on French, astonishingly. 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 both languages.

Overall, the results point out that intermediate task training can be useful for less frequent classes, and depending what class might be prioritized, warrants a closer look despite overall first less promising results.

5.3. Discussion

As an overall result, the task of understanding indirect answers to polar questions remains quite challenging. Once the answer is not a clear *yes* or *no*, results are low. Nevertheless, we have seen that for these less frequent labels the intermediate task training improved the results sometimes clearly and the inference datasets MNLI and XNLI helped within the scope of cross-lingual transfer.

One explanation for low overall performance results might be the lack of context, especially for subtitles. It was observed that using subtitles without cues of who is speaking is difficult to understand and annotate even for humans. Furthermore, the skewed label distribution makes the task difficult.

6. Conclusion

To address the gap in resources, we introduced IndirectQA, a first dataset to help NLU models understand indirect answers beyond English, namely for French and Spanish. It consists of a total of 1,053 pairs of (implicit) polar questions and indirect answers, extracted from the OpenSubtitles corpus (Lison and Tiedemann, 2016) and manually annotated for six labels.

In contrast to prior work, we suggest to use fortuitous subtitle data for data collection. While using such data, we encountered several challenges, which we document in this paper, from lacking speaker information to finding triggers of indirect answers – since dialogues in subtitles are optimized to fit to images. Moreover ambiguity due to lack of context is noted, hence we introduced a new label "lacking context". Nevertheless, subtitles constitute interesting natural conversational data, and together with our broader definition of polar questions, we outline a method that we would like to encourage for uptake to extend IndirectQA's language coverage.

Our empirical evaluation shows large differences across labels and, most interestingly, genres, both for human annotation and model prediction. This shows that understanding indirect answers leaves open great potential for model adaptations and further research.

Acknowledgements

We thank Siyao (Logan) Peng and the anonymous reviewers for their feedback on earlier drafts of this paper, as well as Sif Dam Sonniks for their support on earlier parts of this work. This work is supported by ERC Consolidator Grant DIALECT 101043235.

7. Bibliographical References

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 multitask benchmark for evaluating cross-lingual generalization.

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.

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.

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.

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.

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.

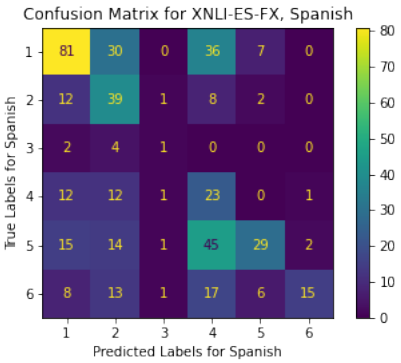

Figure 4: Confusion matrix for the XNLI-ES-FX model, tested on Spanish (Accuracy of 42.92, F1-score of 36.92).

8. Appendix: Confusion Matrices

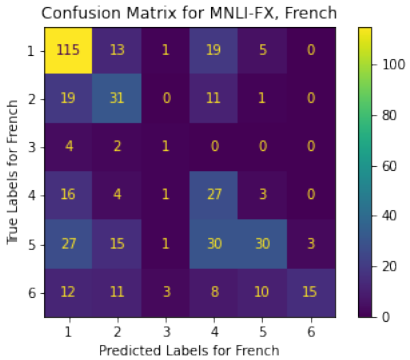

Figure 3: Confusion matrix for the MNLI-FX model, tested on French (Accuracy of 50.00, F1-score of 40.06).