Who Did You Blame When Your Project Failed?

Designing a Corpus for Presupposition Generation in Cross-Examination Dialogues

Maria Francis[1,2], **Julius Steuer**[1], **Dietrich Klakow** [1] **and Volha Petukhova**[1]
[1]Spoken Language Systems Group, Saarland University, Saarbrücken, Germany
[2] Language and Communication Technologies, University of Groningen, The Netherlands
{mfrancis,jsteuer,dietrich.klakow,v.petukhova}@lsv.uni-saarland.de

Abstract

This paper introduces the corpus for the novel task of presupposition generation - a natural language generation problem where a list of presuppositions carried by the given input sentence, in the context of the presented research - given the cross-examination question, is generated. For this, two datasets, PECaN (**P**resupposition, **E**ntailment, **C**ontradiction **a**nd **N**eutral) and PGen (**P**resuppostion **Gen**eration) are designed and used to fine-tune BERT (Devlin et al., 2019) and T5 (Raffel et al., 2020) models for classification and generation tasks respectively. Various corpora construction methods are proposed, ranging from performing manual annotations, prompting the GPT 3.0 model, to augmenting data from the existing corpora. The fine-tuned models achieved high accuracy on the novel Presupposition as Natural Language Inference (PNLI) task which extends the traditional Natural Language Inference (NLI) task, incorporating instances of presupposition into classification. T5 outperforms BERT by a broad margin, achieving an overall accuracy of 84.35% compared to 71.85% of BERT, and specifically when classifying presuppositions (93% accuracy of T5 vs. 73% of BERT). Regarding presupposition generation, we observed that despite the limited amount of data used for fine-tuning, the model displays an emerging proficiency in generation presuppositions reaching ROUGE scores of 43.47, adhering to systematic patterns that mirror valid strategies for presupposition generation, although failing to generate the complete lists.

Keywords: natural language inference, presupposition classification, presupposition generation

1. Introduction

A widely shared assumption in the research debate is that most of the information exchanged in human conversation is implicitly conveyed, where presuppositions represent a large chunk of this implicit information. According to the semantic view, a presupposition represents a condition for the evaluation of a sentence as either true or false. In pragmatics, the concept of presupposition is no longer linked to necessary conditions for the truth evaluation of a proposition, but for the felicity or appropriateness of a speech act (Austin, 1975). The focus shifts from the semantic level of sentences to the pragmatic level of utterances, treating presupposition as a propositional attitude including the 'cognitive context' of the speaker's beliefs, assumptions, and presumptions (Stalnaker, 1975). The presuppositional inferences of an utterance are modelled as beliefs of the dialogue participants when the utterance is produced and processed, and include beliefs about what is already accepted ('shared') by interlocutors.

Analysing cross-examination interactions that are largely based on question-answer exchanges, we observed that pivotal features of questions are based on their presuppositions. Therefore, presupposition identification and generation tasks hold significance in this domain. Presupposition recognition as part of Natural Language Inference (NLI) serves as evaluation of the effectiveness of Natural Language Understanding (NLU) models regarding their 'comprehension' of intricate semantic and pragmatic phenomena. It specifically tests the model's ability to grasp semantic information which is not explicitly expressed, but inferable from an utterance. In question answering, Kim et al. (2021) demonstrate how presupposition verification can be leveraged to formulate responses to unanswerable questions. In the context of cross-examination, generated presuppositions can be used to identify information taken for granted in presumptive questions such as one in the title of this paper, which carries the assumption that the addressee blamed somebody when their project failed. Presumptive questions inherently assume specific facts that cannot be answered without implicitly accepting or explicitly rejecting the proposition contained in their presuppositions.

In this paper, we explore the capability of language models to understand and extract information introduced by presuppositions and to differentiate between presupposed information and information conveyed by other types of inferences. Given a pair of sentences comprising a premise and a hypothesis, traditional NLI classification models

LREC-COLING 2024, pages 17564–17574
20-25 May, 2024.

Question Type	Definition	Relative Frequency (in %)	Example
Leading Question	A question that suggests a particular answer and contains information the examiner is looking to have confirmed (Melilli, 2003)	35.6	Dr. Radetsky, would it be fair to say, sir, that you are no stranger to the courtroom
Loaded Question	A question that has a presupposition that the respondent is not committed to (Walton, 1999)	27.8	Is there any chance you were trying to make it hard for adverse lawyers to find out the details of your testimonial history?
Propositional Question	A question asked in order to obtain the truth of a proposition (ISO, 2020)	15.1	And when you were in Jacksonville last year –do you remember going to Jacksonville to testify?
Check Question	A question asked in order to verify the truth of a proposition where the speaker weakly believes is true (ISO, 2020) In cross-examinations, often called *confirmatory* A question that leads to answers that can only support a certain point	11.1	In 1999, sir, every time you testified in court or by deposition, it was for a defendant, wasn't it?
Set Question	A question asked in order to obtain information which members of a certain set, described in the semantic content, have a certain property (ISO, 2020)	8.8	Out of these thirty patients you see a day, approximately how many are you prescribing medications for?
Imperatives	In (ISO, 2020) defined as request to provide information. In cross-examination, by phrasing questions in such a way, the speaker put an extra pressure to get the witness intimidated and/or confused (Logogye, 2016)	1.6	Tell us the facts that appear in your report that support the theory that Mr. Mead is genuinely disabled.

Table 1: Questions types annotated in cross-examinations provided with definitions and illustrative examples.

classify an inferential relation between them as either entailment, contradiction or neither (neutral). We introduce the Presupposition as Natural Language Inference (PNLI) task by including presupposition instances into classification. PNLI models are used as intermediate representations to perform presupposition generation in a transfer learning setting. Here, a list of presuppositions carried by a given utterance is generated. Two datasets, one for presupposition classification (called PECaN) and another for presupposition generation (called PGen), were designed and used to fine-tune BERT (Devlin et al., 2019) and T5 (Raffel et al., 2020) models on the PNLI classification and presupposition generation tasks.

The paper is structured as follows. Section 2 introduces the domain of cross-examination, discusses question types and presuppositions carried by them. In Section 3, related work is addressed. We present existing NLI corpora and discuss recent advancements in the NLI task from the classification and generation perspective. Section 4 presents the data collection methods. Experiments applying the collected data are described, and the results and their analysis are detailed. We elaborate upon the methodology for the fine-tuning of BERT and T5 models (Section 5). Finally, Section 6 summarises our findings, addresses limitations and outlines directions for future research and development.

2. Presumptive Questioning in Cross-Examinations

Cross-examination is the formal interrogation of a witness called by the other party in a court to challenge or extend testimony already given. It has been shown that parties involved use different question forms in cross-examinations compared to casual conversations (Seuren, 2019). A question's form may amplify or diminish peoples' tendencies to agree or disagree, to speak openly or save face, and to feel threatened or comfortable. Presumptive questioning may trigger non-representative answers, and addressees may fail to realise that the answers are shaped by the questions asked (Swann et al., 1982). Presumptive questions take "non-default responses" as a given and ask the respondent to search mentally for substantive replies leading to answers that otherwise would be denied with default responses when questions were asked directly (Kellermann, 2007). While leading questions may be useful, e.g. to fresh up memory, respondents can be profoundly influenced by them. For instance, "false confessors" to crimes are highly susceptible by leading questions (Gudjonsson, 1984). Loaded questions presuppose at least one unverified assumption and may put the person being questioned in a disadvantageous and defensive position, since the assumption in the question could reflect badly on them or pressure them to answer in a way that they would not otherwise. In cross-examinations, is therefore important to identify presumptive questions and understand information they presuppose.

Leading and *loaded* questions as form of suggestive interrogations, prevail in court rooms, see Table 1. Consider an example of a loaded question (premise) and its presuppositions (hypotheses):

(1) **Premise**: Doctor, it is very misleading to tell this jury that your legal work is done at four in the morn-

Trigger Type	Example
Factive verbs	Do you recall reading Mr. Kitchen's deposition in which he said that after 1991, Mr. Mead appeared withdrawn, when a person is defeated, your posture is not quite the same, he was a little beaten, do you remember that?
Aspectual verbs	What did he lose when he stopped working? When did the malingering start?
It-clefts	It is something you want to run away from, your history of which cases you've testified in?
Wh-clefts	What I thought I asked — well, I'll try to ask better — is what is your total bill going to be?
Quantifiers	Most attorneys and forensic psychiatrists consider it the responsibility of the forensic psychiatrist to put a spin on the data and highlight and emphasise facts favorable to their side and de-emphasise or even ignore data that are not. Do you agree with that statement?
Focus particles	You've charged $20,000.00 in this case up until recently when you began to prepare for it again.
Definites	In the 450 cases in which you've testified, either by deposition or at trial, in 95 percent of the cases you've testified for the defendants and against the patient?
Possessives	I try to give the same testimony no matter who retains me, but it was Mr. Craig Dennis's firm.

Table 2: Presupposition triggers types observed in cross-examinations and illustrated with examples.

ing and on weekends, isn't it, sir? >>
Hypothesis 1: There is a jury.
Hypothesis 2: There is a doctor addressed in the statement.
Hypothesis 3: The doctor is involved in legal work.
Hypothesis 4: The doctor has informed the jury about the timing of their legal work.
Hypothesis 5: The doctor claimed that their legal work is done at four in the morning and on weekends.
Hypothesis 6: The speaker believes that the doctor's claim is misleading.

Hypothesis 6 in (1) is the presupposition which contains information critical to the addressee's (doctor's) assertion and implies that the doctor's claim may not be truthful or accurate.

Presuppositions may be triggered by a broad range of linguistic items. Three key groups are defined - existential, lexical, and structural presuppositions. An *existential* presupposition presupposes the existence of a reference, e.g. in case of definite descriptions or proper names. A *lexical* presupposition is triggered by a certain lexical item, e.g. factive verbs. A *structural* presupposition arises from the syntactic structure of an utterance, e.g. cleft constructions. Table 2 provides an overview the presupposition trigger types observed in the analysed cross-examinations illustrated with examples.

The presuppositions induced by simple utterances tend to survive under embedding, albeit often in a weakened form. They remain valid even when the trigger is subject to negation, is modalised, questioned or conditioned, see example in (4).

3. Related Work

The adaptation of a 'commonsense' definition of inference - as when "a human reading premise would infer that hypothesis is most likely true. . . [given] common human understanding of language [and] common background knowledge" (Dagan et al., 2005) - has gained considerable traction in the most recent work. It is however unclear which types of inference, semantic or pragmatic, if any, are truly learned by the modern NLI models (Jeretic et al., 2020). This is complicated by the fact that there is also no exhaustive theory of how humans draw inference.

3.1. Related Tasks

Several tasks exist that are conceptually close to the PNLI task. One of the first ones to discern the inferential relationship between two sentences is Recognising Textual Entailment (RTE, see Dagan et al. (2005)). In this context, when presented with a text (T) and a hypothesis (H), the objective is to decide whether T infers H. The RTE task classifies a sentence pair as having either an entailment or a non-entailment relation. Systems participated in the PASCAL Recognising Textual Entailment Challenge achieved accuracy ranging from 50% to 60%.

Recently, the NLI task, conceptually an extension of the original RTE task, has gained considerable research attention. Given a premise (P) and a hypothesis (H), the task becomes categorising the relationship between the two as either entailment, contradiction, or neutral (neither entailment nor contradiction). Both the RTE and the NLI tasks, along with the re-framing them for Question Answering (QA), are included within the GLUE benchmark (Wang et al., 2018), a widely accepted and used evaluation NLU benchmark. In single NLI related task training, BiLSTM performance achieved in terms of accuracy ranges from 53.5%

on textual entailment recognition (RTE), 76.9% when classifying natural language inferences in multi-genre texts (MNLI), and 77.2% when classifying natural language inferences in question-answering data (QNLI).

The task which is conceptually most related to our Presupposition Generation (PGen) task is the task of Generating Textual Entailment (GTE). Here, given an input text (T), the objective is to generate a context that can be inferred from T. Jia (2008) presents an approach where templates for semantically equivalent sentences are created and stored in a rule base. The task is designed to be performed by chatbots which save information acquired during a dialogue as part of the discourse context. The chatbot is expected to discern the novel information from what is already conveyed earlier. A large set of more than 10,000 rules is generated. The inference rule sets (e.g. 231,000 unique rules induced in (Lin and Pantel, 2001)) may easily occupy hundreds of megabytes of working memory which is often seen as an obstacle for real-time generation of chatbot responses. Nevěřilová (2014) presents an alternative rule-based method to GTE based on synonym/antonym replacement, anaphora resolution, and verb valence lexicons. Human based evaluation showed that 47.1 % of paraphrases and entailments were judged as correct, 37.3 % as incorrect, and 15.6 % as non-sense. Kolesnyk et al. (2016) adapts the sequence-to-sequence Recurrent Neural Network introduced in (Cho et al., 2014) for the GTE task, achieving accuracy of 82%. Guo et al. (2017) trained a residual LSTM network and obtained a rate of correctly generated entailments of 78.2%.

Whereas previous work on GTE generate a single output sentence for each input utterance, we aim to generate an ideally complete list of all presuppositions carried by an utterance.

3.2. Related Datasets

The popularity of the NLI task has led to the development of diverse datasets such as SNLI (Bowman et al., 2015), MultiNLI (Williams et al., 2018), SICK (Marelli et al., 2014), XNLI (Conneau et al., 2018) and bilingual Spanish-English esXNLI (Artetxe et al., 2020).

In these datasets, there are very few items where the relation between premise and hypothesis is of presuppositional nature. In the MultiNLI dataset, only eight presupposition pairs were found on which two or more annotators agreed on (Jeretic et al., 2020). Of these eight presuppositions, all but one are of existential type. Even with this small amount of instances of presuppositions in data used for fine-tuning, Jeretic et al. (2020) show that NLI models have the capacity to perform presuppositional reasoning to a certain extent. They introduce ImpPres, a semi-automatically generated dataset of premise-hypothesis pairs comprising both implicature (Imp) and presupposition (Pres) relations. The ImpPres dataset includes eight distinct types of presupposition triggers, all of which appear under negation, modals, interrogatives, and conditionals, which serves to assess a model's grasp of the fact that presuppositions project over these embeddings. As a control, they include entailment relations, as well as versions where the entailment is cancelled under negation. The authors fine-tuned three models on MultiNLI and test them each on the ImpPres dataset: a Bag-of-Words (BOW) model, InferSent (Conneau et al., 2017), and BERT-Large (Devlin et al., 2019). The BERT NLI model and the BOW model perform relatively well on five out of eight sub-datasets, namely those containing presuppositions triggered by the adverbial 'only', cleft constructions, possessives, and questions. The authors assume that this reasoning was successfully learned during the pre-training phase.

Limited datasets are available for the Generating Textual Entailment (GTE) task. To train the LSTM models, Guo et al. (2017) used a subset of the SNLI (tag) dataset where the labels correspond to entailment, omitting non-entailment pairs.

4. Data Collection: Design and Corpus Overview

Initial in-domain data was collected from the cross-examination interactions published by a number of excellent attorneys, transcribed and annotated by Patrick Malone.[1] The corpus consists of 13 interactions comprising 4259 utterances from which 1509 are questions: 34.7% of leading type; 27.8% are loaded questions; 15.1% -propositional questions; 11.1% - check or confirmatory questions; 8.8% - set questions, and 1.6% are direct requests or imperatives. Table 1 provides an overview of annotated questions with definitions and illustrative examples from cross-examination dialogues. Two independent annotators (a linguist and a lawyer) annotated questions and reached a moderate agreement of 0.36 in terms of kappa scores. However, near perfect agreement (kappa scores of 0.8) was observed when annotating presumptive questions: loaded and leading ones.

For all questions in one randomly selected cross-examination dialogue, we manually generated 367 presuppositions, 5.5 presuppositions per question on average. Given the small size of the data and

[1] https://www.patrickmalonelaw.com/useful-information/legal-resources/attorneys/legal-resources-attorneys-injured-clients/cross-examination-transcripts/

the rather high costs of manual generation of presuppositions, the set of presupposition instances was further automatically enriched by: (1) using SpaCY's[2] dependency parser, noun chunker, and NER models; (2) prompting OpenAI's GPT-3.0; (3) exploiting presupposition projection properties; and (4) re-annotating items from existing NLI corpora. Corpora construction methods are detailed in the next sections and summarised in Table 3.

4.1. Generating Presuppositions with Dependency Parser

We first based our approach on trigger detection largely adopting the classification scheme proposed by (Khaleel, 2010). Existential presuppositions and those carried by adverbial clauses were generated as they occur the most frequently in our data, covering about 57% of all cases.

Existential presuppositions are triggered by definite descriptions, proper names and possessives. For example,

(2) **Premise**: By the way, these inconsistencies in your testimony today versus your testimony before, that might raise in your mind a credibility issue similar to what you've told us about Mr. Mead. >>
Hypothesis 1: There is person named Mr. Mead.
Hypothesis 2: There is the Mr. Mead's testimony.
Hypothesis 3: There was the Mr. Mead's testimony before.
Hypothesis 4: There are inconsistencies between the Mr. Mead's today's testimony and their previous testimony.

SpaCy's language model was used to identify noun chunks and to filter indefinite descriptions. Definite descriptions and noun phrases that include a possessive element were combined with the pre-defined patterns *"There is an X"* or *"X has a Y"*. To generate presuppositions from proper names, SpaCY's Named Entity Recognition model was used, which not only extracts named entities but also classifies them as locations, persons, or organisations. We generate the presuppositions with the format *'There is a* $\langle class \rangle$ *called* $\langle entity \rangle$*'*, resulting in presuppositions such as *'There is a country called Spain'* or *'There is a person called Mr. Mead'*.

To automatically detect **adverbial clauses**, an expanded word list comprising the presupposition-triggering adverbs, mostly temporal ones, observed in the cross-examination questions was constructed. When an adverb from the word list was detected in a premise, the adverbial clause was extracted and rephrased into a proposition. For example,

(3) **Premise**: We liked LG's Heart Rate Monitor Earphones best after testing a few brands >>
Hypothesis: We tested a few brands.

All automatically generated presuppositions were manually checked.

4.2. Generating Presuppositions using GPT 3.0

A relatively large set of presuppositions was generated prompting OpenAI's GPT 3.0[3]. We used the `text-davinci-003` model with a maximum output of 100 tokens and a high temperature of 0.6, allowing for diverse outputs. Several prompts were tested, where some resulted in a single hypothesis generated while others returned a list. The prompt that worked most reliably and that was finally used to construct the corpus was *"Generate all possible presuppositions carried by this sentence: PREMISE "*. It was observed that specifically, including the word *'carried'* caused the model to generate a higher ratio of presuppositions to entailments. Per input, about ten hypotheses were generated for each of the 150 premises, resulting in a total number of 1556.

Essentially, we analysed GPT-3's zero-shot performance on the task of presupposition generation. We found that all hypotheses generated were grammatically correct and semantically coherent. By closer inspection, however, we found that most generated hypotheses were entailments, accounting for 84.6% of all output utterances, with presuppositions contributing to another 10.3%. The remaining 5.1% were contradictions and neutral instances. We experimented with the maximum output length of GPT 3.0, testing whether a lower maximum output length would cause the model to generate less entailments, causing the ratio of presuppositions to entailments to increase. A lower maximum output length lead to some of the presuppositions being replaced by entailments since they were more relevant in discourse. We concluded that the model is not aware of the differences between a presupposition and entailment.

4.3. Exploiting the Property of Presupposition Projection

Presuppositions carried by a sentence remain valid even when the proposition of the original sentence is negated, modalised or questioned, see the example in (4). We exploited this property to enrich our dataset accordingly, and tested whether models are able to learn this phenomenon in the fine-tuning experiments.

To negate the premise, SpaCy's dependency parser identified the root of a sentence, whose main verb was further negated. We embedded

[2] spaCy is a library for advanced Natural Language Processing in Python, see https://github.com/topics/spacy

[3] https://openai.com/blog/gpt-3-apps

Collection method	PCaN data						PGen data				
	Total instances		from all hypothesis (in %)				Total instances permise-hypothesis pairs	from all hypothesis (in %)			
	premise	hypothesis	P	E	C	N		Existential	Lexical	Structural	Contextual
Cross-examination dialogue	74	367	100	-	-	-	367	72.2	10.4	5.1	12.3
From SpaCy parsed output		141	100	-	-	-	141	50.0	-	50.0	na
Prompting GPT 3.0	150	1556	10.3	84.6	0.5	4.6	150	*triggers not annotated*			
Projected	1175	235	100	-	-	-	235	10.1	60.6	29.3	na
From existing NLI corpora	7611	7611	13.9	16.0	33.0	37.1	2511	10.1	60.6	29.3	na
Total collected	9010	9910	28.3	23.2	22.6	25.9	3404	17.2	49.9	26.2	6.7
Total used for fine-tuning	2000	2000	25.0	25.0	25.0	25.0	premise: 444 hypothesis: 1532	51.5	28.4	7.8	12.3

Table 3: PCaN and PGen corpora overview. Hypotheses are of type of entailment (E); presupposition (P); contradiction (C); and neutral (N). Note: hypotheses in PGen data are all of presupposition type.

the original premises in modal and interrogative constructions using Nodebox English Linguistics library to change the tense of the root verb to infinitive and to add a modal auxiliary verb, e.g. might, and to create a syntactic construction applicable for imperatives, as illustrated in (4):

(4) **Premise**: You are aware that Dr. Lees-Haley was a contributing editor to Claims Magazine.
Modified (intermediate) premise: You be aware that Dr. Lees-Haley was a contributing editor to Claims Magazine.
Modalised premise: You might be aware that Dr. Lees-Haley was a contributing editor to Claims Magazine.
Interrogative premise: Are you aware that Dr. Lees-Haley was a contributing editor to Claims Magazine.
Conditional premise: If you are aware that Dr. Lees-Haley was a contributing editor to Claims Magazine for a number of years, then you know he has extensive experience in the field.
>>
Hypothesis: Dr. Lees-Haley was a contributing editor to Claims Magazine.

To modify the original premises as the premise of a conditional, GPT-3.0 was prompted to *"Rephrase the following sentence as the premise of a conditional, and generate a conclusion"*. This method has an advantage over the template-based method used when constructing `ImpPres`, since rather high lexical and syntactic variability was achieved.

4.4. Collecting Presupposition Items from the Existing NLI Datasets

For the classification and generation models to effectively learn the essence and patterns of presuppositional reasoning, as well as the differences between inferences of various types, models encountered a variety of (non-)presuppositional instances aggregated from two existing corpora - `MultiNLI` (Williams et al., 2018) and `ImpPres` (Jeretic et al., 2020).

The premises in `MultiNLI` are derived from ten different genres of written and spoken English. For each premise, three distinct hypotheses were defined: an entailment, a contradiction, and a neutral statement. This approach guarantees a balanced distribution of instances across the three designated classes. We manually checked entailments in `MultiNLI`, and relabeled them as presuppositions when applicable.

`ImpPres` was designed to assess the ability of NLI models fine-tuned on `MultiNLI` to perform pragmatic reasoning. It encompasses a wide range of presupposition trigger types under a variety of entailment-cancelling embeddings. However, a template-based generation jeopardised verbal fluency of the sentences. In `ImpPres`, the vocabulary often features rare combinations, resulting in grammatical but nonsensical sentences.

4.5. Converting the PECaN Dataset into the PGen Dataset

The PECaN dataset, designed for the task of presupposition classification, was converted into the PGen data to be used specifically for the task of presupposition generation. Contrary to the PECaN dataset which consists of premise-hypothesis pairs, each training instance in the PGen dataset consists of a premise and a list of presuppositions that are carried by the premise. The subset of PECaN items labeled as 'presupposition' is searched and sorted, whereby hypotheses appearing in presupposition items with the same premise are collected and stored in a list. These hypothesis lists are examined for completeness, and any missing presuppositions are manually filled in. In total, PGen consists of 444 instances, each consisting of a premise and a list of on average 3.4 hypotheses carried by it. Table 3 provides an overview of the different methods used for corpora construction, including frequency details for the inference and presupposition types.

5. Application of the Collected Data

As an intermediate step towards presupposition generation, we investigated the ability of language models to differentiate between presuppositions

Task	Model	Accuracy (%)
NLI	BERT	69.64±0.88
	T5 zero-short	84.82
	T5 finetuned	**86.20**±1.20
PNLI	BERT	71.85±0.81
	T5 zero-short	-
	T5 finetuned	**84.35**±0.94

Table 4: Performance of BERT and T5 fine-tuned models on the NLI and PNLI tasks.

and the other types of inference, in particular entailments. The model performance was specifically assessed on the four-class classification.

5.1. Fine-Tuning Experiments

Using the `PECaN dataset`, two models were fine-tuned on our 4-class NLI task: BERT (Devlin et al., 2019) and T5 (Raffel et al., 2020). BERT is a decoder-only transformer model pre-trained on a masked language modelling task. BERT can be fine-tuned on classification tasks by extending it with a classification head, using the contextualised embedding of a [CLS] token as input to the head. T5, or Text-to-Text Transfer Transformer, is an encoder-decoder model designed for transfer learning and is pre-trained on a suite of language tasks, including the standard NLI one based on the `MultiNLI` dataset. We used the base version of both models from HuggingFace transformers (Wolf et al., 2020), and fine-tuned them with the hyperparameters as suggested by Devlin et al. (2019): 10 epochs (BERT-base) and 8 (T5-base); $2e{-}10$ learning rate and warmup of 1% . We initialised the parameters of BERT's classification head with five different random seeds, and used five-fold cross-validation for T5. For each model, we report the average performance in terms of accuracy, as well as standard error estimates.
`PECaN` data was split into 80% train, 10% validation and 10% test sets, with classes equally distributed over the three sets. Given the novelty of the defined fine-tuning task, there is no benchmark model that we can directly compare our results with. Therefore, both models, BERT and T5, were trained on the standard NLI task of three-way classification to establish a baseline.

5.2. Results

The performance of T5 is markedly better than that of BERT on both the NLI and PNLI tasks. The model fine-tuned on the PNLI task, however, tends to classify instances as entailment, i.e. it falsely predicts instances of contradictions and neutrals as entailments, where the biggest difference is observed in the 'neutral' class, see Figure 1.
In the case of T5, there are very few cases where instances of contradiction (2%), entailment (4%), or neutral (1%) classes were confused with presuppositions. Our output analysis confirmed the conclusion made in previous research that presupposition instances exhibit a rather high lexical and syntactic overlap with their premises (Jeretic et al., 2020; Kabbara and Cheung, 2022). Many structurally triggered presuppositions were mostly constructed from a subset of the vocabulary used in the premise. Vocabulary found in presupposition and entailment instances mostly originated from a legal setting compared to the vocabulary of contradiction and neutral items. Our dataset augmentation approach of generating new premise - hypothesis pairs by converting the premise of an existing pair into a negated, interrogative, conditional clause further reinforced this property of the dataset. The models learned to take advantage of these recurring patterns, resulting in a high performance on the test set, but making it hard to generalise over less frequent instances. During the fine-tuning of BERT on the PNLI classification task, our models are consistent with the model having learned to classify presuppositions solely based on the premise. Per presupposition instance, the dataset contains five instances between which the propositions and the vocabulary were almost identical, and some of these instances inevitably ended up in training data, while others ended up in the test set. To mitigate the influences of these factors, we randomly selected one of the five premises under entailment-cancelling embeddings per hypothesis.

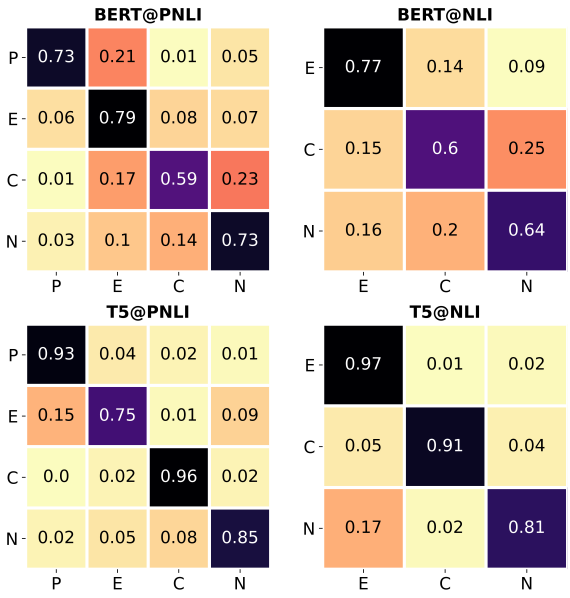

Figure 1: Model performance of BERT and T5 in terms of accuracy on the PECaN task, both including (left column, PNLI) and exlcuding (right column, NLI) presupposition items.

Looking at the by-class performance of the T5

Model	ROUGE-1	ROUGE-2	ROUGE-L
T5 (CNN/DM)	43.47	20.30	40.63
T5 (PGen)	50.49±2.58	37.52±2.29	45.62±2.15

Table 5: Performance on presupposition generation in terms of ROUGE scores.

models (Figure 1), it can be observed that T5 outperforms BERT by a broad margin, classifying presuppositions with 93% accuracy compared to the BERT accuracy of 73%. T5's performance on presupposition classification significantly surpasses that of the entailment and neutral classes, which achieve 75% and 85% accuracy respectively, and is comparable with the accuracy on contradiction classification of 96%. By contrast with the BERT performance on entailment classification, T5 more often confuses entailment with presuppositions. This can be attributed to the fact that T5 is pre-trained on the `MultiNLI` dataset and our `PECaN` dataset includes many `MultiNLI` instances. Thus, it is likely that there are instances in our test set that the T5 model has already seen during pre-training. Moreover, aforementioned lexical and structural similarities between premise and hypothesis impacted the performance as well. To reduce the proportion of these pattern similarities, one solution could be to (para)rephrase the premise and its hypotheses, or to generate adversarial instances by negating the hypothesis of the presupposition instances, and to integrate them into the dataset as instances of either contradiction or neutral classes.

5.3. Presupposition Generation: Quantitative and Qualitative Analysis

The T5 model fine-tuned on PGEN generally failed to generate complete lists of hypotheses, i.e. a set of **all** presuppositions carried by the premise. However, the completeness and quality of these lists needs to be further assessed in full-fledged human-based evaluation experiments. We performed a primary detailed analysis of the model performance on the test data.

From the 89 input utterances in the test set, 12.2% of output hypotheses were presuppositions, 23% entailments, 13% neutral, and 6% belonged to the contradiction class. The model generated about 15.3% ungrammatical outputs, which were mostly incomplete utterances. Some hypotheses duplicated the premise, either as-is, or with some discourse markers removed, e.g. interrogative elements in tag questions.

While the ratio of successfully generated presuppositions was rather low, the generated output utterances follow a set of observable strategies. In many cases, output hypothesis were lexically or semantically very close to their respective target presupposition. Nevertheless, we concluded that despite having a limited amount of training data, the model was able to develop a logic for presupposition generation. The following effective strategies were observed: 1) extracting a complement clause from the complex premise, and using it as the hypothesis; 2) exploiting the entailment-cancelling embedding of the input utterance; 3) removing interrogative elements and other verification markers from the input utterance.

Presuppositions triggered by adverbial clauses, cleft constructions or factive verbs were generated by extracting an adverbial clause or other complement clause from the input premise. This appeared to be a reliable strategy for presupposition generation and was successfully learned by the model. Seven out of the twelve generated presuppositions were produced following this strategy. For example,

(5) **Premise**: If you are aware that Dr. Lees-Haley was a contributing editor to Claims Magazine for a number of years, then you know he has extensive experience in the field. >>
Hypothesis: Dr. Lees-Haley was a contributing editor to Claims Magazine for a number of years.

A considerable portion of the generated utterances were a result of learning from entailment-cancelling embeddings of the input utterance, with the most frequent changes involving switching from a conditional or interrogative statement to the 'original' one. For example, when the input utterance is a conditional, the model often extracts a clause from the conditional premise. In itself, this is a valid strategy, as it reflects the design of our training dataset, i.e. the majority of the input utterance that are conditionals were generated by re-embedding the original presupposition-carrying clause as the premise of a conditional.

When the model encountered an interrogative utterance, it often returned the proposition embedded in the interrogative as a statement. This strategy works well for most Set and Propositional Questions.

Many cross-examination input utterances are Check (Confirmatory) Questions, typically asked in the form of tag questions. Tags were omitted in the hypothesis. This strategy is effective and valid as long as the question tag contains a factive verb, such as *"do you agree?"* or *"do you recall?"*.

On occasion, if the input utterance was negated, our model returned a non-negated version of the input utterance, resulting in an output that directly contradicts the input.

6. Discussion, Limitations and Conclusion

In this paper, we have introduced two novel datasets, Presupposition Classification (PECaN)

and Presupposition Generation (PGEN), used to perform presupposition classification and generation tasks respectively. Both will be made available for the research community to deploy. Given the limited data size, the popularity and the success of the pre-trained models, BERT and T5 models were fine-tuned to extend the standard three-class NLI task for presupposition classification. Both models demonstrate rather high accuracy in classifying presupposition instances, where T5 with accuracy of 93% significantly outperforms BERT (73%). However, it is still unclear whether models were able to learn patterns underlying presuppositional reasoning, or whether the high performance is largely attributed to the data structure and its nature. The latter reflects the concern prevalent in previous research that trained and fine-tuned models may take advantage of reoccurring lexical and structural patterns unique to the presupposition class. In our data, presupposition instances exhibit a rather high degree of lexical overlap with their premises. The vocabulary of the premises and hypotheses of the presupposition and entailment classes mostly feature a legal domain, while contradictions and neutral classes largely do not. The modification of entailment-cancelling embeddings of premises incorporating modalised and conditional clauses introduced additional 'systematicity' into the data, along with frequent semantic inconsistencies and grammatical errors; all this is applicable exclusively for the presupposition class. Effects of these patterns on model generalisation became apparent during model evaluation and detailed output analysis.

Regarding presupposition generation, we concluded that despite the limited amount of data used for fine-tuning, the model displays an emerging proficiency in generation presuppositions. The generated hypothesis often adhere to well-defined patterns that mirror valid strategies for presupposition generation. This outcome confirms the model's capacity to formulate reasonable strategies for accomplishing this task effectively, endorsing presupposition generation to the promising and fruitful research endeavor. The goal of generating lists of hypotheses was however only partially achieved and requires further development efforts.

The complexity of the NLI tasks, the diverse array of triggers and reasoning patterns calls for enhancement in corpora size, structure and content. A more sophisticated strategy is required in generating premise-hypothesis pairs with presupposition and entailment relations, but also contradiction instances ensuring the between- and within-class balance. Future work will focus on identifying additional data sources from diverse domains and exploring further methods that reduce the reliance on manual annotations.

A serious limitation of the presented and related studies pertains to the lack of an adequate automated assessment metric for the generation task. ROUGE as a token overlap-based metric does not appropriately reflect the quality of the generated presuppositions. An automatic semantically adequate evaluation metric is required. For this, Word Mover's Distance (WMD, (Kusner et al., 2015)), or any other metric based on word2vec embeddings, can be considered. We also aim to create a standardised human-based qualitative evaluation procedure to assess the criteria of completeness of a generated list of hypotheses, semantic adequateness, linguistic fluency, and grammaticality of their members.

In conclusion, presupposition classification and generation are highly promising tasks that may find extensive application in fields of question answering and dialogue systems design. We are optimistic that the corpora we have made available will prove valuable to the broader NLP community, and that this work will catalyse renewed endeavors in the realm of generative language models.

6.1. Datasets Availability

The datasets and code are available on https://github.com/uds-lsv/presupposition-generation.

Acknowledgments

The authors are also very thankful to anonymous reviewers for their valuable comments. One of the authors is supported by the Erasmus Mundus Master's Programme "Language and Communication Technologies". [4]

7. Bibliographical References

John Langshaw Austin. 1975. *How to do things with words*, volume 88. Oxford university press.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Lo Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *CoRR*, abs/1705.02364.

[4] https://lct-master.org/

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190.

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.

Gisli H Gudjonsson. 1984. Interrogative suggestibility: Comparison between 'false confessors' and 'deniers' in criminal trials. *Medicine, Science and the Law*, 24(1):56–60.

Maosheng Guo, Yu Zhang, Dezhi Zhao, and Ting Liu. 2017. Generating textual entailment using residual LSTMs. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data: 16th China National Conference, CCL 2017, and 5th International Symposium, NLP-NABD 2017, Nanjing, China, October 13-15, 2017, Proceedings 16*, pages 263–272. Springer.

ISO. 2020. *Language resource management – Semantic annotation framework – Part 2: Dialogue acts. ISO 24617-2, Second Edition*. ISO Central Secretariat, Geneva.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Jiyou Jia. 2008. The generation of textual entailment with NLML in an intelligent dialogue system for language learning CSIEC. In *2008 International Conference on Natural Language Processing and Knowledge Engineering*, pages 1–8. IEEE.

Jad Kabbara and Jackie Chi Kit Cheung. 2022. Investigating the performance of transformer-based NLI models on presuppositional inferences. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 779–785, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Kathy Kellermann. 2007. Persuasive question asking: how question wording influences answers. In *Annual Meeting of the State Bar Association of California, Anaheim, CA*.

Layth Muthana Khaleel. 2010. An analysis of presupposition triggers in english journalistic texts. *Journal of College of Education for Women*, 21(2):523–551.

Najoung Kim, Ellie Pavlick, Burcu Karagol Ayan, and Deepak Ramachandran. 2021. Which linguist invented the lightbulb? presupposition verification for question-answering. *arXiv preprint arXiv:2101.00391*.

Vladyslav Kolesnyk, Tim Rocktäschel, and Sebastian Riedel. 2016. Generating natural language inference chains. *arXiv preprint arXiv:1606.01404*.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question-answering. *Natural Language Engineering*, 7(4):343–360.

CA Logogye. 2016. Syntactic and pragmatic analysis of cross-examination in ghanaian law courts. *Journal of Literature, Languages and Linguistics*, 26:24.

Kenneth J Melilli. 2003. Leading questions on direct examination: A more pragmatic approach. *Am. J. Trial Advoc.*, 27:155.

Zuzana Nevěřilová. 2014. Paraphrase and textual entailment generation in Czech. *Computación y Sistemas*, 18(3):555–568.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Lucas M Seuren. 2019. Questioning in court: The construction of direct examinations. *Discourse Studies*, 21(3):340–357.

Robert Stalnaker. 1975. Presuppositions. In *Contemporary Research in Philosophical Logic and Linguistic Semantics: Proceedings of a Conference Held at the University of Western Ontario, London, Canada*, pages 31–41. Springer.

William B Swann, Toni Giuliano, and Daniel M Wegner. 1982. Where leading questions can lead: The power of conjecture in social interaction. *Journal of Personality and Social Psychology*, 42(6):1025.

Douglas Walton. 1999. The fallacy of many questions: On the notions of complexity, loadedness and unfair entrapment in interrogative theory. *Argumentation*, 13:379–383.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

8. Language Resource References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 1–8.

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