Universidad del País Vasco Euskal Herriko Unibertsitatea

Entailment for Zero- and Few-Shot Text Classification in Catalan: Monolingual vs. Multilingual Resources and Task Transfer

Author: Irene Baucells de la Peña

Advisors: Oier López de Lacalle Lecuona

hap/lap

Hizkuntzaren Azterketa eta Prozesamendua
Language Analysis and Processing

Final Thesis

February 2023

Departments: Computer Systems and Languages, Computational Architectures and Technologies, Computational Science and Artificial Intelligence, Basque Language and Communication, Communications Engineer.

Abstract

As the field of NLP continues to evolve and expand in industry, new tasks and languages emerge for which task-specific data for fine-tuning is often scarce or unavailable. Against this background, zero- and few-shot methods are gaining ground. However, most of them have typically been studied in the context of English and often rely on leveraging extensive pre-existing resources, raising questions about their applicability to less resourced languages. The current study investigates the application of one of these approaches, based on transforming the target task into a Natural Language Inference (NLI) task and using an NLI (or entailment) model to solve it, in the context of Catalan, a medium-sized language. Specifically, we address a multi-class text classification problem and ask whether (smaller) monolingual resources can compete with (larger) multilingual resources in such framework, experimenting with different combinations of pre-trained language models (LM) and NLI datasets to gain further insight into the contribution of each. In addition, we explore task transfer learning for potential performance improvements. Our results show that the larger size and richness of multilingual NLI datasets, and to a lesser extent the amount of text seen during LM pre-training, are key to the superior performance of multilingual models in the zero-shot setting, yet the monolingual LM seems to gain significance when the task requires a finer-grained classification. In contrast, in the few-shot setting, the weight of the base NLI dataset appears to decrease considerably and the monolingual LM becomes a stronger option. In turn, task transfer learning significantly improves the monolingual results in the zero-shot scenario, but becomes less relevant in the few-shot scenario. Overall, our study demonstrates the potential and limitations of the approach in resource-limited settings, providing insights into the factors influencing the entailment models' performance and highlighting areas for future improvement.

Contents

List of Figures

List of Tables

1 Introduction

Today's well-known successes in NLP results from the concerted efforts in the fields of machine learning and deep learning focusing on one key ingredient: data. Earlier techniques required not only a large amount of annotated data for the task at hand, but also an explicit and task-specific design of the features to be used. The advent of neural models removed the need for this feature engineering, as these models could learn directly how to extract relevant information from data, and spurred the proliferation of different model architectures, but the need for large amounts of annotated data remained.

A few years ago, however, the introduction of transformer models brought about a new paradigm in NLP: pre-training and fine-tuning. With this approach, vast amounts of unannotated and task-agnostic text are used to train massive language models that encode a wealth of linguistic and world knowledge, and can then be reused for any task by fine-tuning them with supervised data. While this has greatly reduced the need for task-specific data, large amounts of annotated data are still required. Meanwhile, the demand for NLP applications in the industry continues to grow, and the need to apply deep learning approaches to new domains, tasks, and languages, where training data is often scarce or non-existent, has become critical. In recent years, this problem has spawned a field of research aimed at overcoming the difficulties of learning —generalizing— from a limited number of examples: zero- and few-shot learning.

Several approaches have been recently proposed to address the challenge of data scarcity in NLP. Some of them focus on maximizing the learning from the few training examples available, but they generally aim at finding the means to exploit the knowledge already contained in pre-trained LMs. One such novel and promising approach, hereafter referred to as the entailment-based approach, is to reformulate the target NLP task as a textual entailment (TE) task, also known as natural language inference (NLI). This involves classifying whether or not a given sentence, the hypothesis, is entailed by the meaning of another sentence, known as the premise. Once this conversion has been made, an entailment model is used to perform the inference task and the results are mapped back into the output format of the target task. The method has been studied for a wide range of NLP tasks and has been found to be highly effective. Its main advantages include providing a common framework for unifying different NLP tasks, and the ability to leverage large, general-purpose NLI datasets to train the entailment model used for zero- or few-shot settings.

However, the usefulness of the entailment-based approach in data-poor scenarios has been demonstrated primarily for tasks in English, where huge NLI datasets and powerful models are readily available, raising the question of the approach's dependence on these large resources —which seems paradoxical given its intended use in data-scarce scenarios. Our research aims to investigate the feasibility and potential improvements of the entailment-based approach for languages with fewer resources. Specifically, we focus on Catalan, a medium-resource language for which a limited NLI dataset is available, and we investigate a multi-class text classification task (TC), due to its similarity to other classification tasks already studied within the entailment-based framework. Additionally,

we experiment with multilingual pre-trained models to contribute to ongoing debates in the research community about the relationship between resource size and effectiveness, specifically: are multilingual and larger resources more effective than monolingual and fewer resources? This debate could be important to guide the industry's efforts in resource creation. Finally, our research looks at task transfer learning to study the potential improvement of the technique.

1.1 Research objectives and questions

Our **first research objective** is to evaluate the capabilities of the entailment-based approach for zero- and few-shot scenarios for TC in a moderately under-resourced language, Catalan, solely using monolingual resources: a Catalan pre-trained model trained on a Catalan NLI dataset.

Secondly, given that the use of exclusively Catalan monolingual resources imposes two notable constraints, namely the size of the data used to train the monolingual LM and a reduced NLI dataset, we ask ourselves about their implications by investigating the approach's application when using multilingual resources. For this purpose, we examine both the scenario where a multilingual pre-trained model is trained with a Catalan NLI dataset, and the scenario where a multilingual pre-trained model is trained with large multilingual NLI datasets, which, together with the baselines utilized, also provide a point of comparison for our first research objective. The questions are: to what extent does the size of the data matter in this approach? Are larger NLI datasets essential to achieve good performance, even if they are not in the target-task language? At this point, our research intersects with the ongoing debate surrounding the use of monolingual compared to multilingual resources. Notably, Armengol-Estapé et al. (2021) have examined the performance of a medium-sized Catalan language model against state-of-the-art multilingual models on an NLU Catalan benchmark, and have concluded the superiority of language-specific models within the pre-training and fine-tuning paradigm. Our work expands upon this investigation by examining the comparison of monolingual and multilingual resources in the context of the entailment-based approach.

Thirdly, we investigate the potential of a task transfer learning setup to enhance the approach and compare it to the standard procedure involving a generic NLI dataset. Could we improve the method by reusing data from a similar task, transforming it into the NLI format, and using it for the approach? What kind of NLI training provides the greatest benefit?

Our primary research goals have led us to structure our work into three main branches: zero-shot, few-shot, and task transfer learning experiments. Each branch investigates specific secondary objectives. In the zero-shot branch, we experiment with different premise-shortening settings and templates to generate hypotheses in the NLI format conversion, in order to gain a deeper understanding of the robustness of entailment models. In the few-shot and task transfer learning branches, we use different few-shot data regimes to study the model's ability to learn as more data becomes available, and we compare the entailment-based approach to another state-of-the-art method for data-scarce scenarios.

1.2 Structure of the master's thesis

This thesis begins with a comprehensive literature review in Section 2, covering zero- and few-shot learning and the prominent works that adopt an entailment-based approach. Section 3 introduces the core methodology issues, including the study design, an explanation of the approach, the entailment models and the target task explored. Afterwards, Section 4 reports all the experiments conducted, which are divided into two main subsections: zero-shot and few-shot experiments. Each contains a detailed description of the experimental setup and the results obtained. For the zero-shot experiments, we also include a brief examination of the performance fluctuations resulting from different premise-shortening configurations. Within the few-shot subsection, we also include the task transfer learning experiments. Finally, Section 5 summarizes the main findings and conclusions of the experiments by directly addressing the research objectives. The Appendices section at the end provides additional supplementary experiments and graphs.

2 Literature Review

This section provides the necessary background information on which this thesis is built. It starts by introducing the field of zero- and few-shot learning for NLP and its main techniques, wherein we frame the entailment-based approach. Then, we gradually dive into it by first presenting the entailment task and the most common NLI datasets and entailment model training techniques. Finally, we focus on the entailment-based approach for zero- and few-shot learning by providing an overview of related work, with an emphasis on classification tasks —which are targeted in this thesis—, as well as some other NLP tasks that have been addressed in the literature.

2.1 Zero- and few-shot learning

Zero- and few-shot learning (ZSL and FSL, respectively) surge to address a fundamental drawback of the current pre-training and fine-tuning paradigm: while transformer LMs brought task-agnostic architectures, containing general —yet indeterminate— linguistic knowledge, fine-tuning an LM introduces the need for large amounts of labeled data for each specific task (usually, at least from the magnitude of some thousands of examples), which can span a myriad of different domains and languages. This requirement is hard to meet in real scenarios, where existing data are often insufficient (sometimes even non-existent due to data privacy issues, for instance) to train usable supervised models, whose performance is usually excessively weak in these conditions (Schick and Schütze, 2021b), and the possibility of human annotation is too costly or time-expensive to consider. Additionally, according to Brown et al. (2020), the narrowness of the training data distribution and objective compared to the broadness of those seen in the pre-training may damage the generalization capabilities of the model, possibly leading to a greater tendency to rely on superficial correlations.

In this context, ZSL specifically surges to tackle the scenario where there is no data at all, while FSL assumes access to some small amount of it, whether it be one (also known as one-shot learning), two, eight, or more examples, with no clear consensus, nevertheless, on the exact numbers or conditions. The research on zero- and few-shot learning, however, transcends those purely practical objectives mentioned above with the underlying aim of breaking the "conceptual limitation in our current NLP techniques" (Brown et al., 2020), which is the aforementioned need of large amounts of data for each specific task, in search for a broader generalization capability, i.e. a true language understanding model that can directly apply its knowledge to any task or learn new tasks as humans do, from small explanations or a few examples.

ZSL and FSL have been approached from several perspectives, often sharing the idea of eliciting the capabilities of pre-trained transformer-based LMs by reformulating the final task in a different format. The following paragraphs intend to offer a broad overview of the most prominent approaches in the area with the purpose of framing the one investigated in this work.

1. **Prompt-based methods.** Prompt-based methods for ZSL and FSL consist on prompting a pre-trained LM with a specific input —using a format with which it is more familiar due to the similarity with its pre-training objectives— to trigger the answers to the task. Works studying this approach are numerous, as prompt-based learning has emerged powerfully, even as a new paradigm —pretrain, prompt, predict— able to replace the pre-training and fine-tuning one (Liu et al., 2021). The performance of prompt-based methods is typically highly dependent on the fortunate selection of prompts, which turns prompt design into a key issue. Prompts are often manually designed, which requires prompt engineering and not only does it take time and sometimes even domain expertise to achieve good results, but the effort may not necessarily lead to optimal prompts. To address this issue, some works have focused on automatically searching (Shin et al., 2020) or generating prompts (Gao et al., 2021b; Jiang et al., 2020). Also, prompts can be discrete (in natural language) or continuous (Liu et al., 2021; Li and Liang, 2021). Prompt-based approaches have been used for diverse NLP tasks, among others, for knowledge and linguistic probing (Schick and Schütze, 2020; Petroni et al., 2019), classification tasks (Puri and Catanzaro, 2019; Gao et al., 2021b), information retrieval (Chen et al., 2021; Cui et al., 2021), and text generation tasks (Radford et al., 2019; Schick and Schütze, 2021b).

   (a) **Prompting a generative LM with task descriptions and demonstrations.** In the history of recent ZSL and FSL methods in NLP, Brown et al. (2020), GPT-3's creators, are often acknowledged for stunning the research community with the powerful zero- and few-shot capabilities of huge pre-trained LMs. GPT-3, with 175 billion parameters, achieved near-SOTA results on several NLP tasks by providing a task description, a few labeled examples (called demonstrations), and a prompt (often just the unlabeled example) as input to the model, without gradient updates, and letting it complete. For ZSL, only the task description and the prompt are fed to the model. However, Schick and Schütze (2021a) stress two crucial drawbacks of this method, often referred to as in-context learning: its reliance on huge-sized LMs is itself a limitation for its actual use, together with the carbon footprint associated, and the restriction in the number of few-shot examples that can be used, which is imposed by the maximum number of tokens accepted by the LM, seriously affecting the scalability of the method.

   (b) **Reusing the MLM objective of LMs for prompting.** These approaches are based on prompting LMs trained with the masked language modeling (MLM) objective by reformulating the final task as a cloze-question task. In few-shot settings, PET (Schick and Schütze, 2021b,a) and its iterative version, iPET, stands out for text classification and NLI tasks, requiring, however, apart from a few labeled examples, a considerable amount of unlabeled data. In PET (Schick and Schütze, 2021b), a few training examples are reformulated as cloze-phrases using various patterns and verbalizers (mappings from the label to tokens in the vocabulary that are used as possible options to fill the mask in the pattern), each

of which is used to train a separate LM. The resulting models are ensembled together in order to annotate unlabeled data with soft labels, which are finally used to fine-tune a standard classification model. By using different patterns for the ensemble model, PET handles one of the problems of prompt-based learning, which is its dependence on specific patterns used to prompt the LM. In Schick and Schütze (2021a), the authors introduce a modification to use PET in tasks that require predicting more than one token.

Some techniques built over PET are ADAPET (Tam et al., 2021), which removes the need for unlabeled data, and PERFECT (Mahabadi et al., 2022), which does not use handcrafted prompts and verbalizers. Another remarkable approach for few-shot prompt-based fine-tuning is LM-BFF (Gao et al., 2021b), which uses automatically generated prompts (using a generative model, T5) and adds task demonstrations. The author highlights some limitations of the method, which can probably be extended to —at least similar— prompt-based methods: in the first place, not all tasks can be naturally reformulated as cloze-questions; in the second and third place, it favors tasks with shorter input texts and with fewer output classes.

2. **Pivot tasks.** These approaches reformulate the final task into another NLP task that works as a bridge. Although some few works can be found that explore QA as a pivot task (Levy et al., 2017; McCann et al., 2018), NLI, which is the focus of the current work, has received wider attention for its promising results and demonstrated usefulness across diverse classification (Wang et al., 2021) and, more recently, information extraction (Sainz et al., 2021, 2022b,a) tasks, and has even been postulated as a "true language understanding task" (Wang et al., 2021). In the most common methodology, the input text from the original task is used as the premise in the NLI format, and the hypothesis is formed using one or more natural language patterns (or templates) that include a slot for a label description or verbalization, which maps to the possible output classes; an entailment model is then used to determine the probability of each premise and hypothesis pair bearing an entailment relationship, and this is used for the final mapping to an output class. Note that this approach shares some traits with some prompt-based learning methods; notably, the task reformulation requiring a pattern and some sort of label verbalization.

   Entailment-based learning has two main advantages: first, it naturally allows to leverage already trained entailment models for the downstream task, and can therefore be applied to zero-shot settings (Yin et al., 2019) as well as few-shot ones; second, it unifies classification tasks into a common task-agnostic formulation (as NLI), opening the possibility to reuse data from other tasks reformulated as NLI (Wang et al., 2021); third, in classification tasks, it allows to use the representation of the output classes in a more human-like way, instead of converting them to classification indexes. As this work is framed within the entailment-based approaches for zero- and few-shot, a more thorough explanation of it is provided later, in Section 2.1.

3. **Parameter-Efficient Fine-Tuning (PEFT) methods.** PEFT methods for few-shot settings are based on using the limited training examples available for the task to fine-tune only some of the model parameters (already existing model parameters or some added parameters), while the majority of them remain frozen. T-FEW (Liu et al., 2022) is a prominent and recent exponent of this approach, using the T0 model together with a novel technique for parameter-specific model fine-tuning (called $(IA)^3$) —which incorporates some vectors for modifying some of the model's activations— and introducing new loss terms. In the RAFT benchmark, T-FEW surpasses SOTA results and, for the first time, attains an over-human performance. Altogether, T-FEW accomplishes the significant feat of outperforming GPT-3 with fewer computational resources. According to Tunstall et al. (2022), however, despite the radical size decrease when compared to GPT-3, the large dimensions of T-FEW still make it hardly useful in real settings. PEFT methods, nevertheless, have also been used with smaller models, such as BitFit (Zaken et al., 2021). The fact that PEFT allows keeping most of the base LM unaltered, not only adjusted to a target task, and therefore its potential to share knowledge across tasks, makes it particularly interesting in multi-task learning settings, a hot current research direction in few-shot approaches (Mahabadi et al., 2021; Bansal et al., 2022).

4. **Sentence transformer-based methods.** SetFit (Tunstall et al., 2022) is the main recent work in this line of research for zero- and few-shot settings, which radically diverges from the previous methods for being based on sentence transformers (Reimers and Gurevych, 2019), an architecture that obtains sentence embeddings (that allow sentence similarity measurement) by using Siamese and triplet networks to modify pre-trained transformer LM models. In few-shot settings, SetFit first uses the few training examples available to create sentence pairs in a contrastive manner, i.e. for each class, similar (positive) examples are generated by choosing sentences from the same class, and, parallelly, not similar (negative) examples are created by choosing among out-of-class sentences. These contrastive pairs are used to fine-tune a sentence transformer (ST), which then encodes the original input sentences that —together with their respective output classes— will finally serve to train a classification head. At inference time, the new inputs are encoded by the trained ST and then passed to the classification head to output the class prediction.

   SetFit training and inference is fast and does not require prompts, and yet manages to position itself among the top-ranked SOTA methods in diverse classification tasks, and even reaches the first place in some of them. In zero-shot, SetFit can also be used by generating a small synthetic training dataset out of a natural language pattern with a gap that is replaced by each of the output class labels[1]. Note that SetFit shares with entailment-based methods the underlying idea of comparing the relation (similarity in one case and entailment in the other) between input pairs

[1]This application in zero-shot scenarios is proposed in the GitHub repository, available at `https://github.com/huggingface/setfit`

(which comprise two input sentences in SetFit, and an input sentence together with a generated hypothesis in entailment methods), as well as the contrastive data (example pairs) augmentation using positive and negative original examples.

5. **Other methods.** Other methods for ZSL and FSL include data augmentation techniques to resize the few-shot training set (Xie et al., 2019), and the recent STraTA method (Vu et al., 2021), which proposes a combination of task augmentation (generating data out of unlabeled target task examples for fine-tuning a model on an auxiliary task, such as NLI) and self-training. In a different line, some approaches focus on applying the knowledge learned from related tasks, such as the model-agnostic meta-learning algorithm (MAML) (Dou et al., 2019), or using supplementary training on intermediate tasks, such as STILTs (Phang et al., 2018). Additionally, there are works centered on improving training strategies, such as optimization and regularization techniques, to handle the challenges of standard fine-tuning with few examples (Lee et al., 2019; Zhang et al., 2020).

2.2 Textual Entailment

The entailment-based approach, so far framed within the zero- and few-shot research area, builds on the textual entailment (TE) task, which needs to be described before delving deeper into our main topic. In this subsection, we introduce the task, the primary TE datasets, and the main approaches for training entailment models.

2.2.1 The Textual Entailment task

Textual Entailment is the task of deciding whether the meaning of one sentence (hypothesis) follows, i.e. can be inferred, from the meaning of another one (premise). Currently, TE, or —in its complete form— Recognizing Textual Entailment (RTE), is frequently used interchangeably with the term Natural Language Inference (NLI), although the latter strictly refers to broader tasks and is employed in this wider sense in earlier literature (Poliak, 2020). TE surged as a common framework for evaluating systems on the task of recognizing semantic inferences, which is, in essence, the ability to appropriately associate meaning and written language, taking into account its inherent variability and ambiguity. Semantic inference has been claimed to be a crucial skill in linguistic comprehension and production, a core element in Natural Language Understanding (NLU) that is aimed for in NLP systems (Dagan et al., 2013). It is worth noting that, even when the term "entailment" is used to refer to the task, it does not correspond to the definition of logical entailment; the task is rather defined by empirical means, driven by individual's criteria (Dagan et al., 2006). Therefore, for instance, some entailment relationships may rely on general knowledge that is not explicitly stated in the text example.

Given that inference is an essential skill in numerous NLP tasks, the applications of NLI are broad and go beyond the evaluation purposes for which it was first proposed. With regard to evaluation, however, it is worth mentioning that NLI has been used both as a

generic evaluation for NLP systems (Poliak, 2020) and as a specific evaluation measure for certain tasks, such as abstractive summarization (Bora-Kathariya and Haribhakta, 2018). Furthermore, NLI has been used to approach some NLP tasks, often as components of more complex pipelines, such as those that imply checking text validity or consistency against their reference texts, like inconsistency detection in summarization (Laban et al., 2022) or fact-checking (Gao et al., 2021a), and subtasks of QA systems (Paramasivam and Nirmala, 2022). More recently, NLI has been proposed as a unified framework for directly modeling some NLP tasks, an approach that has proven to be very useful in scenarios with little data, as will be explored in Section 2.3.

2.2.2 Textual Entailment datasets

TE (or, most commonly, NLI) datasets have been rising in popularity for evaluation purposes, and later, as mentioned, as approaches to solve NLP tasks. One of the earliest such datasets was FraCas (Cooper et al., 1996), designed to cover a set of semantic phenomena; however, the fact that only contains about 1,000 labeled examples drastically limits its usability to training models. In contrast, the most commonly used NLI datasets today, SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018), respectively contain 570k and 433K premise-hypothesis pairs labeled as entailment, contradiction, and neutral. For each premise, the three kinds of hypotheses were manually written by crowd-source workers. SNLI premises are image captions, and MNLI gathers written and oral data from ten different genres. On its part, the ANLI dataset (Nie et al., 2020) was created to provide particularly difficult examples for current models via adversarial training based on an iterative process involving human and model feedback. Other NLI datasets have been created as reformulations of specific NLP tasks, and some of them are later mentioned, such as the foundational RTE datasets (Dagan et al., 2006; Bar-Haim et al., 2006; Giampiccolo et al., 2007, 2008; Bentivogli et al., 2009a,b, 2011), on account of their stronger link to the literature covered there.

For languages other than English, it merits mentioning the XNLI (Conneau et al., 2018) multilingual NLI dataset, which collects 7,500 NLI pairs from the MNLI corpus translated into 14 languages[2]. In fact, monolingual NLI datasets in other languages are often automated or human translations of English datasets, such as WNLI-es and WNLI-ca[3], the Spanish and Catalan translations, respectively, of the English WNLI dataset (Wang et al., 2018) (composed of 855 highly ambiguous NLI pairs), KorNLI (Ham et al., 2020), a Korean translation of the SNLI, MNLI and XNLI datasets, or AmericasNLI (Ebrahimi et al., 2022), which is again a translation of a subset of the XNLI dataset into 10 indigenous languages of the Americas, extremely low-resourced languages. NLI datasets originally created for a non-English language can also be found, but they tend to be considerably

[2]The 14 languages are the following: French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili and Urdu. With 7,500 NLI pairs for each of these languages and for the original English version, the dataset sums a total of 112.5k pairs.

[3]Available at `https://huggingface.co/datasets/PlanTL-GOB-ES/wnli-es` and `https://huggingface.co/datasets/projecte-aina/wnli-ca`

smaller than those available for English; for example, ASSIN (Fonseca et al., 2016), for Portuguese, consists of 10,000 NLI pairs, and InferES (Kovatchev and Taulé, 2022), for Spanish, has 8,055 NLI pairs. For Catalan, we find Teca[4], with a total of 21,163 NLI pairs using manually-written hypotheses. More details about this dataset are presented in Section 3.3.

2.2.3 Textual Entailment models

Prior to the rise of deep learning in NLP, multiple techniques were explored for building NLI systems. Some early methods include those based on logic and theorem proving (Bos and Markert, 2005), on word overlap (Jijkoun and de Rijke, 2005), on explicit syntactic or semantic knowledge (Pakray et al., 2010; Tatu and Moldovan, 2005), or machine learning (Zanzotto et al., 2009), among many others.

With the emergence of large-scale NLI datasets, deep learning approaches have been able to outperform other methods for developing entailment models. According to Coet (2019), the three main deep learning approaches to training entailment models are, firstly, sentence vector-based methods, also known as Siamese architecture, where the premise and hypothesis are independently encoded and passed to a classification layer that predicts the relationship between them; secondly, sentence matching approaches, that directly model the relationship between premise and hypothesis, usually through attention mechanisms; thirdly, transfer learning techniques, in which a model that has already been trained on a task is fine-tuned to perform the NLI task. Nowadays, state-of-the-art entailment models typically build upon the third approach; in particular, a transformer-based pre-trained model is fine-tuned on large NLI datasets such as SNLI or MNLI, as in other classification tasks. To train Masked Language Models like BERT on NLI data, the premise and hypothesis are concatenated with a separator token in the middle and a classification token at the start of the sequence (Devlin et al., 2018).

2.3 Entailment approach for zero- and few-shot learning

In this subsection, we review the main literature on the entailment-based approach for ZSL and FSL. We first focus on studies dealing with classification tasks, which best fits the context of the current work, and then briefly review those dealing with other NLP tasks.

2.3.1 Classification tasks

One of the pioneering works about entailment-based learning, focused on few-shot settings, is the one by Wang et al. (2021), on which the present work is greatly based. They call their proposed framework EFL (short for "Entailment as Few-shot Learner"), which consists in transforming any classification task into an entailment task. EFL is defined both for binary and multi-class classification tasks for one-sentence input tasks. In the first case, a sentence is chosen as a label description for one of the two classes (e.g., the positive

[4]Available at `https://huggingface.co/datasets/projecte-aina/teca`

class). Each input example is then rephrased in the form of a sentence-pair NLI example (x_i) as $x_i = [CLS]S_1[SEP]S_2[EOS]$, where S1 (the premise) is the input example, S2 (the hypothesis) is the label description, and CLS, SEP and EOS are the classification, separator, and end-of-sequence tokens, respectively.

Each input sentence from the original classification task forms an entailment example if the label description corresponds to its true output class and a non-entailment example otherwise. For the NLI few-shot training, the available examples are reformulated in this way and are used to train the entailment model. At inference time, each new classification example needs to be recast to the NLI format and passed to the entailment model, which outputs whether S2 is entailed by S1 or not. Finally, the model's prediction is mapped to the corresponding output class in the classification task, i.e. if the label description refers to the positive class, the model's prediction of entailment would point towards it, and towards the negative class if the model's prediction is not-entailment. The diagram in Figure 1 shows the binary classification at inference time.

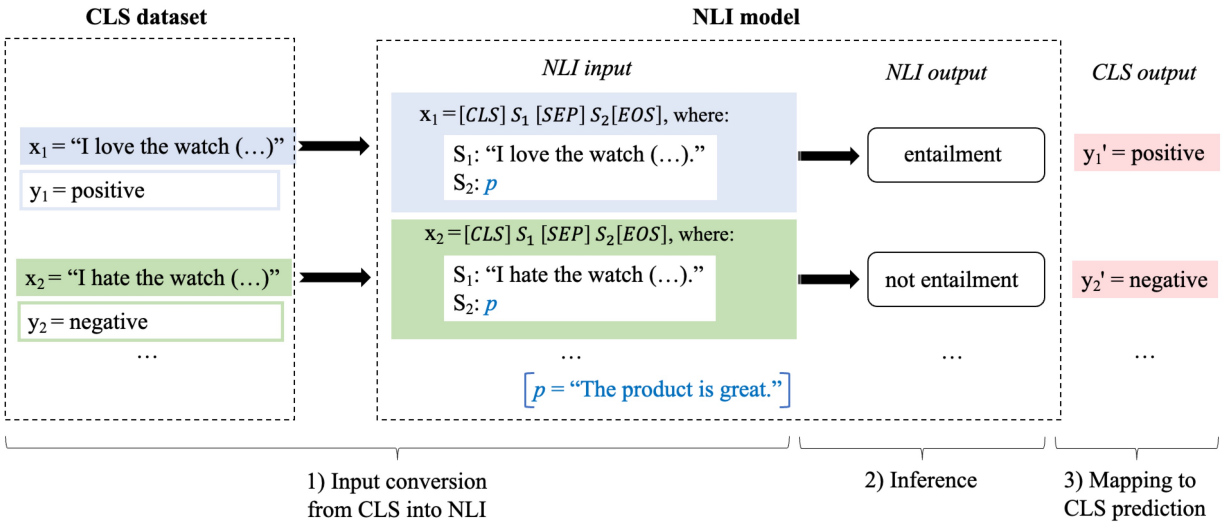


Figure 1: Steps involved in the EFL approach for binary sentiment classification. CLS stands for classification.

In multi-class classification, a label description is needed for each class. Each input sentence from the original classification task will be combined with each label description to form NLI pairs. That is, the number of NLI pairs formed from each original sentence will be equal to the number of classes in the classification task, and only one of them will be an entailment example. At training and inference time, all the possible NLI pairs are used in EFL.

Wang et al. (2021) propose two versions of EFL: in the first one, EFL wo PT, a pre-trained LM is directly fine-tuned on the NLI task with the few available training examples reformulated as entailment; in the second one, EFL, an entailment model is first trained on a general NLI dataset (MNLI) and then fine-tuned on the few-shot training examples

from the downstream task. With 8 examples per class, EFL outperforms the standard fine-tuning (using the same amount of training examples) and the other few-shot techniques considered (majority, LM-BFF, and STILTS) in 15 NLP tasks, with an average 8.2% improvements over them. EFL works significantly better than EFL wo PT in the 8 examples per class setting, but the MNLI pre-training becomes less important when the number of training examples increases to 16 and 32. Regarding the scaling ability of EFL, they demonstrate that the improvements obtained are bigger (compared to the baselines) when the training data size is smaller (8 examples per class). Also, they show that EFL performance benefits from larger pre-trained LMs.

Additionally, Wang et al. (2021) investigate the application of EFL in multilingual settings and unsupervised contrastive data augmentation for the entailment model based on word and span deletions, word reordering, and word substitutions, which brings some improvements to EFL performance. In multilingual settings (where the target task language is multilingual), they also outperform the other baselines by using a multilingual pre-trained LM, XLM-RoBERTa (Conneau et al., 2020), and a multilingual general NLI dataset, XNLI (Conneau et al., 2018), for the training of the entailment model before the few-shot fine-tuning of it. They keep the label description in English and experiment with both translate test method (where the training dataset is in English and the test set is also translated into it before the inference) and translate train method (in which the English training examples are translated into the target languages and added to the training set), and observe similar improvements.

Another study that has a strong connection to our own is the one presented by Yin et al. (2019) for its focus on text classification tasks using the entailment-based approach. They specifically deal with zero-shot scenarios, whose difficulty lies in the diversity of domains (for instance, news articles, reviews, medical records, etc.), aspects (emotions, situations, etc.), and the nature of the labels (they may change in time, be coarse or fine-grained, etc.) that exist in text classification tasks. In addition to proposing the entailment-based method for the task, they point out the dispersion of the literature in this area and therefore present a benchmark composed of datasets from three aspects (topic categorization, emotion and situation detection), and two setups for evaluation: the label-partially-unseen (during training, some labels are used, and the whole label space is used for testing) and label-fully-unseen scenarios (without target task data for training).

They argue that the entailment-based approach offers significant advantages that distinguish it from the standard (supervised) classification formulation: firstly, it allows using a single entailment model in any type of classification problem without the need to specify the number of output classes and using the label names; and secondly, it enables leveraging the output class label names for the task (instead of converting them to mere output indexes). In their experiments, they convert the test sets from each of the classification datasets in their benchmark to the NLI format and pass it as input to three entailment models trained using a pre-trained LM (BERT) with three distinct NLI datasets in English. To cast the classification data as NLI, they use three different templates (referred to as "interpretations") to form the hypothesis, adapting to the particular aspect of the downstream classification task (for example, "this text expresses [X]" for the emotion dataset).

To fill the template gap, they experiment with the label name and with a label description from WordNet (Miller, 1994). As in Wang et al. (2021), at inference time, for each text to classify, the model receives as many premise and hypothesis pairs as there are labels in the task. Then, the maximum entailment probability across all pairs will point toward the predicted class.

Compared to their three main baselines, which are the majority class distribution, the Explicit Semantic Analysis (a dataless classifier that uses Wikipedia to measure the semantic relatedness) (Chang et al., 2008), and Word2Vec (Mikolov et al., 2013) (which uses the cosine similarity between the representations of the text and the labels), the entailment models achieve the strongest results in all of the three classification tasks for the unseen labels of the label-partially-unseen evaluation and for all the labels in the label-fully-unseen evaluation. For the first setting, an additional and more challenging baseline used is an entailment model fine-tuned with the training data (those labels seen) reformulated as NLI, which clearly underperforms the general entailment models in the evaluation of the unseen labels. In the second setting, they also use a challenging baseline consisting of a binary BERT classifier trained on pairs of Wikipedia articles with their corresponding categories, which attains better performance than the general entailment models in the topic classification task. The three entailment models, however, have more robust results across all the classification datasets.

Also, out of the three entailment models, the one trained with a RTE dataset gets overall better results in the first setup, while MNLI and the ensemble of the three (only considered in this setting) in the second, but the interpretation of this phenomenon is left to future work. Regarding the use of label names or WordNet definitions to fill the template gaps, the second always performs worse, but a combination of both is useful in some cases (depending on the NLI dataset used and the classification task).

The promising results obtained by Yin et al. (2019) for zero-shot text classification, however, are further analyzed in a later work (Ma et al., 2021) and some issues with the method are brought to light. The first question they consider is how much NLI data contributed to the results. To investigate the subject, they use the Next Sentence Prediction (NSP) objective of a BERT model without any fine-tuning as a baseline; specifically, they employ as input the same premises and hypotheses (with the same templates and strategy to fill the gaps) used in the entailment-based approach, as well as these premise and hypotheses in reversed order (i.e. the hypothesis followed by the premise). Surprisingly, the NSP baseline overperforms the entailment models in most of the datasets evaluated, which suggests that the pre-trained model already contains the needed abilities for the task. They conclude that the influence of limited NLI datasets could even narrow the semantic coherence of the pre-trained model, which often contains relevant lexical biases that possibly result from relying on superficial lexical co-occurrences, which some other studies have also warned about (Sinha et al., 2021). However, the most abstract tasks, mainly emotion and situation detection, seem to obtain some slight benefit from the NLI datasets, which is however not very significant, according to the authors.

Sinha et al. (2021) also question the stability of entailment models. They mention the high variation across entailment models trained with different NLI datasets, but mainly

emphasize the differing nature between the NLI task and the target classification task. To study this issue, they train various entailment models on the same NLI dataset (MNLI) using different hyperparameters and keep the models with high and very similar results in the NLI task. By testing these different runs over the classification task, they report very strong variation across them, with absolute differences between the worst and the best result of about 23 points. This points to a reduced capability for generalization in out-of-distribution scenarios.

Finally, they explore whether a more robust entailment model would be helpful. First, they randomly shuffle the tokens in the input and check that the performance is not significantly affected, which could be explained, as mentioned, by a high reliance on superficial lexical patterns. Then, they experiment with three techniques on the MNLI data that have been proposed in the literature to improve the generalization capabilities of entailment models. Overall, the techniques do not significantly change the results, which seems to demonstrate that not only word patterns, but broader linguistic capabilities, are used by entailment models; it also suggests that more robust models would not be particularly beneficial to the entailment-based approach. In general, as has been made clear, the conclusions are not very hopeful regarding the method.

Other works that address classification tasks from the entailment-based approach are the ones from Sainz and Rigau (2021) and Obamuyide and Vlachos (2018), both focused on the zero-shot setting. Obamuyide and Vlachos (2018) deal with the relation classification task, in which the existence or absence of some specific relation type between two tokens or spans (for instance, "city of birth", "child", etc.) needs to be determined. The premise in this case is a sentence mentioning the two entities considered in the relation, and the hypothesis is the relation description. As a base model, however, they use Enhanced Sequential Inference Model (Chen et al., 2017), based on the BiLSTM architecture, which is commonly replaced by the Transformer one in more recent works. They use the label-partially-unseen approach, as defined in Yin et al. (2019), and also experiment with a general NLI dataset (MNLI) as a source of supervision and with a combination of both training regimes. The general NLI dataset already achieves reasonable performances, but the results obtained with each training dataset seem to vary across different evaluation datasets.

Sainz and Rigau (2021), at its turn, perform domain labeling of WordNet synsets (a multi-class classification problem) from 3 different approaches based on leveraging the knowledge already contained in pre-trained LMs: first, through the Masked Language Modeling objective (where some patterns with a fill-in mask are used to prompt the model, without predefining or restricting the possible output tokens); second, via Next Sentence Prediction, as used in Yin et al. (2019); third, by using an LM trained with a general NLI dataset, MNLI. Note that the three approaches use a formulation that needs two sentences as input, the second one of which is generated through templates. Their experiments show that the NLI-based method greatly outperforms the NSP-based one (the MLM does not allow for comparison given the free form of its predictions) and the two systems that previously attained SOTA results. Besides, through experimentation with different templates, they find that very short ones perform worse.

2.3.2 Other NLP tasks

In addition to classification tasks (including, among others, topic, sentiment and polarity classification, yes/no QA, etc.), the entailment-based approach been used to address other NLP tasks. Yin et al. (2020), for instance, recast the tasks of **QA** (specifically, the task consists of short text stories with some questions associated and their corresponding multiple-choice answers) and **coreference resolution** of pronouns as textual entailment problems in a few-shot setting. In the first case, the text is used as the premise, and the hypothesis is generated by transforming the question with each candidate's answer into an affirmative sentence, being only "entailment" the one bearing the correct answer. In coreference resolution, the original sentence acts as the premise, and the hypothesis is formed by replacing the pronouns in the sentence with the candidate entities (which are provided in the dataset). The study, however, does not employ the entailment-based approach later presented in Wang et al. (2021), but a matching-based method that uses a cross-task nearest neighbor layer to learn from both a general NLI dataset and the training examples from the target task.

Information extraction (IR) tasks have also been recently addressed. Sainz et al. (2021) deal with the **relation extraction** (RE) task (using the TACRED dataset (Zhang et al., 2017)) by reformulating it as a TE task through handcrafted templates used for generating the hypothesis of each of the possible output relations. Specifically, they use more than one template for each possible output relation. In addition, to guarantee the feasibility of the technique in real-world settings, they limit manual labor per relation to a maximum of 15 minutes. Along with templates, since some may potentially correspond to multiple relations, some entity constraints are used as well. In the zero-shot scenario, they directly use an entailment model, trained with a general NLI dataset (MNLI), to determine which of all the possible hypotheses generated for each example instance has the maximum entailment probability.

In the few-shot scenario, the available data (1, 5, and 10 percent of the training and development data) is reformulated into NLI by creating, in addition to the entailment pair (with the verbalization corresponding to the correct relation), one neutral pair and one contradiction pair by randomly sampling one of the incorrect relations and using one of its corresponding templates. This is due to the fact that MNLI uses the three options as output classes. The results are indeed promising: the few-shot entailment models outperform other SOTA systems used for the task with the same amount of data, while, in zero-shot, the performances are comparable to those of supervised training with 10% of the data. Additionally, the full-shot scenario demonstrates that the method is able to keep learning when more data is provided to a greater extent than other SOTA systems.

In a later work, a more complex information extraction task is explored, **Event Argument Extraction** (Sainz et al., 2022a), obtaining results that are on par with the current best in ACE and WikiEvents datasets in zero- and few-shot settings. Similarly, as in RE, each event argument is verbalized using templates (as before, with a maximum of 15 minutes per argument) to form the hypothesis, and the one obtaining the highest entailment probability among those that meet some particular type constraints is mapped to the pre-

diction. Their research also studies multi-source learning, where the training data of the different IR tasks are sequentially used to train the entailment model used to solve the final task, and demonstrate the effectiveness of the approach, a noteworthy achievement in a field where different annotation schemes make transfer learning challenging. Based on their experiments, they show that the brief time spent writing templates yields better results than the same time spent annotating. Additionally, they demonstrate that using several entailment datasets (MNLI, FEVER, SNLI, and ANLI) to train the entailment model significantly boosts results in zero-, few-, and full-shot settings.

These findings serve as a foundation for a later work (Sainz et al., 2022b), where a new annotation workflow for IE tasks is presented to replace the current one (in which a scheme is defined and then the annotation is performed) consisting of creating templates for verbalizations interactively by testing their effectiveness in zero-shot during their design. The authors also present a toolkit, ZS4IE, that makes this possible, which consists of a candidate generation module that varies according to the task (NER relies on POS for the candidate selection, RE relies on NER, etc.), a label verbalization module —that uses the verbalization templates written by the specialist as well as the candidates to fill in the template—, and an entailment model used for the inference.

Several works recognize the entailment task as a common NLP framework, usually with the aim of evaluating the reasoning capabilities of models, through the creation of several datasets that are themselves reformulations of other tasks or linguistic resources. In Dagan et al. (2006), the first RTE PASCAL challenge dataset was created as a benchmark with the purpose of evaluating the semantic ability of systems (the ability to detect that, from various texts using different words, the same meaning can be inferred) needed across NLP tasks. For this purpose, they manually annotated text pairs corresponding to seven different NLP tasks (IR, QA, reading comprehension, etc.).

The RTE challenges inspired further works. White et al. (2017) automatically recast the task of semantic role labeling, anaphora resolution, and paraphrase detection, to the NLI format. At its turn, Khot et al. (2018) and Demszky et al. (2018) automatically convert a QA dataset into NLI, which, according to the authors of the second reference, has the advantage of using multi-sentence reasoning and makes the task particularly difficult for machines. Finally, Poliak et al. (2018) rephrase 13 NLP datasets covering 7 different semantic phenomena into NLI; the authors stress that, unlike human-generated NLI datasets where the annotators freely create hypotheses, NLI datasets created from other resources incorporate various types of inference capabilities which can be identified based on the source task recasted.

3 Methodology

This section introduces the basic elements that compose the methodology used throughout this thesis. After outlining the basic experimental design, we summarize how the entailment-based approach works for text classification, providing real examples from the TC dataset used. Then, we briefly explain the classification task addressed in the experiments and the Catalan NLI dataset, and describe the main features of the entailment models used in the approach. Finally, we refer to the main metrics used in the experiments.

Our main research questions, stated in Section 1.1, have determined the backbone of our experimental design to be divided into three main branches with the common goal of solving the Catalan TC task: zero-shot, few-shot, and task transfer learning. The first two branches correspond to the names of the two main experimental subsections and address the first and second research questions by comparing monolingual and multilingual entailment models, whose general scheme is shown in Table 1. Specifically, for our first research question, which is to test the suitability of the entailment-based approach in zero- and few-shot settings when using purely Catalan monolingual resources (substantially smaller than the English resources often used in the literature), we use model 1 in the table: a monolingual LM trained on a monolingual NLI dataset. For the second research question, where we ask about the performance of the approach when using multilingual resources, we use a multilingual pre-trained LM and explore two options, corresponding to model 2 and model 3 in the table, respectively: with training on the monolingual NLI dataset and with training on multilingual NLI datasets. This allows us to understand the influence of resource size (both the pre-trained LM and the NLI dataset used for training) and language specialization on the approach in both zero- and few-shot scenarios.

The third branch is presented within the few-shot experiments and relates to the third research question, where we seek potential improvements of the entailment approach through task transfer. For that purpose, we create Wikicorpus, a classification dataset in Catalan built from Wikipedia summaries, and convert it into NLI data to fine-tune the pre-trained LM. The detailed experimental setups for each branch are given in the corresponding subsections of Section 4, before the results are reported.

Entailment model	**Pre-trained LM**	**NLI dataset**
model 1	monolingual (ca)	monolingual (ca)
model 2	multilingual	monolingual (ca)
model 3		multilingual

Table 1: Scheme of the three entailment models used in our zero-shot and few-shot experiments according to the languages used in the LM pre-training and NLI dataset.

3.1 The entailment-based approach

Essentially, the entailment method for solving TC tasks consists of converting the task data into the TE format and feeding it into an entailment model to be solved. There is little variation in the main aspects of methodology found in related literature. However, given that we address a multi-class TC task, the main works on which we base ourselves are those of Wang et al. (2021) and Yin et al. (2019), which deal specifically with classification tasks in zero-shot and few-shot settings, respectively, and we also take terminological and methodological guidance from Sainz and Rigau (2021) and Sainz et al. (2021, 2022a).

Regardless of whether it is a zero- or a few-shot setting, the same process is performed at inference time which requires, as a first step, that the TC data is converted into the NLI format —note that the TC data consists of a text with its corresponding label (the output class), whereas the NLI data consists of a premise, its associated hypothesis, and the NLI output class (the "entailment" or "non-entailment" label[5]). The conversion at inference time is accomplished as follows: each TC example generates a number of premise-hypothesis pairs equal to the number of labels in the task, all using the same premise, i.e. the text from the TC task, but different hypotheses, each consisting of a sentence indicating that the text belongs to one of the possible labels. To create the hypothesis, two elements are needed: a template, which is a sentence containing a fill-in element for the label (for example, "This text is about {label}."), and a label verbalization, which is the mapping from the label to a word or description to be replaced in the template. Once the TC example has been converted into a set of NLI examples, the entailment model receives it as input and returns entailment and non-entailment probabilities for each. To obtain the output for the classification task, the NLI pair having the maximum entailment probability will be chosen, and the label verbalization used to form its hypothesis will be mapped to the original label to obtain the final prediction. The whole process is illustrated in Figure 2 with a real example from the Catalan TC dataset, TeCla[6].

Being then the same process used at inference time, the difference between zero- and few-shot settings refers to the amount of training data from the final TC task used to fine-tune the entailment model. While no training data from the target task is used in the case of zero-shot, in few-shot a limited amount of TC examples, reformulated as NLI examples, are used to fine-tune the entailment model. For the reformulation into NLI, each text from the available TC data and its gold label are used to create one entailment pair, and some or all of the remaining (incorrect) labels can be used to create non-entailment pairs. The ratio of non-entailment examples for each entailment example created is therefore a configuration decision, as it is the template and label verbalization used to recast the data.

[5]Instead of these two labels, some NLI datasets use a triple distinction as the output class: entailment, neutral and contradiction.

[6]Available at `https://huggingface.co/datasets/projecte-aina/tecla`

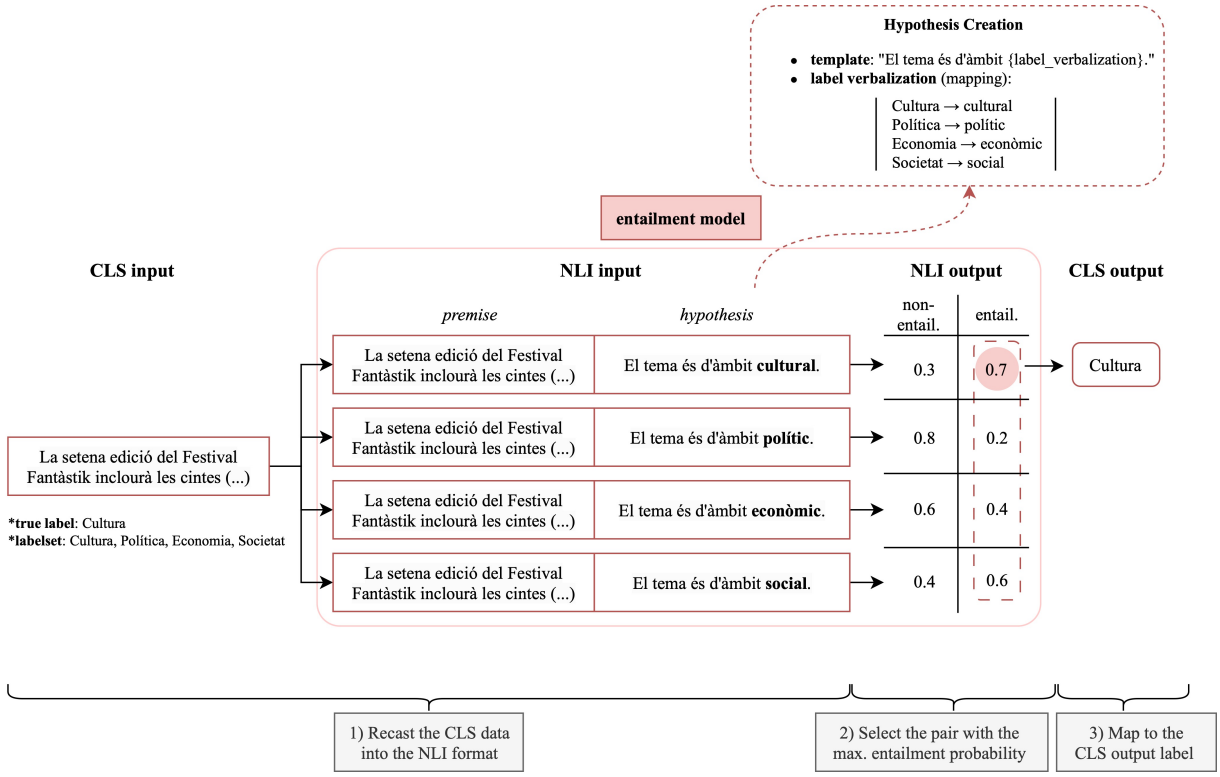


Figure 2: Steps involved in the entailment-based zero- and few-shot text classification at inference time. The example is directly extracted from the Catalan TC dataset used in this work (TeCla), and the template and label verbalizations are among the ones used in the experiments. CLS stands for classification.

3.2 TC dataset

This work focuses on multi-class text classification, a specific classification problem in which the inputs are texts (instead of words or single sentences), and the output classes are numerous and mutually exclusive labels[7]. The dataset used in all the experiments is TeCla (which stands for "Text Classification"), a free and publicly available dataset for the Catalan language —there are few alternatives of public, large text classification datasets in Catalan to date[8]. TeCla is a collection of news articles (title, subtitle and body merged into a compact text) from ACN (Agència Catalana de Notícies), a Catalan news agency, associated with categories corresponding to different news sections. The length of the texts are from one to a few paragraphs long, with an average of 484 tokens using the Catalan monolingual LM (RoBERTa-base-ca-v2) tokenizer and 577 tokens using the multilingual LM employed in this work (XLM-RoBERTa-base).

We used the second version of TeCla, developed and published as part of this work to

[7]This task differs from multi-label classification, where text can have multiple output labels.

[8]WikiCat, a text classification dataset obtained through a selective crawling of the Catalan Wikipedia, was published after the start of this work and is significantly smaller than TeCla.

achieve greater class exclusivity through stricter curation criteria that excluded examples simultaneously belonging to two categories, since the first version contained some potential class overlap. To avoid significantly reducing the size of the dataset due to the harsher filtering criteria, no examples were removed for failing to meet a minimum number per class, which inevitably leads to a more unbalanced dataset. In addition, unlike the previous version, the new version uses a hierarchical class structure to take advantage of both the section and subsection categorizations of each article (which were used in the original data). That is, each article is labeled with a coarse-grained class (called label 1 in the dataset) with 4 possible options, and a fine-grained class (label 2) with a total of 53 classes, limited to the options allowed for its coarse-grained class. This dual categorization naturally provides two levels of difficulty in the classification task.

The dataset is divided in a stratified manner into train, development and test splits, with 80%, 5%, and 15% of the total data, respectively. Table 2 shows the total number of examples for each split according to their coarse-grained category. The extremely unbalanced class distribution for the train set (which is proportionally equivalent between splits) is shown in Figure 3.

Coarse-grained labels	**Fine-grained labels**	**train**	**dev**	**test**
Cultura	Arts, Castells, Cinema, Equipaments i patrimoni, Festa i cultura popular, Gastronomia, Llengua, Lletres, Música, Teatre	11,921	746	2,233
Economia	Agroalimentació, Comerç, Comptes públics, Empresa, Energia, Finances, Habitatge, Hisenda, Indústria, Infraestructures, Innovació, Logística, Mobilitat, Moda, Noves tecnologies, Treball, Turisme, Urbanisme	16,305	1,018	3,059
Política	Entitats, Exteriors, Govern, Govern espanyol, Parlament, Partits, Política municipal, Unió Europea	25,568	1,599	4,794
Societat	Cooperació, Educació, Esports, Immigració, Judicial, Medi ambient, Memòria històrica, Meteorologia, Moviments socials, Policial, Recerca, Religió, Salut, Serveis Socials, Successos, Trànsit, Universitats	36,906	2,306	6,921
		90,700	5,669	17,007
			113,376	

Table 2: Coarse-grained classes in the TeCla dataset with their corresponding fine-grained ones and the number of per-class (coarse-grained) examples in each split.

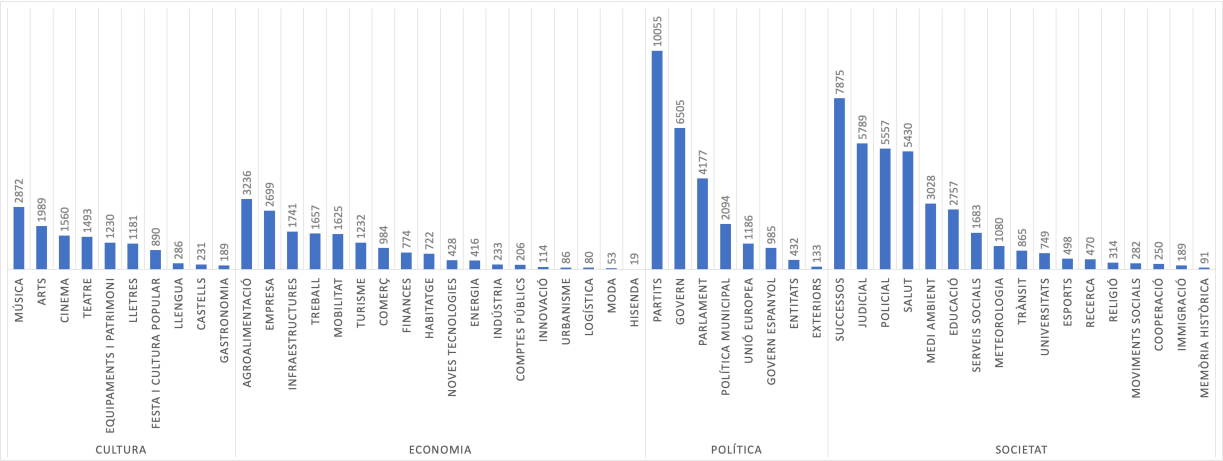


Figure 3: Distribution of the coarse-grained and fine-grained classes in the train partition of the TeCla dataset.

3.3 Catalan NLI dataset

Teca is the main Catalan NLI dataset existing to date[9], with a total of 21,163 premise-hypothesis pairs divided into training, development, and test partitions of 16,930, 2,116, and 2,117 examples each, respectively. The texts used as premises were either sentences from the Catalan textual corpus[10] or headers from a Catalan news site, VilaWeb, and have an average length of 20.6 tokens calculated with the RoBERTa-base-ca-v2 tokenizer. For each of them, one entailment, one neutral, and one contradiction hypothesis were manually written by annotators.

3.4 Entailment models

The entailment models used to fill the design specifications from Table 1 are shown in Table 3, along with a report of the main features of the NLI datasets used to fine-tune their pre-trained LM[11]. In Table 4 we present as well the two pre-trained LMs, a Catalan monolingual LM and a multilingual LM, which are both RoBERTa-base models that share a pre-training scheme and model architecture. For reference, in Table 5 we include the standard fine-tuning results of both LMs for both the coarse-grained TC and the fine-

[9]As mentioned in Section 2.2.2, there also exists WNLI-ca, the translation of the English WNLI dataset (from the GLUE benchmark), consisting of 855 NLI pairs.

[10]Available at `https://huggingface.co/datasets/projecte-aina/catalan_textual_corpus`

[11]The already fine-tuned models RoBERTa-ca-Teca (RoBERTA-base-ca-v2-te) and XLMR-SMAX (XLM-RoBERTa-base-SNLI-MNLI-ANLI-XNLI) can be found in the HuggingFace library, at `https://huggingface.co/projecte-aina/roberta-base-ca-v2-cased-te` and `https://huggingface.co/symanto/XLM-RoBERTa-base-snli-mnli-anli-xnli`, respectively. XLMR-Teca (XLM-RoBERTa-base-te), on the other hand, was fine-tuned for this work with a learning rate of 1e-5, 10 maximum epochs, and a batch size of 16. It achieved 79% accuracy in the NLI Teca dataset, while RoBERTa-ca-Teca reached 83% accuracy.

grained TC using the full training and development sets; the models were trained with a learning rate of 3e-5, 10 maximum epochs, and a batch size of 16, as this is the configuration used for the few-shot experiments.

Entailment model	**Pre-trained LM**	**NLI dataset**	**NLI dataset num. example pairs**
RoBERTa-ca-Teca	RoBERTa-base-ca-v2 monolingual (ca)	Teca monolingual (ca)	21,163
XLMR-Teca	XLM-RoBERTa-base multilingual	Teca monolingual (ca)	
XLMR-SMAX		SNLI, MNLI, ANLI, XNLI (14 languages in total, but mainly English) multilingual	570k+433k+ 169k+112k = 1284k

Table 3: Entailment models used in the zero- and few-shot experiments with their respective pre-trained LM and NLI dataset(s).

LM	**Train. data size**	**Languages**	**Num. params**
RoBERTa-base-ca-v2	34.9GB	Catalan	125M
XLM-RoBERTa-base	2.5TB (10.84GB of Catalan data)	100 languages (including Catalan)	270M

Table 4: Basic characteristics of the two pre-trained models (monolingual and multilingual) used in the experiments.

LM	**Coarse-grained labels**	**Fine-grained labels**
RoBERTa-base-ca-v2	96.3	80.3
XLM-RoBERTa-base	95.8	79.1

Table 5: Standard fine-tuning results (weighted F1) for the coarse-grained and fine-grained tasks of the TeCla dataset.

3.5 Metrics

The primary metric used in this work to compare the performance of the entailment models in the TC task, as well as to select the best checkpoint for training entailment models, is the weighted average F1, where the F1 is computed separately for each class and then averaged considering the class proportion. F1, the harmonic mean of precision and recall, is particularly valuable for summarizing the performance of a model when the class balance is uneven and/or when multiple classes intertwine. Its weighted version reflects the impact of the number of examples per class on the final result.

4 Findings

This section describes the experiments conducted to address the research objectives. Its two main subsections, for zero- and few-shot experiments, respectively, contain the detailed methodology followed, which builds on the basic outline already presented in Section 3, as well as a thorough exploration of the results obtained.

4.1 Zero-shot experiments

4.1.1 Experimental setup

The diagram in Figure 4 illustrates the scheme employed in the zero-shot experiments. Note that, for each of the three entailment models, the coarse-grained and fine-grained categorizations from the TeCla dataset are considered as separate tasks (as mentioned above, they offer two levels of difficulty, the first with only 4 coarse-grained labels being much easier than the second with 53 fine-grained labels), and each of them is explored in two settings: the first uses the whole original text from the TC dataset as a premise; the second uses the first sentence from the text as the premise, which corresponds to the title of the article. The motivating factor for these two settings, which aim to explore the influence of premise length, is the fact that NLI datasets typically use one-sentence premises. The numbers in red on the right refer to the set of templates and verbalizations used to transform the TC examples into NLI at inference time, collected in Table 6 (circle 1) and Table 7 (circle 2).

The twelve templates in Table 6 were generated by making slight linguistic modifications to the initial template (denoted as "original" in the right column), which corresponds to a generic template that has been often used in other works, and also incorporates some other template types retrieved from the literature, such as the bare label and the QA format. This set of templates allows exploring the robustness of the entailment models to small variations. The label verbalizations used with these templates are simply the lowercase versions of the label names (except for some proper nouns, whose capitalization has been maintained), since this is the most practical scenario, requiring less manual effort. However, the purpose of the second set of templates and label verbalizations, in Table 7, which are only used for the coarse-grained task because only those categories admit adjectivation, is precisely to investigate the effect of a different label verbalization (in particular, the adjectivized form of the labels), which imply different templates to properly fit them.

For the test set evaluation, the template and setting that yielded the best results for each model in the development set were used. The performance of entailment models is measured against the following three baselines:

- **Prompt-based approach.** A text with a masked token is input to a Masked Language Model (MLM) for it to fill the gap; since we need the model to output some defined classes, instead of letting it predict the most likely token from the whole vocabulary, we limit the output space so that it only returns the probabilities

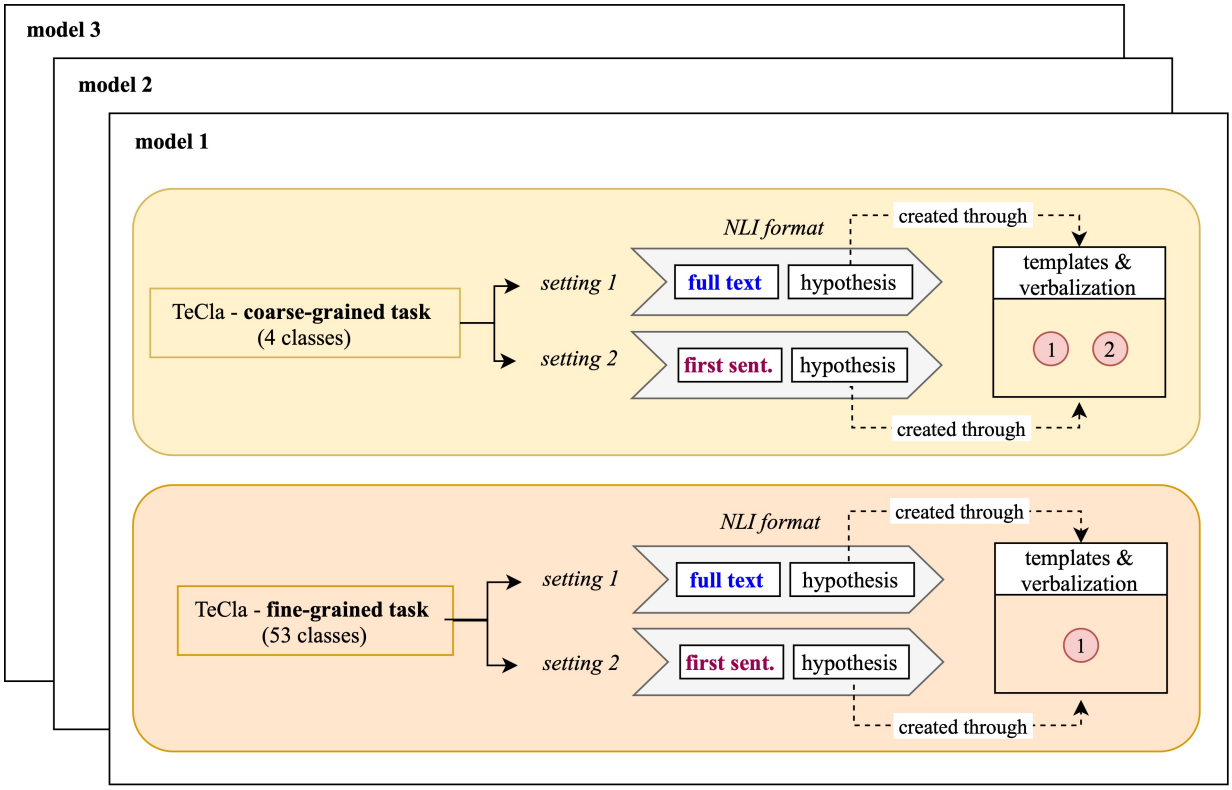


Figure 4: Structure of the zero-shot experiments' design. The circled numbers inside the "templates verbalization" box refer to the tables underneath.

Templates	1	Aquest text tracta sobre {label}.	*original*
	2	Aquest text va sobre {label}.	*verb change*
	3	Aquest text és sobre {label}.	
	4	Aquest text tracta de {label}.	*preposition change*
	5	El text tracta sobre {label}.	*article change*
	6	Aquest exemple tracta sobre {label}.	*noun change*
	7	Aquest article tracta sobre {label}.	
	8	Això tracta sobre {label}.	*noun phrase change*
	9	ø Tracta sobre {label}.	
	10	Aquest text tracta sobre {label}	*punctuation change*
	11	{label}	*only label*
	12	Pregunta: El text tracta sobre {label}? Resposta: Sí.	*QA form*
Label verbalization	Original label names, all lowercased except for 3 label names corresponding to proper nouns in the fine-grained task: "Unió Europea", "Parlament", "Govern".		

Table 6: Set of templates (1) with their corresponding label verbalization, applicable to coarse-grained and fine-grained tasks.

Templates	13	És un tema {label}.
	14	Aquest text tracta un tema {label}.
	15	El tema és de caire {label}.
	16	El tema és d'àmbit {label}.
	17	L'article és de caire {label}.
Label verbalization	Label names converted into their adjective form and lowercased, as in the following mapping (original: verbalization): - Cultura: cultural - Política: polític - Economia: econòmic - Societat: social	

Table 7: Set of templates (2) with their corresponding label verbalization, only applicable to coarse-grained categories.

for the specified labels[12]. Analogous to the NLI approach, each text example from the classification dataset (the premise in the NLI format) is concatenated with a template containing the masked token, and then passed to the LM as input. We used RoBERTa-base-ca-v2 as the pre-trained LM, and adopted the following procedure (analogous to that used for the entailment models) to select the best possible template and premise-shortening options: in the development set, we evaluated the same set of templates utilized for the entailment models by replacing the gap for the label with the masked token (for instance, "Aquest text tracta sobre $< mask >$."), and then chose the best combination for the test set evaluation, which was the full premise setup with template 16 (adj) and template 4 (N) for the coarse- and fine-grained tasks, respectively.

Compared to the entailment approach, the prompt-based method poses a significant challenge to dealing with multi-mask tokens as labels, since the MLM outputs the probability of each individual token in the next sentence position, and the joint probability of multiple consecutive tokens is always smaller and not directly comparable to that of a single-token prediction. Schick and Schütze (2021a) suggest calculating the probability of each token in its position, and then inserting the token with the highest probability into the gap and calculating the probability of the remaining token. However, in our experiments, this approach resulted in much higher probabilities for that label when compared to the single token labels. To address the problem, we additionally tried the following methodology: we converted all the tokens from multi-token labels into consecutive masks and calculated the probability of each one at its position (keeping the remaining tokens in the label masked). To determine the final probability for that label, we experimented, in the development set, with obtaining the mean, maximum, and minimum of those partial probabilities, and finally used the minimum in the test set, which was found to greatly outperform the other options.

[12]To implement this approach, we used the FillMask pipeline from the Hugging Face API, which includes the mentioned functionalities.

- **Majority.** This model assigns the most common label[13] from the training set to all examples.
- **Uniform.** A classifier that assigns classes randomly, with each class having an equal chance of being selected[14].

4.1.2 Main results

The three entailment models were evaluated in the TeCla development set according to the scheme presented in Figure 4. The summarized results across all templates for the two premise length settings (full premise and first sentence as premise) are shown in Table 8 and Table 9 for the coarse- and fine-grained tasks, respectively. Additional graphics can be found in Appendix A that provide more in-depth information. The results allow us to investigate three main issues: the performance comparison between our three entailment models, and the impact of the premise shortening and template variations in their performance.

Premise type	Entailment model	max w-F1	min w-F1	mean w-F1	mean m-F1	best template	worst template
full premise	RoBERTa-ca-Teca	58.7	26.8	42.9	43.7	temp. 16 (adj)	temp. 11 (N)
	XLMR-Teca	57.0	25.0	44.5	41.5	temp. 7 (N)	temp. 17 (adj)
	XLMR-SMAX	**71.8**	51.4	62.8	59.1	temp. 13 (adj)	temp. 11 (N)
first sentence	RoBERTa-ca-Teca	59.7	32.5	41.7	46.1	temp. 16 (adj)	temp. 9 (N)
	XLMR-Teca	66.0	42.6	57.3	55.0	temp. 7 (N)	temp. 14 (adj)
	XLMR-SMAX	67.0	52.0	57.4	54.8	temp. 13 (adj)	temp. 10 (N)

Table 8: Zero-shot development results for the full and first sentence as premise setups across templates in the coarse-grained task. The metrics used are the maximum, minimum, and mean weighted F1 score (w-F1), as well as the mean macro F1 (m-F1).

Premise type	Entailment model	max w-F1	min w-F1	mean w-F1	mean m-F1	best template	worst template
full premise	RoBERTa-ca-Teca	24.0	14.1	20.5	15.6	temp. 7 (N)	temp. 11 (N)
	XLMR-Teca	28.0	5.5	20.7	18.0	temp. 7 (N)	temp. 11 (N)
	XLMR-SMAX	31.6	16.6	28.8	24.3	temp. 8 (N)	temp. 11 (N)
first sentence	RoBERTa-ca-Teca	**36.1**	25.9	32.4	21.4	temp. 7 (N)	temp. 8 (N)
	XLMR-Teca	24.6	9.4	19.5	12.8	temp. 5 (N)	temp. 11 (N)
	XLMR-SMAX	24.8	8.5	21.8	17.4	temp. 3 (N)	temp. 11 (N)

Table 9: Zero-shot development results for the full and first sentence as premise setups across templates in the fine-grained task. The metrics used are the maximum, minimum, and mean weighted F1 score (w-F1), as well as the mean macro F1 (m-F1).

- **Which are the best-performing entailment models in the ZS setting?**

In the coarse-grained task, XLMR-SMAX obtains the best performances (both maximum and means across templates are the highest), and it specifically achieves those results in

[13]In the TeCla dataset, the class distribution is the same for the training, development and test sets.
[14]For the majority and uniform models, we used the DummyClassifier from the Sklearn library.

the full premise setting. If we consider the best result from each model, XLMR-Teca reaches the second-best performance, and RoBERTa-ca-Teca ranks the last, both in the first sentence premise setting. However, the distance with respect to the overall leading model is especially poignant in the latter, with respectively 5.5 and 21.1 absolute points of difference in the mean weighted F1, and 5.8 and 12.1 in the maximum weighted F1.

In the fine-grained task, the overall tendencies change substantially with respect to the coarse-grained. Firstly, the highest maximum and mean weighted F1 are attained by RoBERTa-ca-Teca under the first sentence setting, in marked contrast to its last place ranking in the coarse-grained task. In the second place, at a relatively close distance from it, XLMR-SMAX, in the full premise setting, is 4.5 and 3.6 absolute points beneath in the maximum and mean weighted F1, respectively, although it slightly improves the mean macro F1. Finally, XLMR-Teca, with the full premise, is 8.1 and 11.7 absolute points in the mentioned metrics far from the best model.

The above results allow for the following interpretation. In the coarse-grained task, which deals with a few generic labels, both the large size of the pre-trained model and the NLI training data seem to contribute decisively to improving the task performance, regardless of the small Catalan input seen. In short, the vastness of data seems to be playing the most crucial role in acquiring the general-domain inference abilities required for the task. However, the size is not as determinative in the fine-grained task, whose numerous narrow classes demand a more sophisticated discrimination ability. The specific language-related knowledge provided by the pre-trained monolingual model appears to be the most relevant factor (instead of the NLI training dataset) for RoBERTa-ca-Teca to outperform the other models, given that XLMR-Teca lags significantly behind it.

- **How does the shortening of the premise influence the performance of the entailment models?**

The full premise being richer in terms of informativeness than the first sentence, XLMR-SMAX has been found to possess the highest capacity to benefit from it in both coarse- and fine-grained tasks. In contrast, RoBERTa-ca-Teca is able to achieve higher results in the first sentence setting, which is particularly noteworthy for the fine-grained task, where the highest and mean weighted F1 surpasses the ones obtained for the full premise scenario by 12.1 and 11.9 absolute points, respectively. This apparent preference towards shorter premises is further analyzed in Section 4.1.3 through some additional experiments performed on this model using different premise shortenings. On its part, XLMR-Teca exhibits less consistency with regard to its premise-shortening preferences: it obtains better results in the first sentence setting for the coarse-grained task and in the full premise setting for the fine-grained task.

A deeper examination of XLMR-Teca's behavior in the coarse-grained task, however, reveals additional insights into the role of the pre-trained LM and NLI dataset that compose each model. In the full premise setting, the two models sharing the monolingual NLI training dataset (RoBERTa-ca-Teca and XLMR-Teca), obtain very similar results in the maximum F1 and means; parallelly, in the first sentence setting, the two XLM models are

the ones obtaining very similar results. One plausible explanation is the following: firstly, the monolingual NLI dataset seems to be responsible for limiting the model's capability in utilizing the whole textual premise when it is very long, probably due to a more pronounced change in the data distribution from the NLI training data and the testing data —since the monolingual NLI dataset only included one-sentence premises, while some of the NLI datasets in XLMR-SMAX include much longer premises.

Secondly, the fact that the entailment models trained with the XLM pre-trained model achieve more competitive results for the coarse-grained task supports the idea that the knowledge encoded in the pre-trained model is the major contributor to the entailment model's ability for the task, which aligns with the explanation for the superior performance of the monolingual model in the fine-grained task.

The tendencies identified in the coarse-grained task are still present in the fine-grained one, though to a lesser extent. This is because XLMR-Teca has more inconsistent performance across templates, resulting in higher maximum and lower minimum F1 scores in the full premise setting compared to the monolingual model.

- **How does the template variations affect the performance of the entailment models?**

The results reveal significant fluctuations in the model's performance based on the template utilized, a trend also observed in other studies referenced in the literature review. Besides, each entailment model exhibits different template preferences, with few coincidences between them. However, in the coarse-grained task, the best-working template for each model is consistent across both the first sentence and full premise settings, while this only happens for RoBERTa-ca-Teca in the fine-grained task. The best template that repeats the most in both tasks, specifically in RoBERTa-ca-Teca and XLMR-Teca, is "Aquest article tracta sobre {label}." (which could be translated as "This article is about {label}."). This template is a variation of the standard one, as can be checked in Table 6, which replaces "text", an unspecific textual reference to the hypothesis, with "article", a semantically more restricted noun that better fits the genre of the text used as context. Regarding the worst template, while there are variations between the full premise and first sentence setups within the same model in the coarse-grained task, there is almost a consensus across models and setups in the fine-grained task: the worst template, except in one case, is the bare label ("{label}").

The most inconsistent model across templates is XLMR-Teca, showing the highest standard deviations and the overall lowest performances in both tasks. Additionally, its performance drops dramatically with the adjectival labels (used only in the coarse-grained task)[15], reaching its worst F1 results, conversely to the other model's tendencies, which obtain their best results in the two premise shortening setups using one of those templates. In fact, the templates with adjectival labels lead to significant improvements for RoBERTa-ca-Teca and XLMR-SMAX of 14 and 6 absolute points in the two premise setups for the former, and 8 in each of the setups for the latter.

[15]This phenomenon is best appreciated in the graphics in appendices A.

As regards the linguistic variations in the templates (considered only in relation to the nominal labels), the dispersed and unpatterned results make it difficult to obtain straightforward conclusions. Some templates that work relatively well for one task are bad assets for the second one; for instance, template 8, with a pronoun in the place of the noun phrase ("Això tracta sobre {label}.", meaning "This is about {label}."), is one of the best for RoBERTa-ca-Teca in the coarse-grained task, but one of the worst in the fine-grained task. The premise setup also affects decisively each template's performance within each model, particularly in the coarse-grained task, where very similar templates (templates from 1 to 10) obtain differences in the results up to 15 absolute points in weighted F1.

		coarse-grained task		*fine-grained task*	
	Model	**F1 (weighted)**	**F1 (macro)**	**F1 (weighted)**	**F1 (macro)**
entailment models	RoBERTa-ca-Teca	59.7	62.1	**36.3**	26.0
	XLMR-Teca	63.9	61.5	27.0	23.5
	XLMR-SMAX	**71.1**	**69.2**	31.4	**26.3**
baselines	prompt-based	52.4	55.6	22.8	16.0
	majority	23.5	14.5	2.2	0.4
	uniform	25.5	23.2	2.3	1.4

Table 10: Zero-shot test results for the coarse-grained and the fine-grained tasks, comparing the performance of the entailment models against the baselines.

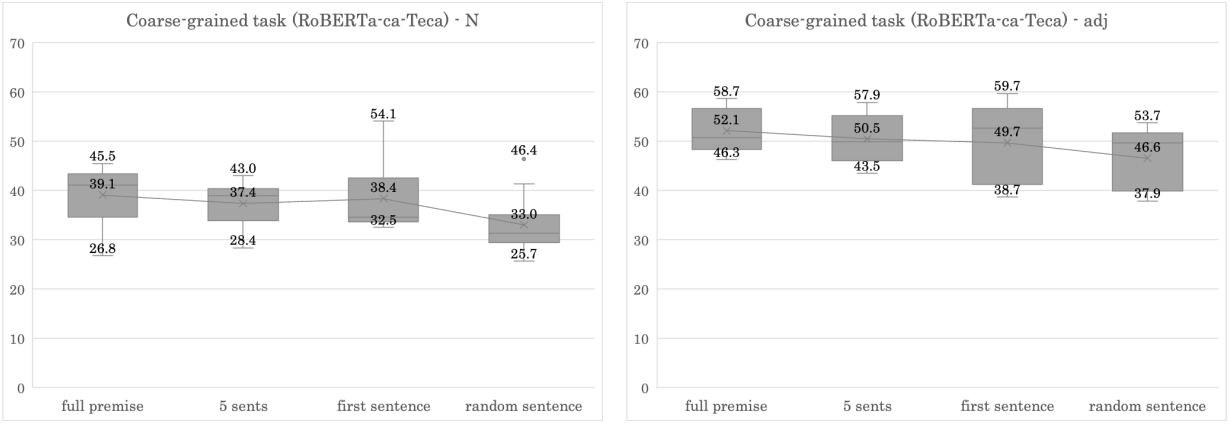
Table 10 summarizes the results (weighted and macro F1) obtained for the coarse-grained and the fine-grained tasks in the test set with the best entailment models from the development set[16] and the baselines. With respect to the development results, the entailment-based models achieve very similar results, probably due to it being of sufficient size and having the same class distribution as the test set. The results clearly show that the three entailment models significantly outperform the baselines in both the coarse-grained and fine-grained tasks. The best entailment model translates into a 35.7% and 59.2% of relative improvement in the weighted F1 with respect to the prompt-based baseline in the fine-grained task[17], and even the worst entailment model implies a 13.9% and 18.4% of relative improvement.

4.1.3 A further look at the premise shortening

In the development results, RoBERTa-ca-Teca model generally yielded better performances in the setting with the first sentence as premise, particularly in the fine-grained task. To

[16]For each entailment model, the best-working template and premise shortening setup in the development set is used in the test set evaluation.

[17]We must note that the prompt-based approach, despite the problems with multi-token labels referred to in the experimental setup, is able to obtain weighted F1 results comparable to the entailment-based approach. However, it has particularly low macro F1, which a further exploration revealed to be probably due to some bias related to the approach, since there are 15 classes for which the model did not output a single prediction. Among those, multi-token classes are not particularly penalized; instead, some of them are the most semantically broad label names ("Economia", "Arts", etc.) as well as the most unnatural or ungrammatical verbalizations (those which worse fit the template).



(a) (b)

Figure 5: Coarse-grained task results for the four premise-shortening experiments with RoBERTa-ca-Teca (a) using the template set with nominal labels presented in 6, and (b) using the template set with adjective labels presented in 7.

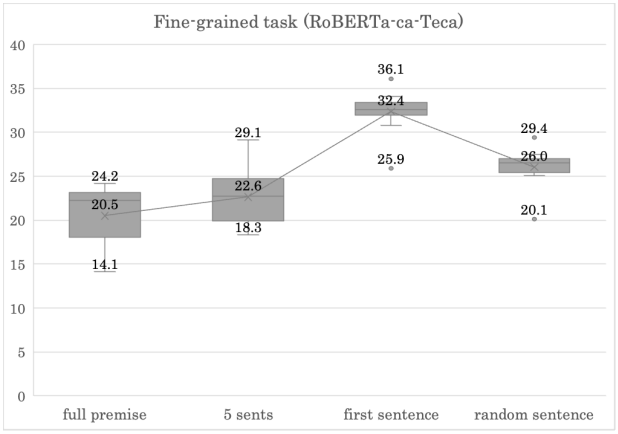


Figure 6: Fine-grained task results for the four premise-shortening experiments with RoBERTa-ca-Teca using the template set with nominal labels template presented in Table 6 (the adjective labels were not used for this task).

further investigate a suspected bias for shorter premise setups, we compare the F1 weighted performances obtained with this model for the whole set of templates presented in Table 6 in four different premise-shortening settings: 1) the full premise, 2) the first five sentences, 3) the first sentence, and 4) a random sentence taken from the text. The results for the coarse-grained and the fine-grained tasks are shown as boxplots in Figure 5 and Figure 6, respectively, where the upper, medium, and lower values of each box represent the maximum, mean, and minimum weighted F1 obtained for the set of templates.

In the coarse-grained task, the results from the template set with nominal and adjective

labels do not show identical trends across different premise-shortening setups, suggesting a relationship between specific premise lengths and templates, and the difficulty of drawing conclusions from each element in isolation. While in the template set with adjective labels mean scores decrease as the premise shortens (and further decrease from the first sentence premise to a random sentence used as the premise), in the template set with nominal labels, using the first sentence as a premise yields better mean scores than using the first five sentences. Additionally, in terms of maximum scores obtained, the first sentence always ranks first, but using a random sentence from the text surprisingly achieves the second-highest F1 score in the template set with nominal labels, suggesting that a preference for shorter texts holds when combined with certain templates. Interestingly, we found that, in the template set with nominal labels, the template that reaches the maximum score in the first sentence and random sentence setups is the bare label, while in the full premise and five-sentence setups it is a sentence hypothesis (the eighth template).

In contrast to the coarse-grained results, in the fine-grained task, all three premise shortening options show superior performance compared to the full premise option in terms of maximum and mean scores, with the random sentence setup achieving the second highest peak and mean. This finding suggests that the one-sentence setting is favored in this task, probably linked to the drastic increase in the number of categories, despite the loss of informative content. Longer sentences appear to introduce noise into the classification task, which detracts from the overall performance.

4.2 Few-shot experiments

4.2.1 Experimental setup

4.2.1.1 Main experiments

The few-shot experiments investigate the performance of our three entailment models when they are further pre-trained with a limited set of training and development examples from the target TC task converted into the NLI format. Four training and development data regimes are explored on the three entailment models: 1-1, 8-4, 16-8, and 32-16, where the first and second digits refer to the number of training and development examples per class, respectively. For greater representativeness of results, for each of the data regimes, 3 samples are generated at random from the training and development sets. The hyperparameters are kept fixed at a learning rate of 3e-5, a batch size of 16, a seed set at 26, and a maximum of 10 epochs, and the development set is used to select the best checkpoint according to the highest weighted F1 score obtained in the classification task.

Each of the three models is trained and evaluated (using the training and development schemes presented) in the two premise-shortening setups (full premise and first sentence as premise), and the best setup (the one with the highest F1 mean results across the three samples) is selected for the final evaluation in the test set. The experiments are carried for TeCla's coarse-grained and fine-grained tasks.

For the conversion of the training and development data into the NLI format, the best templates for each model and premise-shortening setting are used according to the

results from the zero-shot experiments (summarized in Table 4). With regard to the proportion of entailment and non-entailment cases generated, each example in the original classification dataset creates one positive (entailment) premise-hypothesis pair by using the correct output label, and each incorrect label is used to generate a negative (not entailment) premise-hypothesis pair[18]. This choice, as well as the selection of the best NLI training checkpoint based on the TC task results, is examined in greater depth in Appendix B through some additional experiments. Finally, we note that, although the three entailment models are trained with three output classes (entailment, neutral, and contradiction), the recast NLI data is binary (entailment and neutral, or not entailment), in acknowledgement of the difficulty of creating the three-way distinction from TC data. In this case, the new training forces the entailment model to forget the unseen class. In the test evaluation, we compare our NLI approach to the following techniques:

- **Supervised models.** We performed a standard fine-tuning in the target classification task using the available training and development data. We used our two pre-trained LM, monolingual and multilingual, respectively: RoBERTa-base-ca-v2 and XLM-RoBERTa-base.

- **SetFit.** This technique, explained in Section 2.1, has recently obtained SOTA results in various benchmarks, as reported by Tunstall et al. (2022). Unlike other SOTA techniques, this one, as well as ours, does not require overly expensive computing resources or additional unlabeled data. For its implementation, not having found any available sentence transformer (ST) in Catalan at the time of writing this, we used a multilingual ST, paraphrase-multilingual-mpnet-base-v2[19], which uses XLM-RoBERTa-base as the base model (and, as a teacher, the paraphrase-mpnet-base-v2 model) and has been trained on parallel data from over 50 languages, including Catalan, with a maximum sequence length of 128. To train the models, we used the default configuration options from the official library[20]: batch size of 16, 1 epoch, cosine-similarity loss, and 20 iterations to generate sentence pairs.

4.2.1.2 Task transfer experiments

In this section, we also investigate the potential for **knowledge transfer** from a related task to enhance the performance of the entailment-based method for text classification in the zero-shot and few-shot setups in Catalan. To this end, we construct a classification dataset to be used as the task transfer source, herein referred to as Wikicorpus, by scraping a number of Catalan Wikipedia articles and their associated categories. We focus on the monolingual pre-trained model RoBERTa-base-ca-v2 and explore two task transfer learning scenarios. In the first scenario (RoBERTa-ca-Wikicorpus), we propose to pre-train an entailment model on the TC dataset, Wikicorpus, recast as an NLI task. This

[18] In the coarse-grained task, therefore, each classification example generates 4 premise-hypothesis pairs, and 53 in the fine-grained task.

[19] The model is available in the HuggingFace repository at `https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2`

[20] Available at `https://github.com/huggingface/SetFit`

approach differs from our previous methodology, which utilized the general NLI dataset in Catalan (Teca) for pre-training. The second scenario (RoBERTa-ca-Teca-Wikicorpus) involves initial training of the entailment model on the general NLI dataset in Catalan, followed by a fine-tuning on Wikicorpus recast as an NLI task.

Wikicorpus creation. Wikicorpus comprises 21,002 article summaries extracted from the Catalan Wikipedia, generally one paragraph in length, and their corresponding category labels. The dataset includes 67 exclusive classes[21], with no overlap with the coarse-grained and fine-grained categories in the TeCla dataset. To construct Wikicorpus, we selected articles from a variety of subcategories within the 8 main available categories[22] to achieve a wide range of thematic diversity. The selection of categories was performed manually, with priority given to those that appeared to be more consistent in theme, many of which pertained to specific professional disciplines and broad social sciences-related themes. To ensure class exclusivity, we discarded texts that were associated with more than one category. From the total number of articles, a stratified 5% (1,050) was reserved for use as the development set.

Training configurations. In order to train the two entailment models, RoBERTa-ca-Wikicorpus and RoBERTa-ca-Teca-Wikicorpus, we first transformed the Wikicorpus dataset into the entailment format. In this process, to generate the hypothesis, we employed the same approach outlined in Section 3.1, utilizing the seventh template from Table 6 ("Aquest article tracta sobre {label}."), which yielded consistently strong results across both the coarse-grained and the fine-grained tasks for our model. To balance the proportion of entailment and non-entailment hypotheses and avoid the computational cost of multiplying the dataset size by 67 classes (which would be the case if we generated all possible non-entailment hypotheses for each entailment hypothesis), we decided to generate one non-entailment per each entailment hypothesis, thereby only increasing the dataset size by a factor of two. For the training of the entailment models with Wikicorpus, we kept the configurations used in the main few-shot experiments previously presented in this section: we selected the best checkpoint according to the weighted F1 score in the classification task and kept the same fixed hyperparameters.

The experiments explore both zero-shot and few-shot scenarios. To gain a deeper understanding of the contribution of each NLI dataset (whether general or from a related task), we compare our results with those from the following models:

[21]The categories are the following: Administració, Aeronàutica, Agricultura, Antropologia, Arqueologia, Arquitectura, Art, Astronomia, Astronàutica, Biblioteconomia, Biotecnologia, Catàstrofes, Circ, Ciència militar, Ciència-ficció, Ciències ambientals, Ciències de la salut, Ciències polítiques, Conflictes, Cronometria, Cultura popular, Dansa, Dret, Ecologia, Enginyeria, Epidèmies, Esoterisme, Estris, Festivals, Filologia, Filosofia, Fiscalitat, Física, Geografia, Geologia, Gestió, Heràldica, Història, Humor, Indumentària, Informàtica, Jaciments paleontològics, Jocs, Lingüística, Llengües, Llocs ficticis, Matemàtiques, Metodologia, Mitologia, Multimèdia, Museologia, Nàutica, Objectes astronòmics, Pedagogia, Periodisme, Protestes, Pseudociència, Psicologia, Química, Robòtica, Ràdio, Seguretat laboral, Sociologia, Telecomunicacions, Televisió, Teologia, Ètica.

[22]The 8 main categories can be found at `https://ca.wikipedia.org/wiki/Categoria:Classificacions_temÃ¤tiques_principals`

- **RoBERTa-ca-Teca**, from the previous section. The results included are drawn straight from it.

- **RoBERTa-base-ca-v2** directly trained on the NLI task using the target task available data, with no previous NLI pre-training. This allows us to investigate the contribution of the entailment model pre-training before the use of the available downstream data.

For the few-shot training setup in the task transfer experiments, we convert the available training data into the entailment format, as previously described in the methodological chapter, by generating all possible non-entailment hypotheses. Regarding the templates used for generating hypotheses, entailment models trained with Wikicorpus use the same template chosen for converting Wikicorpus into the NLI format, while the best-performing templates from the zero-shot experiments (for coarse-grained and fine-grained tasks, respectively, template 16 and template 7) are used for models trained with RoBERTa-base-ca-v2 directly on the target task data. In all these experiments, we only consider the full premise setup, as it yielded the best overall results in the previous experiments.

4.2.2 Main results

Table 11 and Table 12 show the results of the three entailment models on the development set for the coarse-grained and the fine-grained tasks, respectively. Note that, because of the reduced development set used and the different number of development examples employed across few-shot scenarios, the results cannot be used to judge the model's performance, but instead to compare premise-shortening setups within the same few-shot regime. The tables evidence that, in all the few-shot scenarios and tasks, the three models learn better from the full premise setup than from the first sentence, with only one exception observed in RoBERTa-ca-Teca in the 16-8 configuration. This is not truly a surprise, however, since RoBERTa-ca-Teca has the lowest differences between the full premise and first sentence results in the coarse-grained task: disregarding the one-shot setting, which has huge standard deviations, RoBERTa-ca-Teca has between 1.9 - 4.2 absolute points of difference between setups, while the XLM models range between 7.5 - 15.2. In the fine-grained task, the difference between the results for the full premise and first sentence is intensified for all the models, but the tendency is as well detected: RoBERTa-ca-Teca has the lowest differences (7.5 - 8.8, with respect to 11.7 - 14.4 in the rest of the models).

It is worth noting that, in the fine-grained task, the XLM models (and especially XLMR-Teca) were not able to learn with the predefined hyperparameters, i.e. they yielded F1 scores of around 1 or 2 percentage points. To address the problem, we re-ran the experiments on the fine-grained task for both XLM models with a smaller learning rate, 1e-5, and left the other hyperparameters unchanged, which resulted in all models learning.

The test set performances for the entailment models and the three baselines can be found in Table 13 and Table 14. The results are organized in the form of answers to three key questions.

Model	Premise type	1-1	8-4	16-8	32-16
RoBERTa-ca-Teca	full premise	**77.8 ± 19.2**	**87.1 ± 6.6**	88.6 ± 6.3	**92.7 ± 3.2**
	first sentence	72.2 ± 25.5	82.9 ± 13.8	**90.5 ± 8.4**	89.2 ± 8.5
XLMR-Teca	full premise	**77.8 ± 19.2**	**87.5 ± 6.0**	**91.9 ± 6.1**	**90.8 ± 4.2**
	first sentence	75.0 ± 43.3	78.3 ± 14.2	79.8 ± 10.5	83.4 ± 6.3
XLMR-SMAX	full premise	**88.9 ± 19.2**	**88.4 ± 11.4**	**91.4 ± 4.9**	**92.1 ± 2.9**
	first sentence	80.6 ± 33.7	73.2 ± 15.9	81.1 ± 10.8	83.6 ± 2.8

Table 11: Results for the coarse-grained task over the TeCla development set according to the two premise types examined (full premise or first sentence as premise) and the few-shot data regime: 1-1, 8-4, 16-8, and 32-16, where X-Y mean X examples/class as training data and Y examples/class as development data.

Model	Premise type	1-1	8-4	16-8	32-16
RoBERTa-ca-Teca	full premise	**47.9 ± 3.4**	**63.2 ± 3.3**	**65.3 ± 1.4**	**67.0 ± 1.7**
	first sentence	36.5 ± 2.3	55.6 ± 3.1	56.4 ± 0.9	59.5 ± 0.6
XLMR-Teca	full premise	**37.8 ± 1.1**	**57.2 ± 2.1**	**60.3 ± 2.6**	**63.5 ± 0.1**
	first sentence	27.2 ± 3.1	44.5 ± 0.9	46.3 ± 3.2	51.8 ± 0.3
XLMR-SMAX	full premise	**39.5 ± 3.4**	**56.7 ± 2.1**	**61.1 ± 0.8**	**64.4 ± 0.6**
	first sentence	27.5 ± 2.1	43.6 ± 3.4	46.7 ± 2.9	50.9 ± 0.4

Table 12: Results for the fine-grained task over the TeCla development set according to the two premise types examined (full premise or first sentence as premise) and the few-shot data regime: 1-1, 8-4, 16-8, and 32-16, where X-Y mean X examples/class as training data and Y examples/class as development data.

- **Which is the best-performing method for few-shot setups?**

In the coarse-grained and fine-grained tasks, the best results in few-shot scenarios are achieved by one of the entailment models. Specifically, in the 1-1, 8-4, and 16-8 scenarios, all three entailment models outperform the baselines, with the exception of RoBERTa-ca-Teca in the 16-8 scenario of the coarse-grained task, which falls short of SetFit's results. This discrepancy may be due to the fact that RoBERTa-ca-Teca is the only model trained with the first sentence as premise setup, indicating that the full premise would have probably been more beneficial for the few-shot training. In the 32-16 setting, although the best results are still obtained by the entailment models, the data is sufficient for supervised models to achieve comparable results. Overall, the results suggest that the greatest advantage of the entailment approach is its applicability in contexts where data is extremely scarce.

- **Which entailment models have a stronger ability to learn from scarce data regimes?**

The findings for the coarse-grained task indicate that there is no clear dominance of a particular entailment model across data size regimes. In the most data-scarce scenarios,

	Model	1-1	8-4	16-8	32-16
entailment models	RoBERTa-ca-Teca	56.9 ± 8.8	79.2 ± 3.3	82.4 ± 2.9	**89.2 ± 0.6**
	XLMR-Teca	**63.1 ± 2.0**	**81.5 ± 2.1**	**86.7 ± 1.3**	86.7 ± 2.8
	XLMR-SMAX	56.5 ± 8.3	79.7 ± 7.4	**86.7 ± 1.0**	87.7 ± 2.1
baselines	supervised-RoBERTa-base-ca-v2	28.5 ± 4.4	63.0 ± 9.8	74.7 ± 8.1	83.5 ± 2.9
	supervised-XLM-RoBERTa-base	8.9 ± 22.1	40.6 ± 24.0	65.4 ± 20.9	87.8 ± 1.6
	SetFit	47.7 ± 6.8	79.0 ± 6.2	84.7 ± 3.2	87.0 ± 1.7

Table 13: Results for the coarse-grained task over the TeCla test set in the following few-shot data regimes: 1-1, 8-4, 16-8, and 32-16, where X-Y mean X examples/class as training data and Y examples/class as development data. Each cell is the mean and standard deviation across three different samples.

	Model	1-1	8-4	16-8	32-16
entailment models	RoBERTa-ca-Teca	**48.5 ± 4.2**	**60.2 ± 1.4**	**62.4 ± 1.3**	**63.2 ± 1.3**
	XLMR-Teca	41.2 ± 4.1	51.3 ± 0.7	56.7 ± 1.8	60.7 ± 0.5
	XLMR-SMAX	40.0 ± 3.1	53.8 ± 2.2	57.0 ± 0.8	60.0 ± 0.9
baselines	supervised-RoBERTa-base-ca-v2	-	50.0 ± 5.1	54.4 ± 4.1	61.8 ± 2.4
	supervised-XLM-RoBERTa-base	-	44.0 ± 2.7	51.6 ± 1.9	61.3 ± 0.6
	SetFit	22.0 ± 3.6	50.3 ± 0.6	53.3 ± 1.2	56.7 ± 1.5

Table 14: Results for the fine-grained task over the TeCla test set in the following few-shot data regimes: 1-1, 8-4, 16-8, and 32-16, where X-Y mean X examples/class as training data and Y examples/class as development data. Each cell is the mean and standard deviation across three different samples. A dash indicates that the model has not been able to learn in the given setting.

1-1 and 8-4 setups, XLMR-Teca significantly stands out with the highest results, while RoBERTa-ca-Teca and XLMR-SMAX perform around 6 and 2 absolute points lower, respectively. At the 16-8 stage, XLMR-SMAX catches up to XLMR-Teca, and both models show little improvement as more data is added. Notably, RoBERTa-ca-Teca demonstrates the greatest ability to improve as more data is provided, reaching the highest F1 score in the 32-16 scenario. This results in a 2.5 and 1.5 performance boost over XLMR-Teca and XLMR-SMAX, respectively.

In the fine-grained task, the results are more straightforward and reveal tendencies similar to those observed in the zero-shot scenario. The RoBERTa-ca-Teca model outperforms the XLM models across all data ratios; however, the gap between them gradually narrows as more data is made available. This trend can be observed starting with an absolute difference of 8.5 and 7.3 with respect to XLMR-Teca and XLMR-SMAX, respectively, in the 1-1 scenario, and eventually reaching a difference of 3.2 and 2.6 when the data ratio is 32-16.

- **What are the baselines' performances compared to the entailment models?**

Supervised models exhibit a lower capacity for learning in low data regimes, but display a more accelerated progression as more data becomes available. Additionally, they tend to

be particularly unstable in the most extreme few-shot scenarios, showing high variations across samples; notably, in the one-shot setup of the fine-grained task, they were unable to learn, neither with a learning rate of 3e-5 nor 1e-5. In contrast to supervised learning, SetFit exhibits a greater ability in learning from scenarios with limited data, obtaining overall higher results in those setups. Moreover, it does not experience instability issues in one-shot experiments. In this data regime, however, SetFit consistently lags behind the entailment models by approximately 10 and 20 points in the first and second tasks. In the remaining data regimes, in the coarse-grained task, SetFit achieves results that are similar to those of the entailment models, yet it never surpasses the best-performing entailment model. On the other hand, in the fine-grained task, SetFit's performance is generally significantly below that of all the entailment models.

4.2.3 Task transfer results

We present the results of the development set in Table 15 and Table 16, just in order to provide a comprehensive report, and those of the test set in Table 17 and Table 18, where we included the few-shot baselines to facilitate an easy overview of all the results. We attempt to address the following questions: is pre-training on an NLI dataset beneficial in the few-shot scenario? Does task transfer learning lead to an improvement in the performance of the entailment-based method? It is worth noting that there are significant differences between the results for the coarse- and fine-grained tasks. As such, the findings for these tasks are discussed separately.

In the coarse-grained task, the performance of the entailment model trained solely with final-task data is inferior to that of the entailment model pre-trained with other NLI data. The decrease in performance ranges from 10 to 30 points, but is less pronounced as the amount of final-task training data available decreases. Among the models that were pre-trained with additional NLI data, the two entailment models trained with Wikicorpus (with or without a previous training on Teca) generally improve results compared to the RoBERTa-ca-Teca model. Specifically, the RoBERTa-ca-Wikicorpus model is the best among the two in all few-shot data regimes. The highest gains from task transfer learning are observed in zero- and one-shot experiments, with a 15.3 improvement in zero-shot performance achieved by the model trained only with Wikicorpus when compared to the RoBERTa-ca-Teca model. It is important to note that the one-shot setting in the coarse-grained task does not typically yield better results than the zero-shot setting. In fact, in some cases it leads to a significant decrease in performance, especially for the XLM model trained on multilingual NLI datasets.

In the remaining few-shot scenarios (from 8-4 to 32-16) in the coarse-grained task, the differences between the RoBERTa-base-ca-v2 models pre-trained with different NLI datasets become marginal —but is yet relatively far above the model without this additional NLI pre-training. Furthermore, the best monolingual model with task transfer learning is now able to outperform the multilingual entailment models and the few-shot baselines, except for the 8-4 setup, where the XLMR-Teca model is still 0.6 better.

In the fine-grained task, the augment in the number of classes seems to modify some of

Model	1-1	8-4	16-8	32-16
RoBERTa-base-ca-v2	38.9 ± 27.1	55.8 ± 11.6	67.5 ± 5.5	79.9 ± 9.1
RoBERTa-ca-Teca	77.8 ± 19.2	**87.1 ± 6.6**	90.5 ± 8.4	**92.7 ± 3.2**
RoBERTa-ca-Wikicorpus	77.8 ± 19.2	85.9 ± 8.6	88.3 ± 9.5	91.1 ± 4.8
RoBERTa-ca-Teca-Wikicorpus	**88.9 ± 19.2**	84 ± 11.7	**90.7 ± 8.2**	89.5 ± 4.9

Table 15: TeCla development results (weighted F1) for the coarse-grained task using four different NLI pre-trainings for RoBERTa-base-ca-v2. In each of the four data-regime setup, the results reported are the mean and standard deviation across three training and development samples.

LM	1-1	8-4	16-8	32-16
RoBERTa-base-ca-v2	9.2 ± 5.8	**63.7 ± 1.8**	65.2 ± 2.3	**67.9 ± 1.2**
RoBERTa-ca-Teca	47.9 ± 3.4	**63.2 ± 3.3**	**65.3 ± 1.4**	**67 ± 1.7**
RoBERTa-ca-Wikicorpus	51 ± 6.8	63.1 ± 1.2	65.3 ± 3.9	67.2 ± 1
RoBERTa-ca-Teca-Wikicorpus	**52.5 ± 5.3**	63.4 ± 1.8	**64.9 ± 2.3**	67.4 ± 0.9

Table 16: TeCla development results (weighted F1) for the fine-grained task using four different NLI pre-trainings for RoBERTa-base-ca-v2. In each of the four data-regime setup, the results reported are the mean and standard deviation across three training and development samples.

	Model	zero-shot	1-1	8-4	16-8	32-16
monolingual entailment models	RoBERTa-base-ca-v2	-	36.1 ± 9.5	45.0 ± 10.5	65.9 ± 4.4	78.1 ± 8.2
	RoBERTa-ca-Teca	59.7	56.9 ± 8.8	79.2 ± 3.3	82.4 ± 2.9	89.2 ± 0.6
	RoBERTa-ca-Wikicorpus	**75.0**	**74.8 ± 0.6**	80.9 ± 5.8	**87.7 ± 0.9**	**89.6 ± 0.1**
	RoBERTa-ca-Teca-Wikicorpus	66.4	66.5 ± 0.9	79.9 ± 4.8	86.5 ± 1.1	88.2 ± 1.6
multilingual entailment models	XLMR-Teca	66.0	63.1 ± 2.0	**81.5 ± 2.1**	86.7 ± 1.3	86.7 ± 2.8
	XLMR-SMAX	71.8	56.5 ± 8.3	79.7 ± 7.4	86.7 ± 1.0	87.7 ± 2.1
baselines	supervised-RoBERTa-base-ca-v2	-	28.5 ± 4.4	63.0 ± 9.8	74.7 ± 8.1	83.5 ± 2.9
	supervised-XLM-RoBERTa-base	-	8.9 ± 22.1	40.6 ± 24	65.4 ± 20.9	87.8 ± 1.6
	SetFit	-	47.7 ± 6.8	79 ± 6.2	84.7 ± 3.2	87.0 ± 1,7

Table 17: TeCla test set results (weighted F1) for the coarse-grained task using four different NLI pre-trainings for RoBERTa-base-ca-v2 against the multilingual entailment models and baselines from the previous sections. In few-shot, the results are the mean and standard deviation across three training and development samples within each data regime. A dash indicates that the model has not been able to learn in the given setting.

the trends observed in the coarse-grained task. Specifically, when training for the NLI task using only final-task data, only poor performance is observed in zero- and one-shot settings, but the results with the remaining data regimes align with those from any kind of NLI pre-training. Nevertheless, the best results are still always achieved by some of the models with an NLI pre-training. Besides, in line with the findings from the coarse-grained task, transfer learning demonstrates the most significant improvements in zero- and one-shot settings, with the model RoBERTa-ca-Teca-Wikicorpus achieving the greatest improvements, where

	Model	zero-shot	1-1	8-4	16-8	32-16
monolingual entailment models	RoBERTa-base-ca-v2	-	13.8 ± 6.7	60.8 ± 2.8	62.1 ± 1.3	63.6 ± 2.1
	RoBERTa-ca-Teca	36.3	48.5 ± 4.2	60.2 ± 1.4	**62.4 ± 1.3**	63.2 ± 1.3
	RoBERTa-ca-Wikicorpus	49.1	51 ± 2.9	**60.9 ± 0.3**	60.9 ± 0.4	64.2 ± 0.9
	RoBERTa-ca-Teca-Wikicorpus	**49.8**	**53.4 ± 2.4**	59.7 ± 1.2	61.5 ± 1.1	**64.3 ± 0.4**
multilingual entailment models	XLMR-Teca	28.0	41.2 ± 4.1	51.3 ± 0.7	56.7 ± 1.8	60.7 ± 0.5
	XLMR-SMAX	31.6	40.0 ± 3.1	53.8 ± 2.2	57.0 ± 0.8	60.0 ± 0.9
baselines	supervised-RoBERTa-base-ca-v2	-	-	50.0 ± 5.1	54.4 ± 4.1	61.8 ± 2.4
	supervised-XLM-RoBERTa-base	-	-	44.0 ± 2.7	51.6 ± 1.9	61.3 ± 0.6
	SetFit	-	22.0 ± 3.6	50.3 ± 0.6	53.3 ± 1.2	56.7 ± 1.5

Table 18: TeCla test set results (weighted F1) for the fine-grained task using four different NLI pre-trainings for RoBERTa-base-ca-v2 against the multilingual entailment models and baselines from the previous sections. In few-shot, the results are the mean and standard deviation across three training and development samples within each data regime. A dash indicates that the model has not been able to learn in the given setting.

the results increase by 13.5 and 4.9 points with respect to RoBERTa-ca-Teca in these two setups, respectively.

However, from 8-4 to 32-16, there are only slight, non-significant variations observed across all RoBERTa-base-ca-v2-based entailment models, further reinforcing the coarse-grained trend and interpretation that the use of additional NLI datasets for pre-training becomes less important as more final task data is available. When compared to the multilingual entailment models and baselines, where NLI monolingual models already yielded the best results, task transfer learning greatly contributes to accentuating this superiority in zero- and one-shot settings, with only slight contributions in other settings.

5 Conclusion

This section provides a summary of the main findings of our research and their relevance to the objectives and questions addressed in the study; it also discusses its limitations and identifies areas for future research. Our results provide insight into the effectiveness of an entailment-based approach in zero- and few-shot settings on a Catalan TC dataset. Specifically, by experimenting with monolingual and multilingual resources, we investigated how the amount of text seen in the LM pre-training and the size of the NLI dataset affect this approach. Furthermore, through our task transfer experiments, we explored a potential improvement of the method over a monolingual entailment model. Finally, an additional contribution of this work is the creation of a new, revised version of the Catalan TC dataset, TeCla.

In **zero-shot** settings, the implementation of an entailment-based approach using a fully monolingual entailment model (i.e. trained with a monolingual LM and a monolingual NLI training dataset) was found to achieve quite reasonable performances, capable of improving the best baseline considered by more than 10 points in average. When compared to entailment models trained with multilingual LMs, significant differences were observed in the performance of the coarse-grained (4 categories) and fine-grained (53 categories) tasks of the Catalan TC dataset, TeCla, highlighting the strong influence of task specificities on the performance of entailment models. In the coarse-grained task, the multilingual LM trained with several multilingual NLI datasets achieved an outstanding 71.1 in weighted F1, followed at 7.2 points lower by the multilingual LM trained with the (smaller) Catalan NLI dataset, and finally, at a closer distance (4.2 points below), the Catalan LM trained with the Catalan NLI dataset. However, in the fine-grained task, the Catalan LM trained on the monolingual NLI data performed the best with a score of 36.3, followed at 4.9 of distance by the multilingual model using the multilingual NLI data, and finally, at 4.4 points below the latter, by the one using the monolingual data.

These findings suggest that the larger size of the pre-trained LM and NLI training dataset contributes significantly to improving the results in the coarse-grained TC task, with the NLI dataset providing the largest improvements. In contrast, in the fine-grained task, the monolingual LM provides improvements over a larger multilingual LM, although the larger NLI dataset is still a better contribution than the monolingual NLI dataset. Based on these results, we argue that this change in tendencies from one task to another could be due to the language-specific knowledge needed for the task: for the coarse-grained task, the knowledge required was not as language-dependent and could therefore benefit from a larger resource size, even though it was not language-specific, while this type of knowledge was relevant for the fine-grained task, and was essentially retrieved from the LM.

Regarding experiments with different template and premise-shortening setups, the results revealed unstable performances across variations in these areas affecting all entailment models, which probably highlights inherent limitations of the entailment-based approach. Indeed, different templates seem to have an impact on the model's ability to take ad-

vantage of longer premises; in other words, certain templates work better with specific premise-shortening setups. With regard to premise shortenings in particular, the multilingual entailment model trained on multilingual NLI datasets demonstrated the highest capacity to benefit from the full premise setting, while the monolingual entailment model performed best when using the first sentence as the premise. Our exploration suggests that the uniform distribution of premise length in the monolingual general-domain NLI dataset may be limiting the model's ability to deal with varying premise lengths. Increasing the size and diversity of the NLI dataset could help to overcome this issue.

In the **few-shot** experiments, the entailment-based approach (one or all of our entailment models) always achieves the best results compared to the baselines, i.e. supervised models and another few-shot SOTA technique, SetFit, across all few-shot data regimes, proving its worth in these scenarios for solving TC tasks. However, entailment models offer the greatest advantage in contexts where data is more scarce, since the gap with the other systems narrows as more training data becomes available. In particular, the difference between the best entailment model and the best baseline system is, in the coarse-grained task, 15.4 in the one-shot regime, and then around 2 in the 8-shot and 16-shot settings; in the fine-grained task, 26.5 in one-shot, then 9.9 in 8-shot and 8.0 in 16-shot. It should be noted, however, that the one-shot setting does not generally improve the zero-shot results in the coarse-grained task, but it does in the fine-grained task, where the number of training examples is larger —since the NLI conversion multiplied each example by the 53 classes in the task. On the other hand, with respect to the baselines, the supervised models have the fastest learning progression as more examples are added, approaching the entailment models already in the 32-16 setup. For its part, the performance of SetFit generally lags slightly behind the worst entailment model, except in the one-shot setup, where the gap is larger.

When comparing the entailment models, as in the zero-shot setting, the trends are not consistent between the coarse-grained and fine-grained tasks. In the coarse-grained tasks, there is no clear dominance across all few-shot regimes; interestingly, however, in the most extreme data-scarce regimes, the multilingual model fine-tuned with monolingual data outperforms the others[23]. Its results then converge with those of the multilingual model trained with several multilingual NLI datasets. However, the monolingual model achieves the best results in the 32-16 setup and shows the greatest ability to improve as more data is added. In the fine-grained task, as happened in zero-shot, this model consistently outperforms the multilingual models (by around 7 points in the one-shot setting, 6 in 8-shot, 5 in 16-shot, and almost 3 in 32-shot), which achieve similar results. Regarding the premise shortening setups, the three entailment models now appear to learn better from the full premise, and this is intensified in the fine-grained task, although the difference between the full premise results and the first sentence results is greater in the monolingual model.

[23]From the zero- to the one-shot setting, this model worsens its performance by almost 3 points, and the same happens to RoBERTa-ca-Teca, while the same pre-trained model (XLM) fine-tuned with the multilingual NLI datasets worsens its performance by over 15 points, proving serious difficulties in generalizing with the given training data. The monolingual NLI dataset, either due to its smaller size or to being in the same language as the target task, seems to favor the model's capabilities to learn from very few examples.

This suggests that the difficulty in learning from the full premise observed mainly in the models fine-tuned with the monolingual NLI dataset is widely overcome in the few-shot setting.

Overall, in the few-shot results, the NLI pre-training of the model seems to lose relevance in the model's performance, and this is more accentuated as the amount of training data on the target task increases. With the exception of the one-shot and eight-shot settings, where the multilingual model fine-tuned on the monolingual dataset performs best, the results of the two multilingual models tend to converge and differ from the behavior of the monolingual model. Thus, the pre-trained LM apparently gains more importance in these settings, and the monolingual LM, despite being pre-trained on less text, shows the highest ability to learn from new data. This is particularly evident in the fine-grained task of our experiments, where the model's language-specific knowledge is especially useful, as was the case in the zero-shot experiments.

The use of **task transfer learning** significantly improved the performance of the entailment-based approach in zero- and one-shot settings, when compared to models trained solely on the general NLI dataset. However, as the number of training examples for the target task increases, the importance of task transfer learning decreases. Improvements were also observed in other few-shot regimes, particularly in the coarse-grained task, where there are fewer training examples (the same for each class, but has fewer classes). Concerning the best task transfer learning setup, the model trained solely on the TC dataset transformed into NLI, without the general NLI dataset, performed better in the coarse-grained task, whereas the combination of both datasets generally yielded better results in the fine-grained task, suggesting that different tasks benefit unequally from different NLI pre-training datasets.

In summary, the entailment-based approach applied in this work has proven to be an effective technique for tackling a TC problem in the Catalan language; two main advantages that should be noted are its ability to use an already existing (non-target task) NLI dataset, which makes it particularly useful in zero-shot settings, and the data augmentation performed on the few target task examples available in their conversion to NLI, which gives the model more robustness. In the zero-shot setting, the size of the NLI dataset becomes a critical factor: a larger NLI dataset not only improves the model's inference capabilities, but also reduces the potential for bias —for instance, with respect to the model's ability to understand premises of different lengths. The larger size of the LM is also beneficial, but the language-specific knowledge from the monolingual LM appears to be even more valuable for certain tasks. The size factor becomes less important in the few-shot setting, and the monolingual LM generally shows comparable or often higher learning abilities. Finally, the presented task transfer learning setup is found to bring further gains to the approach, especially in the most data-scarce regimes.

The experiments also highlighted the **inherent limitations of the entailment-based method** used in this work. First, in the zero-shot scenario, when using an entailment model trained on a general NLI dataset, the variations in the template and label verbalization used, as well as in the premise length, have the potential to drastically alter the results, with no reliable way of predicting any of these parameters in advance without the aid

of a development set. Note that this is also an inconvenient of prompt-based methods requiring manually written prompts. However, we found that this problem could be at least partially avoided by using a related-task dataset recast as an entailment to train the entailment model given that the configurations chosen for the NLI pre-training can also be used in the zero-shot scenario, achieving more stable results. Second, a task that deals with many output classes multiplies the number of premise hypotheses generated at inference time and thus significantly slows down the inference process; this does not happen in SetFit, for example, which also yields competitive results. Third, a short or insufficiently diverse NLI pre-training dataset may introduce some bias in zero-shot performance that alter, for instance, the model's ability to deal with long premises if not seen before. Finally, the method demonstrated its strength in extremely data-poor scenarios, but its learning capabilities slow down as more data becomes available and are soon surpassed by other methods.

The **limitations of the present study** include the use of only one TC task, which provided some provisional findings and insights, but may not necessarily be generalizable to other text classification tasks. Furthermore, the study only examined the differences between coarse- and fine-grained TC tasks superficially and proposed an interpretation based on linguistic specificities, but a more in-depth analysis is needed to fully understand these differences, including a deep look at the errors of each model. Additionally, no hyperparameter search was conducted for the entailment models, and their influence on the results was not considered. Lastly, the study only compared an entailment-based approach with some baselines and SOTA techniques for zero- and few-shot settings, but more methods should be included to fully evaluate the usefulness of this approach in relation to others.

Future work should aim to address the limitations of the entailment-based approach, such as reducing the dependence on handwritten templates and verbalizations for the entailment-based method, particularly in zero-shot settings. One potential solution could be to incorporate advances from prompt-based learning, such as automatic retrieval or generation in natural language or in continuous embedding space. Moreover, since the size and diversity of the NLI dataset plays a crucial role in zero-shot settings, efforts should be made to overcome the limitations of existing NLI datasets, especially for less-resourced languages, by (automatically, if possible) augmenting the dataset to include a wider variety of examples (for instance, in terms of premise length and register variation). Lastly, the possibilities of task transfer learning within the approach should be further explored, particularly in the context of new classification tasks such as sentiment analysis, which may require more abstract inference capabilities. This could include determining what kind of data is most useful for each task when reformulated as NLI and what knowledge is useful in each case, or, perhaps more ambitiously, identifying data that can be efficiently reformulated as NLI to provide broader inference capabilities to enable its use across tasks.

References

Jordi Armengol-Estapé, Casimiro Pio Carrino, Carlos Rodriguez-Penagos, Ona de Gibert Bonet, Carme Armentano-Oller, Aitor Gonzalez-Agirre, Maite Melero, and Marta Villegas. Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4933–4946, 7 2021. doi: 10.48550/arxiv.2107.07903. URL https://arxiv.org/abs/2107.07903v1.

Trapit Bansal, Salaheddin Alzubi, Tong Wang, Jay-Yoon Lee, and Andrew Mccallum. Meta-adapters: Parameter efficient few-shot fine-tuning through meta-learning. 2022.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. The second PASCAL recognising textual entailment challenge. 2006.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. The fifth PASCAL recognizing textual entailment challenge. In *TAC*. Citeseer, 2009a. URL https://tac.nist.gov/publications/2009/additional.papers/RTE5_overview.proceedings.pdf.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. The sixth PASCAL recognizing textual entailment challenge. In *Text Analysis Conference*, 2009b.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. The seventh PASCAL recognizing textual entailment challenge. 2011.

Rajeshree Bora-Kathariya and Yashodhara Haribhakta. Natural language inference as an evaluation measure for abstractive summarization. *2018 4th International Conference for Convergence in Technology, I2CT 2018*, 10 2018. doi: 10.1109/I2CT42659.2018.9057819.

Johan Bos and Katja Markert. Recognising textual entailment with logical inference. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 628–635, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics. URL https://aclanthology.org/H05-1079.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya

Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. 5 2020. URL http://arxiv.org/abs/2005.14165.

Ming-Wei Chang, Lev Ratinov, Dan Roth, and Vivek Srikumar. Importance of semantic representation: Dataless classification. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, page 830–835. AAAI Press, 2008. ISBN 9781577353683.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1152. URL https://aclanthology.org/P17-1152.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. *WWW 2022 - Proceedings of the ACM Web Conference 2022*, pages 2778–2788, 4 2021. doi: 10.1145/3485447.3511998. URL http://arxiv.org/abs/2104.07650http://dx.doi.org/10.1145/3485447.3511998.

Aurélien Coet. Deep learning for natural language inference: A literature review. 2019.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Holger Schwenk, Ves Stoyanov, Adina Williams, and Samuel R. Bowman. Xnli: Evaluating cross-lingual sentence representations. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 2475–2485, 2018. doi: 10.18653/V1/D18-1269. URL https://aclanthology.org/D18-1269.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. pages 8440–8451, 7 2020. doi: 10.18653/V1/2020.ACL-MAIN.747. URL https://aclanthology.org/2020.acl-main.747.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Josef Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, Steve Pulman, Ted Briscoe, and Karsten Konrad Fracas. Using the framework. The FRACAS consortium. 1996.

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. Template-based named entity recognition using bart. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845, 6 2021. doi: 10.48550/arxiv.2106.01760. URL https://arxiv.org/abs/2106.01760v1.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3944 LNAI:177–190, 2006. ISSN 03029743. doi: 10.1007/11736790_9.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Zanzotto. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6: 1–222, 7 2013. ISSN 19474040. doi: 10.2200/S00509ED1V01Y201305HLT023/SUPPL_FILE/DAGAN_CH1.PDF.

Dorottya Demszky, Kelvin Guu, and Percy Liang. Transforming question answering datasets into natural language inference datasets. 9 2018. URL `http://arxiv.org/abs/1809.02922`.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:4171–4186, 10 2018. doi: 10.48550/arxiv.1810.04805. URL `https://arxiv.org/abs/1810.04805v2`.

Zi Yi Dou, Keyi Yu, and Antonios Anastasopoulos. Investigating meta-learning algorithms for low-resource natural language understanding tasks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 1192–1197, 2019. doi: 10.18653/V1/D19-1112. URL `https://aclanthology.org/D19-1112`.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.435. URL `https://aclanthology.org/2022.acl-long.435`.

E Fonseca, L Santos, Marcelo Criscuolo, and S Aluisio. Assin: Avaliacao de similaridade semantica e inferencia textual. In *Computational Processing of the Portuguese Language-12th International Conference, Tomar, Portugal*, pages 13–15, 2016.

Jie Gao, Hella Franziska Hoffmann, Stylianos Oikonomou, David Kiskovski, and Anil Bandhakavi. Logically at factify 2022: Multimodal fact verification. *CEUR Workshop*

Proceedings, 3168, 12 2021a. ISSN 16130073. doi: 10.48550/arxiv.2112.09253. URL https://arxiv.org/abs/2112.09253v2.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 3816–3830, 2021b. doi: 10.18653/V1/2021.ACL-LONG.295. URL https://aclanthology.org/2021.acl-long.295.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third PASCAL recognizing textual entailment challenge. 2007. doi: 10.5555/1654536.1654538. URL https://dl.acm.org/doi/10.5555/1654536.1654538.

Danilo Giampiccolo, Hoa Trang Dang, Bernardo Magnini, Ido Dagan, Elena Cabrio, and William B. Dolan. The fourth PASCAL recognizing textual entailment challenge. In *Text Analysis Conference*, 2008.

Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. KorNLI and KorSTS: New benchmark datasets for korean natural language understanding. *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*, pages 422–430, 2020. doi: 10.18653/V1/2020.FINDINGS-EMNLP.39.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020. ISSN 2307387X. doi: 10.1162/TACL_A_00324.

V. Jijkoun and M. de Rijke. Recognizing textual entailment using lexical similarity. Information and Language Processing Syst (IVI, FNWI), 2005.

Tushar Khot, Ashish Sabharwal, and Peter Clark. Scitail: A textual entailment dataset from science question answering. In *AAAI Conference on Artificial Intelligence*, 2018.

Venelin Kovatchev and Mariona Taulé. InferES : A natural language inference corpus for Spanish featuring negation-based contrastive and adversarial examples. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3873–3884, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.340.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. SummaC: Revisiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022. doi: 10.1162/tacl_a_00453. URL https://aclanthology.org/2022.tacl-1.10.

Cheolhyoung Lee, Kyunghyun Cho, and Wanmo Kang. Mixout: Effective regularization to finetune large-scale pretrained language models. 9 2019. doi: 10.48550/arxiv.1909.11299. URL https://arxiv.org/abs/1909.11299v2.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. *CoNLL 2017 - 21st Conference on Computational Natural Language Learning, Proceedings*, pages 333–342, 2017. doi: 10.18653/V1/K17-1034. URL https://aclanthology.org/K17-1034.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 4582–4597, 2021. doi: 10.18653/V1/2021.ACL-LONG.353. URL https://www.researchgate.net/publication/353489963_Prefix-Tuning_Optimizing_Continuous_Prompts_for_Generation.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. 5 2022. doi: 10.48550/arxiv.2205.05638. URL https://arxiv.org/abs/2205.05638v2.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. 2021.

Tingting Ma, Jin Ge Yao, Chin Yew Lin, and Tiejun Zhao. Issues with entailment-based zero-shot text classification. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2:786–796, 2021. doi: 10.18653/V1/2021.ACL-SHORT.99.

Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 565–576, 2021. doi: 10.18653/V1/2021.ACL-LONG.47. URL https://aclanthology.org/2021.acl-long.47.

Rabeeh Karimi Mahabadi, Luke Zettlemoyer, James Henderson, Marzieh Saeidi, Lambert Mathias, Veselin Stoyanov, Majid Yazdani, and Meta Ai. Perfect: Prompt-free and efficient few-shot learning with language models. 4 2022. doi: 10.48550/arxiv.2204.01172. URL https://arxiv.org/abs/2204.01172v2.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. The natural language decathlon: Multitask learning as question answering. 6 2018. doi: 10.48550/arxiv.1806.08730. URL https://arxiv.org/abs/1806.08730v1.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *1st International Conference on Learning Representa-*

tions, ICLR 2013 - Workshop Track Proceedings, 1 2013. doi: 10.48550/arxiv.1301.3781. URL https://arxiv.org/abs/1301.3781v3.

George A. Miller. Wordnet: A lexical database for english, 1994. URL https://aclanthology.org/H94-1111.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.441. URL https://aclanthology.org/2020.acl-main.441.

Abiola Obamuyide and Andreas Vlachos. Zero-shot relation classification as textual entailment. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 72–78, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5511. URL https://aclanthology.org/W18-5511.

Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. A syntactic textual entailment system based on dependency parser. *Lecture Notes in Computer Science*, pages 269–278, 2010. ISSN 0302-9743. URL https://www.academia.edu/1148545/A_Syntactic_Textual_Entailment_System_Based_on_Dependency_Parser.

Aarthi Paramasivam and S. Jaya Nirmala. A survey on textual entailment based question answering. *Journal of King Saud University - Computer and Information Sciences*, 34: 9644–9653, 11 2022. ISSN 1319-1578. doi: 10.1016/J.JKSUCI.2021.11.017.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language models as knowledge bases? *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 2463–2473, 2019. doi: 10.18653/V1/D19-1250. URL https://aclanthology.org/D19-1250.

Jason Phang, Thibault Févry, and Samuel R. Bowman. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. 11 2018. doi: 10.48550/arxiv.1811.01088. URL https://arxiv.org/abs/1811.01088v2.

Adam Poliak. A survey on recognizing textual entailment as an NLP evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 92–109, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.eval4nlp-1.10. URL https://aclanthology.org/2020.eval4nlp-1.10.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. Collecting diverse natural language

inference problems for sentence representation evaluation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, 2018. doi: 10.18653/V1/D18-1007. URL https://aclanthology.org/D18-1007.

Raul Puri and Bryan Catanzaro. Zero-shot text classification with generative language models. *ArXiv*, 2019.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 3982–3992, 8 2019. doi: 10.48550/arxiv.1908.10084. URL https://arxiv.org/abs/1908.10084v1.

Oscar Sainz and German Rigau. Ask2transformers: Zero-shot domain labelling with pre-trained language models. *GWC 2021 - Proceedings of the 11th Global Wordnet Conference*, pages 44–52, 1 2021. doi: 10.48550/arxiv.2101.02661. URL https://arxiv.org/abs/2101.02661v2.

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. Label verbalization and entailment for effective zero- and few-shot relation extraction. 9 2021. URL http://arxiv.org/abs/2109.03659.

Oscar Sainz, Itziar Gonzalez-Dios, Oier Lopez de Lacalle, Bonan Min, and Eneko Agirre. Textual entailment for event argument extraction: Zero- and few-shot with multi-source learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2439–2455, Seattle, United States, July 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.187. URL https://aclanthology.org/2022.findings-naacl.187.

Oscar Sainz, Haoling Qiu, Oier Lopez de Lacalle, Eneko Agirre, and Bonan Min. Zs4ie: A toolkit for zero-shot information extraction with simple verbalizations. *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, pages 27–38, 3 2022b. doi: 10.48550/arxiv.2203.13602. URL https://arxiv.org/abs/2203.13602v3.

Timo Schick and Hinrich Schütze. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:8766–8774, 4 2020. ISSN 2374-3468. doi: 10.1609/AAAI.V34I05.6403. URL https://ojs.aaai.org/index.php/AAAI/article/view/6403.

Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. *NAACL-HLT 2021 - 2021 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, pages 2339–2352, 2021a. doi: 10.18653/v1/2021.naacl-main.185.

Timo Schick and Hinrich Schütze. Exploiting cloze questions for few shot text classification and natural language inference. *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 255–269, 2021b. doi: 10.18653/V1/2021.EACL-MAIN.20.

Taylor Shin, Yasaman Razeghi, Robert L. Logan, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 4222–4235, 10 2020. doi: 10.48550/arxiv.2010.15980. URL `https://arxiv.org/abs/2010.15980v2`.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. Unnatural language inference. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 7329–7346, 2021. doi: 10.18653/V1/2021.ACL-LONG.569. URL `https://aclanthology.org/2021.acl-long.569`.

Derek Tam, Rakesh R. Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. Improving and simplifying pattern exploiting training. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 4980–4991, 2021. doi: 10.18653/V1/2021.EMNLP-MAIN.407. URL `https://aclanthology.org/2021.emnlp-main.407`.

Marta Tatu and Dan Moldovan. A semantic approach to recognizing textual entailment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 371–378, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics. URL `https://aclanthology.org/H05-1047`.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, Oren Pereg, and Hugging Face. Efficient few-shot learning without prompts. 9 2022. doi: 10.48550/arxiv.2209.11055. URL `https://arxiv.org/abs/2209.11055v1`.

Tu Vu, Minh-Thang Luong, Quoc V. Le, Grady Simon, and Mohit Iyyer. STraTA: Self-training with task augmentation for better few-shot learning. 9 2021. URL `http://arxiv.org/abs/2109.06270`.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 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, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. Entailment as few-shot learner. 4 2021. URL http://arxiv.org/abs/2104.14690.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I17-1100.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:1112–1122, 2018. doi: 10.18653/V1/N18-1101. URL https://aclanthology.org/N18-1101.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 2020-December, 4 2019. ISSN 10495258. doi: 10.48550/arxiv.1904.12848. URL https://arxiv.org/abs/1904.12848v6.

Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 3914–3923, 8 2019. doi: 10.48550/arxiv.1909.00161. URL https://arxiv.org/abs/1909.00161v1.

Wenpeng Yin, Nazneen Fatema Rajani, Dragomir Radev, Richard Socher, and Caiming Xiong. Universal natural language processing with limited annotations: Try few-shot textual entailment as a start. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8229–8239, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.660. URL https://aclanthology.org/2020.emnlp-main.660.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. pages 1–9, 6 2021. doi: 10.48550/arxiv.2106.10199. URL https://arxiv.org/abs/2106.10199v5.

Fabio Massimo Zanzotto, Marco Pennacchiotti, and Alessandro Moschitti. A machine learning approach to textual entailment recognition. *Natural Language Engineering*, 15:551–582, 10 2009. ISSN 13513249. doi: 10.1017/S1351324909990143. URL https://www.researchgate.net/publication/220597335_A_machine_learning_approach_to_textual_entailment_recognition.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. Revisiting few-sample BERT fine-tuning. 6 2020. doi: 10.48550/arxiv.2006.05987. URL https://arxiv.org/abs/2006.05987v3.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 35–45, 2017. URL https://nlp.stanford.edu/pubs/zhang2017tacred.pdf.

Appendices

A Zero-shot performance visualization

The boxplots in Figure 7 and Figure 8 show a detailed breakdown of the zero-shot results in Section 4.1 for the coarse- and fine-grained tasks, respectively. Each box represents the variation in the weighted F1 score over the set of templates used in each setting. The three numbers in each box are the maximum, the mean, and the minimum scores achieved.

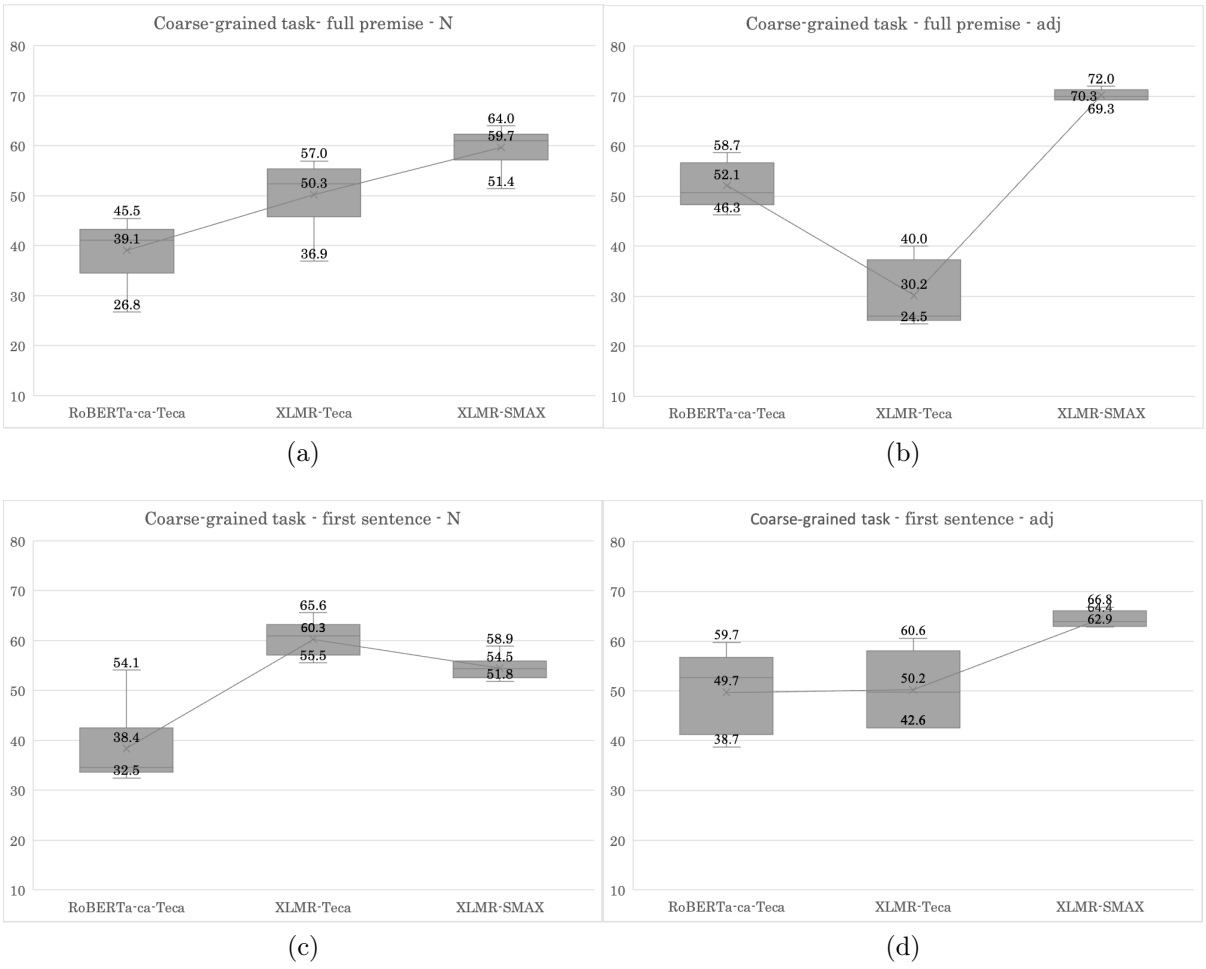
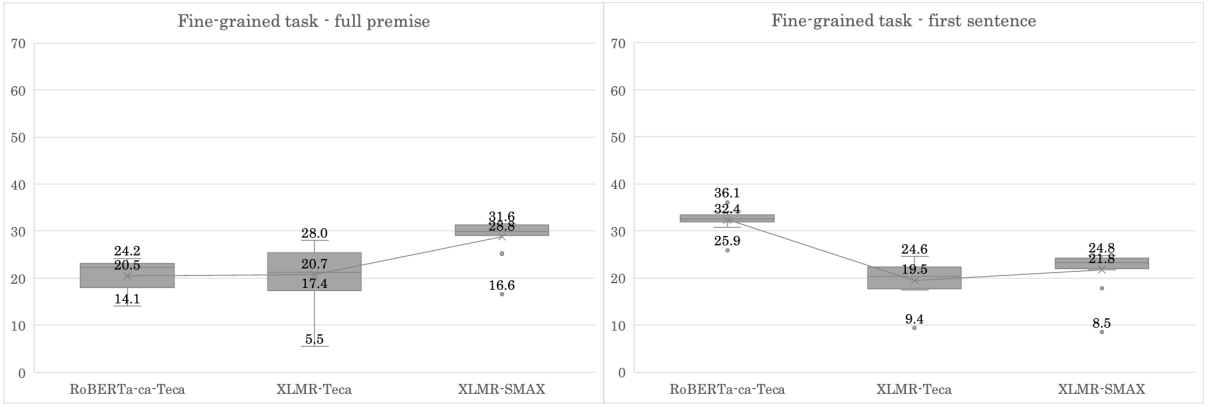


Figure 7: Coarse-grained task performances of RoBERTa-ca-Teca, XLMR-Teca, and XLMR-SMAX in the zero-shot scenario over the TeCla test set. (a) and (b) use the full premise and (c) and (d) the first sentence as premise; (a) and (c) use the template set with nominal labels from Table 6, while (b) and (d) use the template set with adjective labels from Table 7.



(a) (b)

Figure 8: Fine-grained task performances of RoBERTa-ca-Teca, XLMR-Teca, and XLMR-SMAX in the zero-shot scenario over the TeCla test set. (a) shows the results in the full premise setup, and (b) with the first sentence as the premise. In both cases, the template set with nominal labels from Table 6 was used.

B Entailment model's checkpoint selection and negative hypotheses generation strategies

In the few-shot learning experiments conducted, two specific configuration decisions were consistently applied. Firstly, for the training of each entailment model with the available training data, the checkpoint that achieved the highest F1 score in the target task (text classification) on the development set was selected, rather than using the results from the NLI task. Secondly, during the generation of the NLI training data, for each entailment hypothesis (generated using the correct label), all possible negative hypotheses (one for each of the remaining labels) were also generated. To investigate the impact of these decisions, additional experiments were conducted using the RoBERTa-ca-Teca model as the base entailment model: in the 8-4, 16-8, and 32-16 few-shot setups, we converted the available data to the entailment format by creating one non-entailment hypothesis per each entailment one, and we keep the best checkpoint both according to the classification and to the NLI task. These results were then compared to those obtained from the initial experimental setup.

The results of the experiments in the coarse- and fine-grained tasks are presented in Table 19. In the coarse-grained task, there is minimal fluctuation in the results across experiments within each training data regime, and the best-performing model among the three configurations changes at each step. In contrast, in the fine-grained task, the results significantly improve when the best checkpoint is selected based on the classification task performance (by 4.3, 11.5, and 19.0 points compared to the best checkpoint selected according to the NLI task performance). This impact becomes increasingly noticeable as more data becomes available, and the model becomes increasingly unstable (as indicated

		coarse-grained task			*fine-grained task*		
ratio of negative hypotheses	**ckp. selection strategy**	**8-4**	**16-8**	**32-16**	**8-4**	**16-8**	**32-16**
1 negative hip. per positive hip.	best ckp. according to the NLI task	**79.4 ± 4.0**	82.5 ± 1.4	88.1 ± 1.2	41.9 ± 8.4	42.2 ± 18	38.9 ± 17.6
	best ckp. according to the CLS task	78.9 ± 3.7	**83.8 ± 1.9**	87.6 ± 2.2	46.2 ± 0.4	53.7 ± 4.4	57.9 ± 2.7
all possible negative hip. per positive hip.	best ckp. according to the CLS task	79.2 ± 3.3	82.4 ± 2.9	**89.2 ± 0.6**	**60.2 ± 1.4**	**62.4 ± 1.3**	**63.2 ± 1.3**

Table 19: Test set results for the coarse- and fine-grained tasks obtained with RoBERTa-ca-Teca in three few-shot setups (8-4, 16-8, 32-16) using three different decisions with respect to the ratio of negative hypotheses created for training and to the checkpoint selection strategy.

by the high standard deviations obtained).

Furthermore, when the model is trained using all possible non-entailment hypotheses, which in this case implies 53 hypotheses for each example, the results further improve by an astounding 14.0 in the 8-4 setup, by 11.5 in the 16-8 setup, and by 5.3 points in the 32-16 setup. Overall, the results suggest that both choices become particularly important when the number of categories is large. In such cases, the results on the target task are more reliable than the NLI task, and the model appears to be able to benefit from the augmented number of examples in the dataset.